INFORMATION SPREADING BARRIERS IN NEWS

Abdul Sittar

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Prof. Dr. Dunja Mladenić, Jožef Stefan International Postgraduate School and Jožef Stefan Institute, Ljubljana, Slovenia

Evaluation Board:
Asst. Prof. Bernard Ženko, Chair, Jožef Stefan Institute, Ljubljana, and Faculty of Information Studies in Novo mesto, Slovenia
Asst. Prof. Senja Pollak, Member, IPS and Jožef Stefan Institute, Ljubljana, Slovenia
Prof. Dr. Bojana Dalbelo-Bašić, Member, FER, University of Zagreb, Croatia

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Abdul Sittar

INFORMATION SPREADING BARRIERS IN NEWS

Doctoral Dissertation

ANALIZA PREPREK ZA ŠIRJENJE NOVIC

Doktorska disertacija

Supervisor: Prof. Dr. Dunja Mladenić

Ljubljana, Slovenia, January 2024

I would like to dedicate this thesis to my loving parents. ...

Acknowledgments

Thanks to Allah who is the source of all the knowledge in this world, and imparts as much as He wishes to anyone He finds suitable.

To write a dissertation is a mighty undertaking, I would like to thank my supervisor, Prof. Dunja Mladenić for being a fantastic, insightful advisor. She has supported me throughout my work on this dissertation with her patience and knowledge. Without her guidance, motivation, and support this dissertation would not have been completed. One could not wish for a more kind, accessible and friendlier supervisors than her.

I am thankful to my parents for their support, prayers, love, and care throughout my life. They have played a vital role in achieving this milestone. I extend my thanks to my brothers, my sister, in-laws, relatives, and friends for their continuous support and prayers. My wife has always been wonderful to me and extended her wholehearted support, especially during my Ph.D. studies, which I could not have completed without her.

The assistance of the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 812997 in providing financial support is most appreciated.

In the Artificial Intelligence department at the Jožef Stefan Institute, I would like to thank my co-supervisor Marko Grobelnik, and project manager Dr. Polona Skraba Stanic. Many thanks to my colleagues in the department and spread across Slovenia and Pakistan and the rest of the world.

Finally, I am thankful to the Jožef Stefan Institute, and the Jožef Stefan International Postgraduate School in Ljubljana, Slovenia.

Abstract

News spreads in many patterns, structures, and dynamics that change throughout time. For a variety of reasons, certain news is only covered in a particular area. Language, economy, geography, politics, time zone, and culture are just a few of the many barriers that prevent news from reaching a larger audience. Observing these barriers reveals what may influence the spreading of news reporting on different events.

The primary objective of this study is to develop methods and approaches to analyzing news spreading barriers, with a particular emphasis on the above mentioned five barriers (linguistic, economic, geographic, political, and cultural). The aim of the analysis of geographical and time zone barriers is to identify the influence of time zone and geographic position of news publishers on news spreading across the globe. Analysis of political barrier is carried out to see the influence of political alignment of news publishers on news reporting. The aim of the analysis of cultural and economic barriers is to identify the impact of locations of news publishers with different cultures and economies on news spreading. Analysis of the linguistic barriers is performed to determine the influence of publishing language on news propagation.

This thesis focuses on three interconnected issues to the news spreading barriers. The first issue involves adopting information cascade theory for news articles and event-centric news dissemination analysis. The evaluation of the improved topic modeling and the strategy for comprehending political and economic contrasts in news reporting constitute the second issue. The third issue is to profile the news spreading barrier, a task to classify news texts based on the stylistic choices of their news publishers.

A methodology is presented for the analysis of information propagation in news across different barriers in different domains. To deal with the analysis of monolingual and multilingual cascading analysis, different visualizations are incorporated. For all the five barriers (linguistic, economic, geographic, political, and cultural), various analytical methods are employed in this methodology. News related to three different domains (natural disasters, climate change, and sports) is explored. The findings revealed that 1) the scope of a specific event significantly affects the news spreading across languages, 2) the geographical size of a news publisher's country is directly proportional to the number of publishers and articles reporting on the same information, 3) countries with shorter time-zone differences and similar cultures tend to propagate news between each other, 4) news related to global warming comes across economic barriers more smoothly than news related to FIFA World Cup and earthquakes and 5) events which may in some way involve political benefits are mostly published by those publishers which are not politically neutral.

A methodology is presented to comprehend the political, and economic disparities in news reporting and to compare the sentiments expressed in various newspapers across various geographic regions. In order to increase the quality of topics without changing the fundamental structure of Latent Dirichlet Allocation (LDA), an improved topic modeling technique is suggested that employs LDA with varying length of words and articles' pooling depending on queries. The COVID-19 news is used as an example in a variety of political

and economic situations. Our findings imply that news reporting by newspapers with different political alignments supports the reported content. Also, economic issues reported by newspapers depend on the economy of the place to which a newspaper belongs to.

An approach is presented for automatic barrier profiling using news meta-data. All the data about news is obtained from the Event Registry global media monitoring system and enhanced with metadata about the publishers fetched from Wikipedia. To annotate the news articles across different barriers, annotation procedures are set. To deal with the similarity between metadata of news articles, Euclidean distance is calculated and automatic annotation is carried out. Barrier classification is performed using news headlines, common sense inferences and sentiments. Evaluation is carried out using machine learning classical classification methods, deep learning and transformer-based methods.

Povzetek

Novice se širijo po različnih vzorcih ter imajo različno dinamiko širjenja skozi čas. Pokritost nekaterih novic je zaradi številnih razlogov omejena na določeno regijo. Novice potencialno prečkajo številne prepreke, kot so jezikovne, gospodarske, geografske, politične, časovne in kulturne. Opazovanje teh preprek nam daje vpogled v to, kaj lahko vpliva na širjenje novic o različnih dogodkih.

Glavni cilj te raziskave je razviti metodologijo za analizo preprek za širjenje novic s poudarkom na zgoraj omenjenih šestih preprekah. Cilj analize geografskih in časovnih preprek je ugotoviti vpliv časovnega pasu in geografskega položaja založnikov novic na širjenje novic po svetu. Analizo političnih preprek smo naredili, da bi ugotovili morebitni vpliv politične usmeritve založnikov novic na njihovo poročanje. Cilj analize kulturnih in ekonomskih preprek je ugotoviti vpliv lokacij založnikov novic iz različnih kultur in gospodarskih pogojev na širjenje novic. Analiza jezikovnih preprek je pokazala vpliv jezika, v katerem je novica napisana, na širjenje novic.

Disertacija se osredotoča na tri probleme, povezane s preprekami za širjenje novic. Prvi problem vključuje uporabo teorije informacijskih kaskad in analizo dogodkovno usmerjenega širjenja novic. Drugi problem je tematsko modeliranje in pristopa k razumevanju političnih in ekonomskih razlik v poročanju novic. Tretji je profiliranje preprek za širjenje novic, zasnovano na stilu poročanja.

Predlagana je metodologija za analizo preprek za širjenje novic (informacij, o katerih novice poročajo) na različnih domenskih področjih. Uporabili smo različne analitične metode in različne načine vizualizacije. V doktorski nalogi obravnavamo novice, povezane s tremi domenskimi področji (naravne nesreče, podnebne spremembe in šport). Rezultati so pokazali, da 1) obseg določenega dogodka pomembno vpliva na širjenje novic v različnih jezikih, 2) je število založnikov in člankov, ki poročajo o isti informaciji, neposredno sorazmerno z geografsko velikostjo države založnikov, 3) države s krajšimi razlikami v časovnih pasovih in podobno kulturo med seboj širijo novice, 4) novice, povezane z globalnim segrevanjem, lažje naletijo na gospodarske prepreke kot novice, povezane s svetovnim prvenstvom v nogometu, in novice o potresih in 5) dogodke, ki lahko na nek način vključujejo politične koristi, večinoma objavljajo tiste založbe, ki niso politično nevtralne.

Predlagana je metodologija za razumevanje političnih in ekonomskih razlik v poročanju novic ter primerjava sentimenta med časopisi na različnih geografskih območjih. Za izboljšanje kategorizacije novic predlagamo nadgradnjo LDA metode (angl. Latent Dirichlet Allocation) z uporabo zaporedij besedil različne dolžine. Na primeru novic o COVID-19 smo pokazali razlike v poročanju v odvisnosti od političnih in gospodarskih okoliščin založnika. Naše ugotovitve kažejo, da založniki z različnimi političnimi usmeritvami poročajo različno o isti vsebini. Pokazalo se je tudi, da je izbira gospodarskih vprašanj, o katerih poročajo časopisi, odvisna od gospodarske situacije v državi založnika.

Predstavili smo pristop k samodejnemu profiliranju preprek z uporabo metapodatkov novic. Vse podatke o novicah smo pridobili iz javno dostopnega sistema za spremljanje medijev Event Registry. Za njih smo pridobili dodatne metapodatke o založnikih iz Wiki-

pedije. Za potrebe označevanja člankov glede na različne prepreke smo definirali postopek za označevanje. Podobnost med metapodatki novic smo izračunali z uporabo evklidske razdalje. Klasifikacijo preprek smo naredili na osnovi naslovov novic z uporabo metode za zdravorazumsko sklepanje iz vsebine besedila in z uporabo določanja sentimenta novice. Evalvacija vključuje uporabo klasičnih klasifikacijskih metod strojnega učenja, globokega učenja in metod na osnovi Transformer arhitekture.

Contents

Abbreviations

ML	...	Machine Learning
LDA	...	Latent Dirichlet Allocation
LSA	...	Latent Semantic Analysis
OEKG	...	Open Event Knowledge Graph
NLP	...	Natural Language Processing
TFIDF	...	Term Frequency-Inverse Document Frequency
SVM	...	Support Vector Machine
OSN	...	Online Social Networks
CTMC	...	Continuous-time Markov Chain
SLR	...	Systematic Literature Review
GDP	...	Gross Domestic Product
UK	...	United Kingdom
TV	...	Television
UNESCO	...	United Nations Educational Scientific and Cultural Organization
NYT	...	New York Times
NBC	...	National Broadcasting Company
GDELT	...	Global Data of Events, Location, and Tone
PSB	...	Public Service Broadcaster
SNS	...	Social Network Services
HNCD	...	Hofsted National Culture Dimensions
UAE	...	United Arab Emirates
DMOZ	...	Directory Mozilla
DTW	...	Dynamic Time Warpping
TM	...	Topic Modelling
CSV	...	Comma-separated values
JSON	...	JavaScript Object Notation
HDI	...	Human Development Index
SARS	...	Severe Acute Respiratory Syndrome
IBM	...	International Business Machines Corporation
IDE	...	Integrated Development Environment
UTC	...	Coordinated Universal Time
NCCA	...	National Commission for Culture and Arts
WHO	...	World Health Organization
WHONASA	...	World Health Organization - National Aeronautics and Space Administration
PDI	...	Power Distance
IDV	...	Individualism
MAS	...	Masculinity vs Femininity
UAI	...	Uncertainty avoidance by individualism
LTO	...	Long and Short Term Orientation
IVR	...	Indulgence vs Restraint
POS	...	Part-of-speech

BERT	...	Bidirectional Encoder Representations from Transformers
NRC	...	National Research Council
VADER	...	Valence Aware Dictionary and Sentiment Reasoner
SA	...	Sentiment Analysis
CNKI	...	China National Knowledge Infrastructure
NPMI	...	Normalized Point-wise Mutual Information
SNS	...	Social Network Sites
IDT	...	Innovation Diffusion Theory
TAM	...	Technology Acceptance Model
CNN	...	Convolution Neural Network
GRU	...	Gated Recurrent Unit
BiGRU	...	Bidirectional Gated Recurrent Unit
LSTM	...	Long Short-Term Memory
BiLSTM	...	Bidirectional Long Short-Term Memory
DPA	...	Deutsche Presse-Agentur
RNN	...	Recursive Neural Networks
PAN	...	Plagiarism and Authorship Analysis
API	...	Application Programming Interface
URI	...	Uniform Resource Identifier
URL	...	Universal Resource Locators
GB	...	Geographical Barrier
LB	...	Linguistic Barrier
PB	...	Political Barrier
SMOTE	...	Synthetic Minority Oversampling Technique
CNN	...	Convolutional Neural Network
LR	...	Logistic Regression
CB	...	Cultural Barrier
EB	...	Economic Barrier
GNN	...	Gated Neural Network

Chapter 1

Introduction

1.1 General Introduction

News is information regarding recent happenings. It may be accessed through a variety of media, including print and electronic communication. Various social media platforms and online newspapers are a few instances of electronic communication on news and are highly interactive, easy to use, and user-friendly [1]. Furthermore, with respect to economic aspects, the newspaper industry has a long history of being seen as a reliable and strongly related to advertising and financial success. However, the future of newspapers has been called into doubt with the advent of the Internet, propaganda, and communication networks. Due to market competition, consumers and advertisers are under pressure to switch from printed media to online newspapers. A similar study has found that there is a significant and strong relationship between print and online presence. The printed version has characteristics such as business year, price, number of issues per week, and classified advertising rates whereas online newspapers have features such as interactivity, daily unique users, mobile internet users, and user-generated content [2].

The news media serve as the main source of information for the public regarding domestic and foreign events including policy, international conflicts and wars, natural disasters, political upheavals and elections, economic crisis, cultural and sporting events, and health emergencies and pandemics. They are very important for society as they inform citizens about the events around them and how they can impact their life. Their importance becomes more crucial and indispensable in times of health crises such as the recent COVID-19 pandemic [3]. The use of news media is also associated with civic engagement and the cultivation of social capital. Citizens who use news media are more likely to trust their community, participate in community groups, engage in political discourse with neighbors, and have higher levels of social capital than those who do not consume local media. Also it has been shown that newspapers have a positive relationship to readers in different ways such as, sense of community, political participation, and civic engagement [4].

In addition, the convergence of the newspaper and the Internet creates the online newspaper market. Technologically, the Internet is an inherently global medium, which causes what Frances Cairncross [5] calls "the death of distance" because any information published online is equally accessible by Internet users worldwide . This revolution has opened the debate on the effects of new media, separating euphoric scholars from the more skeptical ones. In this respect, one may claim that the growing development of communications via the Internet and its subsequent use as a medium for publishing the digital versions of most printed newspapers, has led to substantial changes in the newspaper business. On the other hand, agenda-setting theory suggests that, beyond the dissemination of information, the news media can raise public awareness and the political importance of issues through

extensive press coverage [6].

There are multiple threats to the media industry such as economic, editorial, and technological challenges. The media industry, especially newspapers, are facing changed business conditions. These challenges appear when competitors launch innovation in their product that competes with their products and services. Moreover, low distribution cost and speed with which news items reach readers are also a main part of the priorities of the industry. One of the most affected sectors is journalism and, specifically, the newspaper business. Digital media and the internet have brought about a revolution in journalism.

Besides, there are several problems attached to local, regional, and international newspapers such as poor reading culture, low literacy level, high cost, and fewer circulation [7]. One important factor for its success can be a change in the political and cultural identities. Some approaches have been adopted that have reduced attention to government policy and elections [8]. On the other hand, one of the factors for the public not being much aware of the city politics is that there is a lack of professional expertise in coverage of local government news [9]. These are the intensified struggles in our current media landscape concerning technological, social, and political dynamics [5].

Before disseminating the news, newspapers have to pass through gatekeepers. Gatekeeping is a process that involves a complex series of operations that extend throughout the entire news production and dissemination. The process of news-gathering, news writing, and dissemination has come under scrutiny on a large scale because people's sense of reality is influenced by what gets into the news and what gets left out [10]. Besides local news events, the central and crucial position is given to the editors in the world of foreign affairs that works as gatekeepers. Hence, media gate-keeping is one of the most enduring areas of research in media sociology [11].

The factors influencing coverage of news events include news values, journalists' professional routines, audience consideration, and organizational influences. The topic of news coverage has long been discussed considering various explanations of foreign news coverage. One is context-oriented which looks at the relationship of events with contextual variables such as trade relations, cultural relevance, political involvement, and geographical proximity [12]. The second one is content-oriented that reflects corporate preferences such as market incentives, the constraints of the political and economic environment, party systems, wealth, and education [11].

1.2 Information Spreading

The process of information traveling from a sender to a set of receivers via a carrier is commonly referred to as diffusion [13]. It has attracted a great number of researchers in recent years, concentrating on topics including misinformation, trust, rumors, and so forth. There are mainly two ways to information diffusion by newspapers and online social networks. Information diffusion through online social networks is an interesting area of research [14], because they play an important role in spreading information on a large scale. They spread information from one individual or community to another and hence draw an increasing attention of scholars and governments for its importance especially in disaster response. Understanding its dynamics can help governments to disseminate information effectively [15]. It has been widely recognized that information diffusion is a more complex contagion than the spread of infectious diseases [16]. It has changed rapidly in recent years with the emergence of social media which provides online platforms for people worldwide to share their thoughts, activities, and emotions, and build social relationships [17]. Many studies on social media, particularly twitter analytics, emphasize the aspects of information diffusion, information spread, and influencers [13].

One of the threats to newspapers is news circulation that is falling due to social and technological changes. There are other problems attached to this factor: such as poor reading culture, low literacy level, and high-cost [1]. Similarly, the stability of the newspaper industry concerning advertisers and the economy is at risk due to the advent of the internet, and communication networks [8]. Also the local and regional press has different threats to their survival. The regional newspapers were more scandalous, less professional, and of a poor quality than the national newspapers. Also they have been operating under conditions of severe financial stress and decline in recent years [18], [19].

Journalists' decisions and contextual and systemic factors are among the major influencing factors to news spreading. These contextual and systemic factors of news flow across national borders can be separated into four groups: (1) economic interaction; (2) presence of international news agencies; (3) traits of nation; and (4) cultural and geographic proximity to the US [20]. Regarding journalists, there are five levels that influence the journalists' news decisions: individual level, media-routine level, organizational level, social-institution level, and social-system level. These levels include different aspects such as general patterns of communication work, characteristics of organizations, governments, advertisers, social structures, ideologies, and cultures [21].

Personal interactions with the political leaders and background knowledge of journalists have a strong correlation with news making [21]. However, the influence of different factors varies depending upon the issues that journalists deal with. Since they have to report on uncertain and unexpected events, various factors influence it at different levels. Studies showed that government officials have a major influence on how the media cover related events. Additionally, the other characteristics of journalists' reporting also influence international events [22]. Many scholars have investigated the factors that influence how journalists cover certain news events, in an attempt to enhance our understanding of the production of news stories [21]. In discussing the relationship between journalists and news sources, it is important to note that journalistic objectivity plays an important part in journalists' use of sources in their presentation of news. Also, they are regarded as objective when they let high-profile sources dictate the news, whereas journalists are viewed as biased if they come up with conclusions based on their expertise.

Furthermore, numerous factors affect journalists' news decisions other than news sources. First, journalists' characteristics can influence their coverage of certain issues. Factors at the individual level include journalists' personalities, values, religion, experiences, attitudes, and role conceptions. In particular, studies showed that journalists' political orientation and professional experience influence how they cover certain issues. The characteristics of an individual news organization can also affect journalists' news decisions. Some studies suggest that journalists become socialized to their employing media organization's political ideology and way of thinking as they adjust to their work environment.

Over the years, scholars have studied the relationship between the news prominence of a country and its physical, economic, political, social, and cultural characteristics [23]. Communication scholars have long been interested in identifying the key determinants of what makes foreign countries newsworthy and why some countries are considered more newsworthy than others. Research on international news flow has identified dozens of variables that correlate with foreign nation visibility in the news, but most of this research has been cross-sectional, focusing on the visibility of foreign nations at a particular point in time. Given the difficulty of gathering longitudinal data, relatively little news flow research has systematically examined whether and to what extent foreign nation visibility and the factors that influence it have changed over time. Specifically, scholarship has typically only addressed why some countries get more news coverage than others at a specific point in time, not how and why the focus shifts over time from one country to another [24].

The studies conducted over the last few decades to investigate the determinants of international news flow are systematically reviewed and analyzed. Their findings can be divided into two broad categories: gatekeeper perspective and logistical perspective. A synthetic profile of the past studies is also provided, which includes the regions and topics covered by international news, the geographic distribution of the studies' locales, and space/news-hole allotted for different regions [25].

News media not only influence the access to information but also influence the way of information presentation and framing, attitude, and shaping of the mind. Media coverage exerts an influence on opinions in different fields, e.g., political, mental illness, and gambling. In cognitive information processing, if the persuasiveness is high enough, then information can lead to a change in people's attitudes [26].

News framing is a process through which journalists select and emphasize certain aspects of social realities [27]. It is represented with linguistic features such as certain words, phrases, and expressions. In different cultures, framing consists of a different set of features. Framing from the perspective of media and individuals, is sometimes reciprocal and influences each other like in agenda-setting-media and public agenda [28]. The assumption behind framing is that how an issue is described in news reports influences how audiences understand it, which is closely related to how a news story is presented. Frames are being used more realistically and frequently by both political actors and journalists to present political reality. Frames are associated with a journalist's cognitive schema about an event, whereas framing is more concerned with such frames as embodied on the discourse level [29].

1.2.1 News Spreading Barriers

News related to different events is engaged in a form of competition for publication at every moment across the world. This does not only mean that the event is worthy enough to be covered but also the existence of an international market of other competing events simultaneously occurring that matters. These competitions can help to explain what are the news barriers. For instance, it may be the existence of continuity or discontinuity in the coverage of event stories [26]. Also, in the past few years, the research on social, economic, and cultural life through news articles has grown exponentially [30].

Journalists are a key part of news spreading. It is still influenced by the remnants of traditions such as the priority of gathered information, and distrust of the internet as a source of knowledge [31]. This relationship that connects journalists and sources affects how journalists can access, interpret and use information. This does not involve only their professional roles but also develops a social relationship. Therefore, relationships that are ever-present behind the content are an invisible yet crucial part of the profession. Understanding the interactions and motives that remain behind the news should shed more light on why the story looks the way it does. Scholars have identified an important emotional role in news media's coverage of international disasters; inviting the audience to care for people in need who are not like us. News media thus play a pivotal role in giving publicity and meaning to these numerous instances of global suffering as it is essentially through media reports that the western world perceives international disasters [32].

Bias is another factor that influences the news spreading. A study on interrogating gender media coverage was conducted by taking distinguished political celebrities. The data consisted of 529 news items which were taken from three Israeli online newspapers, articles, and reports covering publicly renowned figures. Each item of news was labeled into one of three gender-coverage frames. A critical reading of these journalistic texts was also carried out. The findings show that combining the two types of biases, gender and political, increases media sensation and stereotyping [33].

Journalism often involves traveling to cover news across different socio-cultural environments. This traveling could also create an imbalance in news spreading. According to the findings of a comparative study of newspaper travel sections in Australia, Canada, New Zealand, and the United Kingdom, travel journalism frequently reproduces the imbalances found in foreign news flows. Well-known determinants in travel journalism include regionalism, advanced nations, cultural closeness, the role of large neighbors, and diversity of coverage. Simultaneously, a country's tourist behavior plays a role, but it is frequently overshadowed by other factors [34]. It is also notable that distance does not always lead to indifference, nor does proximity always lead to identification and pity.

The different approaches to coverage of news events are influenced by several factors that include journalists' professional routines, news values, audience considerations, and organizational influences. Over the years, scholars and researchers have presented a variety of theoretical explanations about various influences on foreign news coverage. Among these studies, theoretical thinking can be grouped into two perspectives: context-oriented versus content-oriented. The former looks at the origin of a foreign news event and its relationship with such contextual variables as trade relations, cultural relevance, political involvement, and geographical proximity [35].

1.2.1.1 Language Barrier

News media has a very important role in the integrity of a society. It is a really important and challenging task to cover news across language boundaries when more than one language is spoken in a region or country. Multilingual countries depend on institutions like journalism that contribute to integration not only at the regional and local levels but also at the national level in fostering understanding for others and a common sense of identity [36].

Language is used to represent a certain culture. It promotes cultural diffusion for people to use language to communicate with each other. It is deeply rooted in national cultures and traditions and cannot be extended without the culture. Thus, ethnic nationality and values are maintained by language. Several studies performed coverage of different issues across different languages. A contrasting critical analysis of Iranian and British newspapers investigated how both represented Iran's nuclear power program following different sociopolitical patterns. Another aspect of language analysis is to develop a bilingual entity lexicon. This can be done by clustering the news based on similarity computation. This could also help to grasp the current international and regional hot events after grouping news that is written in different languages [37].

1.2.1.2 Cultural Barrier

Culture is a complex social-specific topic that lacks a commonly accepted definition [38]. There are different definitions of culture in literature. Culture is "patterned way of thinking, feeling and reacting that constitutes the distinctive way of life of a group of people." Similarly, Hofstede [38] considers culture as the collective programming of the mind that distinguishes the members of one group or category of people from others. Hofstede's national cultural dimension (HNCD) theory has been widely applied in studies of national and organizational culture. It has a profound impact on an international events, such as the Arab Spring [39]. Huang and Chang found that internet users in countries that have a similar language, religious beliefs, social norms, and economic development were likely to visit similar websites. Similarly, Gevorgyan and Manucharova [40] found that the web design preferences of users have direct relations with the individualistic and collectivistic dimensions of HNCD. Similarly, user perception has been investigated on chatbot-driven

news and chatbot journalism across United States and the United Arab Emirates.

The news selection process takes place within a complex framework shaped by socio-cultural, organizational, and psychological variables. The most important and influential factor of news selection is the theory of newsworthiness that integrates concepts of perception, professional and organization routines, as well as the anticipation of audience interests [22].

When news events happen in countries that are geographically or culturally close, domestication of news is done to make the coverage more relevant to the local audiences. News outlets employ their staff, correspondents, or stringers for coverage of regional news as this would allow more editorial direction and control, helping to domesticate the events for the publication's key audiences. Thus, economically strong publications are most likely to rely less on international news agencies to cover regional news [35].

As it has been argued before, cultural factors appear to have a growing influence on national identities in times of political and economic globalization [41]. Globalization is also a mega-trend from the economic and cultural point of view [41]. The relatively stable national cultures are sources of global diversity. At the macro level, different nations compete on the global stage to reshape the dominant cultural norms and the perceived world according to their preferred cultural frames. The news coverage from the world's major international news agencies embedded with different national cultural frames are valuable resources for scholars to track and examine the symbolic competition among different nation-states [39]. Translation study as a discipline borrows heavily from linguistic, cultural, and sociological studies and is dominated heavily by various dichotomies.

Several hypotheses have been explored given the importance of cultural values such as the countries with English as their national language can lead to a greater amount of news coverage on the web [42]. Determinants of global news flow are not consistent between countries due to cultural differences, countries with similar cultural backgrounds communicate better and might result in heavier news flow. It has been explored that there should be common connections among people of different religions, cultures, political views, races, national origins, and mutual respect [39].

The relationships between cultural factors and information seeking and use have been confirmed widely [22]. A study performed an analysis of the impact of different dimensions on Hofstede's cultural dimensions on the effectiveness of change management processes. The findings of this study significantly present proof that cross-cultural dynamics influence change management initiatives and key drivers [43].

When international news is translated to another language, then cultural transformation happens. A study presents various elements playing a role in news production in a language that differs from the one in which the news is ultimately presented to the customer. It is completely justified to say that news translation may significantly affect people's perception and interpretation of the surrounding world [44]. It has focused on how competitive cultural dynamics have contributed to shaping the evolution of online journalism. In an analysis of the evolution of American newspapers, it has appeared that the cultural factors have contributed to the more conservative and less successful path that these newspapers have had in comparison to sites not affiliated with traditional news firms. The study of the relation between social culture and industry technologies shows that industry technologies and social culture lead to better organizational performance mainly in the dimensions of masculinity, power distance, and individualism [45].

1.2.1.3 Geographical Barrier

It has been indicated that there is a strong influence of geographic distance on international news flow. A hypothesis related to the geographic proximity factor has been investigated

saying that a nation's geographic distance to the US can predict its amount of news coverage on the web [42]. Numerous studies have focused on detecting the events and locating them geographically. To study real-world events, and find relationships between locations based on their coverage, several models are proposed. These models can also provide insights about bias at different levels of news coverage in the world [46].

Based on theories of news values and news factors, research on international news flows and international news geography repeatedly suggests that large-size effects are at play. For transregional coverage, this effect has hardly been examined. According to data from regular reports on public radio and public TV news in Switzerland, however, similar mechanisms can be assumed [36]. Several new factors have emerged that can influence the methodological approach to online news geography [47]. Online news geography shows that local advertisers are not particularly interested in long-distance users. An industry trade analysis of the top 100 markets showed more than one in three active internet users visit newspaper web sites. These factors suggest a reexamination of online newspapers' local and long-distance usage is essential to fully understand the role of geography [47]. According to the findings of a study, online newspaper penetration is higher in the local market, but the local market accounts for less than half of total traffic, implying that the size of the long-distance readership is larger than previously anticipated. Larger newspapers typically have a larger online audience, but all newspapers receive a significant portion of their online traffic from sources other than the print market. Online or offline, geography is still important. It has been found that the theoretical and practical implications of geography are key factors in online news economics.

Online news geography has been explored extensively using a survey of sixty-four online newspapers. The review of the media economics literature found little empirical research on online newspapers' usage in local and long-distance markets. This secondary data analysis of 136 U.S. online newspapers' audience investigates how geography differentiates online newspapers' local and long-distance markets in terms of penetration, size, relative importance, and the relationship between circulation and online usage. The findings of various studies show that the countries with geographic proximity communicate better and might result in heavier news flow [42]. A study explains the effect of geographic distance on online social interactions to understand the interplay between the social characteristics of friendship ties and their spatial properties. There are a plethora of applications and systems that mine online social interactions to provide suggestions, offer recommendations, and filter information. Information related to Social Network Services (SNSs) can even be used to improve existing distributed systems and applications [48].

Most newspapers and other media use geographic bias to promote their content. But geographic and demographic bias can lead to an inaccurate and incomplete view of the news in a country. Normally geographical and temporal features of news are investigated for such problems. The flow of news is influenced by external and internal factors. Political systems and economic pressures are one of them. It is generally assumed that international news coverage reflects the power structure among nations. There are more influences involved in international events coverage such as organizational factors, the local community's power structure, and corporate characteristics. A great deal of research has focused on explaining the variations of news coverage in different nations and regions [49].

1.2.1.4 Economic Barrier

When news related to specific events is not covered or spread to certain types of economies, then it is said that there is an economic barrier. Economic power is a determinant of news flow. For example, the magnitude of economic interactivity between the US and other countries affects the news flow and this influence seems to escalate as the impact of trade

volume wanes slightly. The geographic size and population of a nation are positively related to the nation's news quantity only in the media of some European countries but not in others [42].

Today, with the increasing level of globalization of news organizations in the context of greater political and economic interdependence among nations, one could assume that the role of foreign news would increase in terms of its importance. A third longitudinal finding that warrants emphasis has to do with the relationship between trade flow and news flow. Their results show that in the post–World War II era trade flow has become less predictive of visibility in the New York Times (NYT). The trade variable was slightly significant in the most recent period, but over time it seems to have lost its status as a stable determiner of foreign nation visibility in the NYT. Furthermore, the NBC data indicate that trade flow was never a significant (positive) predictor of visibility [24]. Studies show that trade is positively linked to the volume of news coverage. The level of newsworthiness in the media is affected by the size of countries such as the bigger territory with more populations that seem to carry more weight that can be translated into voluminous coverage.

1.2.1.5 Political Barrier

The news is nowadays mostly affected by political barriers. When news related to specific events is not covered or spread by the news publishers with a specific political alignment, then it is considered that there is a political barrier. News diversity is one of the important aspects that have high priority as a result of political results and is investigated by governments and regulatory bodies [6]. Comparison of international events on the importance of news diversity has documented a plethora of research. Based on this understanding, many studies consider the diversity of political (elite) actors represented in the news media. They investigate the diversity of speakers in the news coverage produced by individual outlets to determine whether there is a balanced representation of political interests.

Other scholars in this field are examining the geographical perspectives taken by the news media [50]. Theoretically, political interests can be stratified geographically, with elite interests concentrated at the national level and non-elite interests at the local, regional, or even global levels [6]. Studies showed that sources have a major influence on the media's coverage of news events, particularly foreign affairs issues. There are some other interesting findings in this study. Journalists' political orientation showed a significant correlation with journalists' perceptions.

Political coverage is strongly influenced by interactions between journalists and political actors. Especially for political actors, these interactions present an opportunity to increase their influence on the news. However, what strategies political actors use in their attempts to steer political journalists when exchanging with them has not been studied comprehensively and on a broad basis [51]. Furthermore, some studies suggest that interactions are dominated by political actors, while others conclude that journalists are at least equally influential [52]. The media are the lens through which citizens gather political information on the performance of political institutions and adjust their attitudes accordingly. Scholars proposed contrasting theories about media's effect on political trust [53].

Earlier studies that sought to understand media coverage of politicians and political issues have identified a wide variety of factors influencing whether and how actors and issues are covered. On the one hand, these factors can be identified in the journalistic news production process [54]. Interest groups compete for visibility in the news media, which is important for political influence. However, actors in the same policy field tend to form close bonds and collaborate when their interests overlap. Existing source research has overwhelmingly focused on journalist-source relationships, with far less attention paid to how news content is negotiated between sources themselves [55]. Media coverage of political

activities around the world is powerfully influenced by interactions between information gatekeepers and other actors in the political sphere. Maurer and Beiler [52] affirmed that for political actors, such interactions present the chance to intensify their influence on the news.

1.2.2 Profiling of News Spreading Barriers

The success of a newspaper is measured by the number of copies distributed on an average day. So news agencies/publishers always want to have more viewership of their content. High levels of circulation mean more advertising revenues. Profiling the news spreading barriers can show us the limitations in reporting on issues related to different events. It can also be helpful in the context of numerous real-world applications, such as event-centric news analysis, suspicious news detection, and content recommendations to readers and subscribers. Current journalism requires many financial, technical, and ethical changes to improve its business model [2].

Viewership of news is one of the main pillars of success for news publishers. To increase it, significant efforts are required at different stages such as input-output analyses would be required to assess the different stages of news production [22]. Another similar process is that the new reporting has to pass through gatekeepers. As we explained earlier in Section 1.1 gate-keeping is a process that consists of a series of operations that extend throughout the news production and dissemination process. It can be studied on many levels of analysis, with many different research methods [10]. Another term is news framing through which journalists select and emphasize certain aspects of social realities [27]. News framing is represented with linguistic features such as certain words, phrases, and expressions. In different cultures, framing consists of a different set of features. It is closely related to a presentation of a news story and is based on an assumption of how an issue is modeled in news reports to how it is understood by people. Framing from the perspective of media and individual is sometimes reciprocal and influences each other like in agenda-setting-media and public agenda [28].

News framing, and news reporting with agenda-setting, are interrelated terms. News spreading is directly related to the way news is reported about any events. There are many studies available that research understanding the framing of news reporting on different topics and in different cultures. Framing of European political and economic news found similarities between media in the Netherlands, Denmark, Germany, and Britain with more emphasis on the conflict over economic frames [12]. Another study suggested that US media relied more on the military conflict frame, and Swedish media emphasized the responsibility and anti-war protest frames [12]. Framing of political issues in local news media is an important research problem that is analyzed by taking militarized conflict zones (Kashmir region) as an example.

Setting an international agenda for political conflicts has been studied in detail in the past. It has been investigated whether media countries, conflict involvement, and crisis phases influence the employment of the issue framing and strategy framing by taking the news coverage from China, the US, Singapore, and Ireland [27]. When news is reported on the same event by other publishers, many barriers may stop it from spreading further. These barriers could be political, economic, linguistic, cultural, and geographical, as explained in Section 1.2.1. In this context, the automatic barrier profiling in news spreading is getting attention as an important research problem that requires a method to improve news spreading. Regarding translation study, heavy material is borrowed from cultural, sociological, and linguistic studies. The linguistic discipline focuses on text differences in written or oral form, translated and source text [56].

1.3 Aims and Hypothesis

Hypothesis 1: *Depending on the nature of an event, there will be variations in information spreading behavior across different barriers.*

Prior state of the art: Various aspects of information spreading have been studied in the past such as temporal aspects of sports-related information spreading, information diffusion in online social networks, and modeling of information spreading under disaster. The existing research on information cascading focuses on social networks and relies on social media features (e.g., retweets, sharing).

Aim: Our study focuses on modeling events in different domains (sports, natural disasters, and climate change) and we have introduced a novel approach based on cascading to news spreading. We aimed to analyze multiple influences on the news spreading in three types of events (earthquakes, FIFA World Cup, and global warming) belonging to three different domains. We focus on information cascading and cross-lingual information spreading across geographical, economic, time zone, political, and cultural barriers.

Hypothesis 2: *The topics present in news related to the COVID-19 pandemic vary according to the publisher's political alignments and the economic situation of a specific area. Moreover, the topics that have strong relationships with each other will have similar trends over time.*

Prior state of the art: Previous studies used pooling based on other parameters and applied pooling based on hashtags on Twitter datasets, while some of them proposed a scheme for pooling tweets into longer documents based on conversations. These studies, however, do not conduct experiments on large timespans that may provide a better overview of the pandemic. On the other hand, numerous studies have investigated the impact of COVID-19 in different countries such as how different discussions evolved and the spatial analysis of tweets addressing the diffusion of information about COVID-19 using a large amount of data from popular social media networks. However, the problem of pooling based on user queries is not explored in news articles.

Aim: The aim of proposing the enhanced topic modeling approach was to improve the quality of topics without modifying the basic structure of standard LDA. For that, we evaluate the model based on the coherence score. Another aim was to analyze the frequent topics across different political alignments and different levels of economic prosperity to find correlations between topics and political alignments and correlations between certain topics and different levels of economic prosperity.

Hypothesis 3: *Common sense-based semantic knowledge and sentiments of news headlines will help to classifying the barriers to the spreading of news.*

Prior state of the art: A vast body of literature exists on how the news media frame the news events and consequently influence public perception of those events. Existing literature posits that framing is often used intentionally to change the perception of content and to cater to this, different computational methods have been applied. However, the detection and classification of such influences based on the stylistic choices of their news publishers are not explored.

Aim: The goal of this work was to define the task of automatic barrier profiling and to design a methodology for automatic annotation of news articles through analysis of metadata of news publishers.

1.4 Scientific Contributions

The main contributions of the thesis are:

1. A novel methodology to analyze the news spreading barriers on different kinds of news events.
This methodology is based on a combination of different types of analytical methods: network and cascading analysis for mono-lingual and multilingual temporal information, respectively; alluvial, and chord diagrams for economic and cultural information, respectively; Google maps for time zone and geographical information;

2. A novel approach to enhance the topic modeling technique and understand political and economic differences in news reporting.
This approach is based on the following items: data collection connected to different political alignments and different levels of economic prosperity; sentiments analysis; topic modeling; and news articles' pooling based on user queries.

3. An approach to barrier profiling by automatically annotating and classifying the news articles for the different barriers.
This approach is based on automatic annotation of news articles using meta-data, common sense inferences and sentiments of news headlines, machine learning classical classification methods, deep learning, and transformer based methods.

1.5 Thesis Structure

The thesis starts with a brief general background and an explanation of the topic. It then explains spreading barriers and the context of the problems related to news spreading barriers. It basically relies on a set of research papers.

Chapter 2 deals with data collection, and analysis of news spreading barriers. It is composed of three separate thematic sections. Section 2.1 deals with the process of compiling corpus from the Event Registry global media monitoring system. This corpus focuses on information spreading in three domains: sports (i.e. the FIFA World Cup), natural disasters (i.e. earthquakes), and climate change (i.e. global warming). It explains the process of comparing subsequent articles via cosine similarity and applying a threshold to classify them into three classes: "Information-Propagated", "Unsure" and "Information-not-Propagated". Section 2.2 deals with the analysis of information spreading in mono- and cross-lingual settings. It presents how force-directed graphs help to visualize pairs of news articles with cosine similarity. It also explains a visualization that has been developed to perform multi-lingual temporal information cascading.

Chapter 3 deals with the applications of barrier detection and classifications. It consists of four separate thematic sections. Section 3.1 deals with the multi-class classification task of news events. It explains how events can be categorized. It also presents the explanation of simple classification methods and neural networks along with a different set of features including character and word n-grams and pre-trained Glove embeddings. Regarding clustering of newspapers, it presents a novel methodology for clustering the daily read newspapers based on political and economic characteristics using different similarity

mechanisms. It is based on the trend lines of Wikipedia-concepts, and a simple count of DMOZ-categories (Directory Mozilla) and Wikipedia-concepts related to the news events. Section 3.2 deals with enhanced topic modeling. It shows the comparison of the novel approach with the previous approach. Related work on Spatio-temporal analysis along with COVID-19 data is explained in it as well. Section 3.3 deals with different textual features. The importance of the varying nature of features has been explained in it. It also explains why understanding news reporting is necessary at regional levels.

Chapter 4 deals with the profiling of news spreading barriers. A novel methodology based on a hypothesis is explained. It includes metadata-based semantic annotation, news headlines, common sense inferences and sentiments (see Section 4.1). The research questions are also explained. Finally the results are presented in the paper Profiling the news spreading barriers using news headline in *Scientific Contributions 1.4-3.*

Chapter 5 concludes the thesis by looking back at the main scientific contributions and clearly stating objectives and direction for future work and research on the topic of information spreading barriers in news.

Chapter 2

Analysis of News Spreading Barriers

There is a lack of research studies that examine, identify and uncover the reasons for barriers to information spreading. Moreover, there is limited availability of news datasets containing news text and metadata including time, place, source, and other relevant information. When a piece of information starts spreading, it implicitly raises questions such as: 1) How far does the information in the form of news reach out to the public? 2) Does the content of news remain the same or change to a certain extent? 3) Do cultural values impact the information especially when the same news will get translated into other languages?

We can help in assessing the significance of an event for a specific language by analyzing monolingual sets of news articles about it, which provides a base for comprehending how important an event is for a specific region. Cross-lingual information cascading, on the other side, enables us to identify the tendency and involvement of each language group in a specific event. Such factors influence the decision to publish news. News has a serious influence on decision-making in any country, and international news has an impact on international relationships and foreign policy as this can change public opinion, which has an impact on important decisions in liberal democracies [25].

The rest of this chapter is structured as follows: Section 2.1 presents the detail of a dataset with its annotation criteria. This dataset has been enhanced with other meta-information to perform a detailed analysis of previously defined barriers (see Section 1.2.1). Section 2.2 presents analysis of news spreading barriers.

2.1 Domain-specific Dataset

The influence of barriers on news spreading varies depending upon the domain. Since this topic is widely known all over the world and is pertinent to cross-cultural studies, numerous studies have been carried out in the past to better understand various aspects of information dissemination. The analytical results show that events spread in different ways depending on the domain; for example, news about natural catastrophes propagates more quickly than news about sports or global warming.

We have composed a corpus that focuses on analyzing information propagation in three different fields. We focused on a combination of rich- and low-resource European languages, in particular English, Portuguese, German, Spanish, and Slovene. The main objective of targeting three different types of events is to potentially analyze different propagation behaviors in our society. These events are sports (FIFA World Cup), natural disasters (earthquakes), and climate change (global warming). These three types of events are also chosen based on their popularity and diversity. A list of sub-events were observed from top websites related to the three events and we selected those which were the most popular

in the countries with the selected national languages. FIFA World Cup, earthquakes, and global warming were found to be the most prevalent, thus a dataset for each was collected. We have represented the cross-lingual news articles by mono-lingual Wikipedia concepts using the Wikifier service. Cosine similarity was calculated between tf-idf representation of news articles across all five languages. First, we excluded those articles which had scored 1.0, as they were considered a copy of the article. We then, for each article, chose an article that had the highest similarity score to it from the list of all articles. After performing this step, we had one similarity score for each article which shows the information spread to a certain extent (if > 0) or not (if 0). To decide about the class label and whether the information is spreading or not, we divided the scores into three intervals. The first is similarity $>= 0.7$, the second is $0.7 >$ similarity $>= 0.4$, and the third is similarity < 0.4. Articles that scored in the first interval were labeled as "Information-Propagated". The second interval was considered as unclear whether the information from the article propagate or not such articles was labeled as "Unsure". The lowest interval was considered as a signal for no propagation and labeled "Information-not-Propagated".

Similar data sources have been used previously but for understanding different research questions. The structure of global news coverage of disasters and its determinants has been revealed by using a large-scale news coverage dataset collected by the GDELT (Global Data on Events, Location, and Tone) project that monitors news media in over 100 languages from the whole world [1]. Significant variables in our hierarchical (mixed-effect) regression model, such as population, political stability, damage, and more, are well aligned with a series of previous research. The results show strong regionalism in news geography, highlighting the necessity of comprehensive datasets for the study of global news coverage [57]. A comparison between GDELT and Event Registry is performed, which monitors news articles worldwide and provides big data to researchers regarding scale, news sources, and news geography. The findings suggest significant differences in scale and news sources, but surprisingly, high similarity has been observed in news geography between the two datasets.

The resulted corpus was presented at Proceedings of the 23th International Multiconference Information Society SiKDD in Slovenia, Ljubljana, in 2020 [58] with a title *A Dataset for Information Spreading over the News* by Abdul Sittar, Dunja Mladenić, and Tomaž Erjavec. In this paper, there are two main contributions: 1) a novel methodology to collect a domain-specific corpus based on semantic similarity between news articles from a news repository, and 2) an annotated dataset encoding the level of information spreading from an article.

[1] https://www.gdeltproject.org/

A Dataset for Information Spreading over the News

Abdul Sittar
Jožef Stefan Institute
Ljubljana, Slovenia
abdul.sittar@ijs.si

Dunja Mladenić
Jožef Stefan Institute
Ljubljana, Slovenia
dunja.mladenic@ijs.si

Tomaž Erjavec
Jožef Stefan Institute
Ljubljana, Slovenia
tomaz.erjavec@ijs.si

ABSTRACT

Analysing the spread of information related to a specific event in the news has many potential applications. Consequently, various systems have been developed to facilitate the analysis of information spreading, such as detection of disease propagation and identification of the spreading of fake news through social media. The paper proposes a method for tracking information spread over news articles. It works by comparing subsequent articles via cosine similarity and applying a threshold to classify into three classes: "Information-Propagated", "Unsure" and "Information-not-Propagated". There are several open challenges in the process of discerning information propagation, among them the lack of resources for training and evaluation. This paper describes the process of compiling corpus from the Event Registry global media monitoring system. We focus on information spreading in three domains: sports (i.e. the FIFA World Cup), natural disasters (i.e. earthquakes), and climate change (i.e. global warming). This corpus is a valuable addition to currently available dataset to examine the spreading of information about various kind of events.

KEYWORDS

Datasets, Information propagation, News articles

1 INTRODUCTION

Information spreading has received significant attention due to its various market applications such as advertisement. did the information about a specific product reach to the public of a specific region? This could be one of the significant research questions. Research in this area considers influential factors in the process of information spreading such as the economic condition of a specific area related to how textual or visual content is helping to advertise a product. Information spreading analytics can also be used in shaping policies, e.g., in media companies to understand if there is a need to improve the content before publishing it. Health organizations may be interested to know the patterns of spreading of a cure for a certain disease. Environmental scientists are perhaps attentive to see whether spread of news about climate changes inside the country is similar to what is being reported internationally.

Domain-specific gaps in information spreading are ubiquitous, and may exist due to economic conditions, political factors, or linguistic, geographical, time-zone, cultural and other barriers. These factors potentially contribute to obstructing the flow of local as well as international news. We believe that there is a lack of research studies which examine, identify and uncover the reasons for barriers in information spreading. Additionally, there is limited availability of datasets containing news text and metadata including time, place, source and other relevant information.

When a piece of information starts spreading, it implicitly raises questions such as:

(1) How far does the information in the form of news reach out to the public?
(2) Does the content of news remain the same or changes to a certain extent?
(3) Do the cultural values impact the information especially when the same news will get translated in other languages?

This paper presents a corpus that focuses on information spreading over news and that hopes to answer some of the above questions (This corpus is published as an online resource at). We present the use of a news repository to produce a corpus and then analyze information propagation. We present a novel methodology for automatically assembling the corpus for this problem and validate it in three different domains. We focused on a combination of rich- and low resource European languages, in particular English, Portuguese, German, Spanish, and Slovene. Three different types of events are targeted in the data collection procedure to potentially involve different information spreading behaviors in our society. These events are sports (FIFA World Cup, 2,695 articles), natural disasters (earthquakes, 3,194 articles), and climate change (global warming, 1,945 articles). The three types of events were chosen based on their popularity and diversity. A list of sub-events was observed from top websites related to the three events and we selected those which were the most popular in the countries with the selected national languages. For sports, a list of countries with their national sports was fetched and then filtered for national language[1], [2]. Based on popularity, we selected the FIFA world cup. Similarly, for natural disasters, lists of natural disasters were collected by country taking the national language into account, for instance, for Slovenia we looked for this country in the natural disaster category on Wikipedia[3]. Earthquakes[4] and global warming[5] were found to be the most prevalent, thus a dataset for each was collected. Table 1 shows the selected events and other related events ordered by prevalence.

Table 1: List of events

Selected events	Other events (ordered by popularity)
Football	Basketball, Baseball, Boxing, Tennis, Cycling
Earthquake	Floods, Tsunamis, Landslides, Hurricane, Volcanic eruptions
Global warming	CO_2 emissions, Chemical consumption

The paper makes the following contributions to science:

(1) a novel methodology to collect a domain-specific corpus from news repository;
(2) semantic similarity between news articles;

Information Society 2020, 5–9 October 2020, Ljubljana, Slovenia

[1] http://www.quickgs.com/countries-and-their-national-sports/
[2] https://www.topendsports.com/
[3] https://en.wikipedia.org/wiki/Category:Natural_disasters_in_Slovenia
[4] https://en.wikipedia.org/wiki/List_of_earthquakes_in_2020
[5] https://www.theguardian.com/environment/2011/apr/21/countries-responsible-climate-change, [6]

(3) an annotated dataset encoding the level of information spreading from an article.

The rest of the paper is organized as follows: in Section 2 we discuss prior work about information spreading; in Section 3 we describe the data collection methodology; Section 4 describes semantic similarity and dataset annotation; and Section 5 gives the conclusions.

2 RELATED WORK

Information spreading is prevalent in our society. It plays a vital part in tasks that encompass the spreading of innovations [9], effects in marketing [6], and opinion spreading [4]. News spreading provides information to consumers that can be used for decision making and potentially contribute to shaping national and international policies. There are several types of media involved, such as print media, broadcast, and internet media. Internet is considered as a building block for connecting individuals worldwide, while news reflects current significant events for people [7]. Apart from news, online social media proved to be a remarkable alternative to support information spreading in an emergency [8, 5]. Social connection plays a vital role in news spreading. Especially the structure of network reflecting who is connected to whom, crucially increases the proportion of information spreading. Network structure analysis comes with a hypothesis related to the strength of the connections, namely that information will spread further in a situation where there exist many weak connections rather than clusters of strong [2].

While, in general, there are not many dataset that would help in modelling information spreading, there are some corpora for detecting the spreading of information about diseases [3] and fake news in social media [10]. There is currently no multilingual dataset of news articles for analysis of information propagation composed from a variety of event-centric information such as sports, natural disasters, and climate changes. This provides additional motivation for our work.

3 DATA COLLECTION METHODOLOGY

In order to collect news originating from different sources, in different languages, and targeting diverse events, we used Event Registry, a platform that identifies events by collecting related articles written in different languages from tens of thousands of news sources [9]. Using Event Registry APIs [7], we fetched a list of articles about each event in the following languages: English, Spanish, German, Portuguese, and Slovenian. Figure 1 shows the data collection process.

Each article was parsed from the JSON response and stored in CSV files. Each article was connected with the available list of relevant information such as the language of the article, event type, publisher, title, date, and time. Figure 2 shows the metadata of articles.

The number of collected articles in each domain varies considerably, and also varies across the languages within each domain. Table 2 shows statistics about each dataset.

4 SEMANTIC SIMILARITY BETWEEN NEWS ARTICLES

We have represented the cross-lingual news articles by monolingual (English) Wikipedia concepts using the Wikifier service[8].

[7]https://github.com/EventRegistry/event-registry-python/blob/master/eventregistry/examples/QueryArticlesExamples.py

[8]http://wikifier.org/info.html

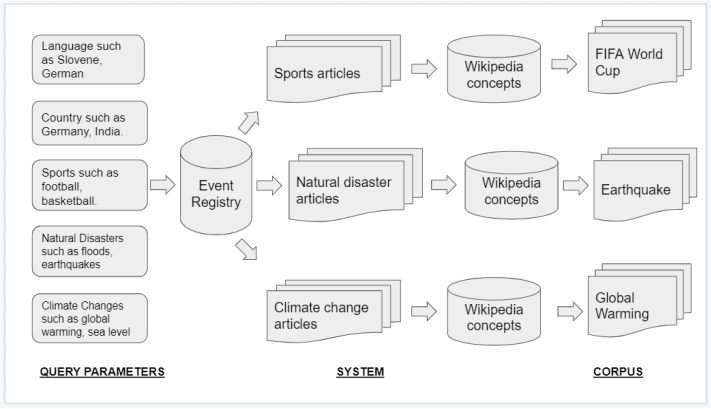


Figure 1: Data collection methodology

A	B	C	D	E	F	G	H	I
Language+I	Event	Weight	Class	Article Title	Publisher	Publishing Time	Website	Article URL
English10	FIFA World Cup	0.991	Information-Propagated	Football: Be	Channel Ne	2020-04-29T16:0	channelne	https://www
English100	FIFA World Cup	0.55	Unsure	Despite beat	OnlineNige	2020-04-28T11:0	news2.on	https://news
English1000	FIFA World Cup	1	Information-Propagated	Norman Hur	Borehamwo	2020-04-10T11:0	borehamv	https://www
English101	FIFA World Cup	0.195	Information-Not-Propagated	Qatar prepa	web4.inside	2020-04-28T10:0	web4.insi	https://www
English102	FIFA World Cup	1	Information-Propagated	Despite beat	Legit.ng	2020-04-28T09:4	legit.ng	https://www
English103	FIFA World Cup	0.199	Information-Not-Propagated	Lungu eulog	Zambia Dail	2020-04-28T09:0	daily-mail.	http://www.
English104	FIFA World Cup	1	Information-Propagated	100 General	My London	2020-04-28T08:5	mylondon	https://www
English105	FIFA World Cup	0.272	Information-Not-Propagated	Nigeria: Oge	allAfrica	2020-04-28T07:4	allafrica.c	https://allafr
English106	FIFA World Cup	0.304	Information-Not-Propagated	What really	Coventry Te	2020-04-28T07:1	coventryt	https://www
English107	FIFA World Cup	0.331	Information-Not-Propagated	From Abdelg	The Nation	2020-04-28T06:4	thenation	https://www
English11	FIFA World Cup	0.906	Information-Propagated	Beckenbaue	Business St	2020-04-29T15:5	business-s	https://www
English110	FIFA World Cup	0.232	Information-Not-Propagated	Analysts' Co	Vancouver	2020-04-27T23:5	whitecaps	https://www
English111	FIFA World Cup	1	Information-Propagated	Indian footb	Scroll	2020-04-27T23:1	scroll.in	https://scrol
English112	FIFA World Cup	0.369	Information-Not-Propagated	Taggart's th	SBS Austral	2020-04-27T22:2	theworldg	https://thew
English113	FIFA World Cup	0.257	Information-Not-Propagated	VAN DIEST:	Toronto Su	2020-04-27T22:2	torontosu	https://toro
English114	FIFA World Cup	0.3	Information-Not-Propagated	Ronaldinho	SBS Austral	2020-04-27T22:1	theworldg	https://thew
English115	FIFA World Cup	0.379	Information-Not-Propagated	Liverpool co	Paisley Gat	2020-04-27T19:1	paisleygat	https://paisl
English116	FIFA World Cup	0.245	Information-Not-Propagated	Manchester	TODAY	2020-04-27T17:2	today.ng	https://www
English117	FIFA World Cup	0.331	Information-Not-Propagated	East Bengal	Indian Expr	2020-04-27T17:2	indianexp	https://india
English118	FIFA World Cup	0.254	Information-Not-Propagated	East Bengal	Firstpost	2020-04-27T17:1	firstpost.c	https://www
English119	FIFA World Cup	0.859	Information-Propagated	General kno	Radio Time	2020-04-27T16:4	radiotime	https://www
English12	FIFA World Cup	1	Information-Propagated	Beckenbaue	timesofmal	2020-04-29T15:4	timesofma	https://time
English120	FIFA World Cup	0.841	Information-Propagated	Argentine st	Legit.ng	2020-04-27T16:3	legit.ng	https://www

Figure 2: Articles with metadata

Table 2: Statistics about dataset

Dataset	Domain	Event type	Articles per Language					Total Articles
			Eng	Spa	Ger	Slv	Por	
1	Sports	FIFA World Cup	983	762	711	10	216	2682
2	Natural Disaster	Earthquake	941	999	937	19	251	3147
3	Climate Changes	Global Warming	996	298	545	8	97	1944

This service uses a page-rank based method to identify a coherent set of relevant concepts from Wikipedia [1]. We retrieved a list of Wikipedia concepts for each article. After representing each article with a list of Wikipedia concepts, the tf-idf score was computed using the popular machine learning library Scikit-Learn[9]. Using the same library, cosine similarity was calculated between tf-idf representation of news articles across all five languages. In the process of computing similarity between the articles, for each article we calculated its cosine similarity to all other articles and stored the results in a CSV file. The results were then sorted based on the publishing time of articles and we kept only the calculations of similarity to articles that are published later that the article in hands. Since we are interested in information propagation, we do not need to compare an article to those articles which have been published before it. As a result, we had a multiple similarity score for each article where each score show the similarity with other articles. Cosine similarity varies between zero and one, zero meaning no similarity and one meaning maximum similarity, i.e., a duplicate article.

[9]https://scikit-learn.org/stable/

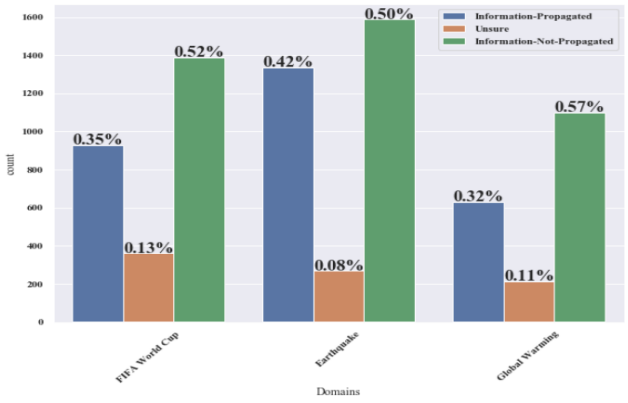


Figure 3: Class distribution for all domains

4.1 Dataset annotations

The results of the semantic similarity calculation were in the form of a table where rows shown the list of articles and columns shown the corresponding similarity score in the range 0..1 with all the other articles. This similarity score was calculated using cosine between TF-IDF representation of news articles (See Section **??**). First, we excluded those articles which had scored 1.0, as they were considered as a copy of the article. We then, for each article, chose an article which had the highest similarity score to it from the list of all articles. After performing this step, we had one similarity score for each article which shows either that the information spread to a certain extent (if >0) or not (if 0). To decide about the class label whether the information is spreading or not, we divided the scores into three intervals. The first is Similarity $\geq$ 0.7, the second is 0.7 > Similarity $\geq$ 0.4, and the third is Similarity < 0.4. Articles that have scores in the first interval were labeled as "Information-Propagated". The second interval was considered as unclear whether the information from the article propagated or not such articles were labeled as "Unsure". The lowest interval was considered as a signal for no propagation and labeled "Information-not-Propagated". For instance, low similarity can be of an article about a sports ground which mentions the population of the city and another article that discusses the population itself. We have manually examined concepts of articles in each class. Figure 3 shows the distribution of class labels in FIFA World Cup, Earthquake, and Global Warming dataset respectively.

4.2 Evaluation of dataset

Each article was annotated with a label based upon the similarity score threshold of each article with other articles (See Section 4.1). For evaluation of the dataset we have checked the content of the corresponding articles which were responsible for a specific class label. We performed the evaluation of labelling by manually inspecting a subset of pairs of articles. If a pair, for instance, were labelled as "Information-Propagated" then two articles should have text discussing more or less the same event, both in mono- and cross-lingual settings.

We have randomly chosen 10 articles with their corresponding articles considering all languages in each class and in each dataset. In this way, we have manually checked 180 articles. Table 3 shows these pairs of articles for evaluation in each dataset. We scanned each article manually for all languages, using Google Translator for Portuguese, German, Slovene and Spanish to translate them into English.

Table 3: Selected articles for evaluation

Domains	Percentage of correctly labelled pairs
Global Warming	100%
Earthquake	93%
FIFA World Cup	100 %

Evaluation results shown that the annotation was significantly related to information spreading. Articles in the "Information-Propagated" class show that most articles were an exact or paraphrased copy of each other, with some articles published within few hours after each other. Articles in the "Unsure" class were typically also relevant to the event but involved extra and different discussions. Lastly, in the third class "Information-Not-Propagated", articles involved only keywords related to event but discussion was about other topics. Moreover, here the gap in the publishing time was quite large.

5 CONCLUSIONS

This paper proposed a methodology and explained the process of data collection from a news repository to provide a corpus for event-centric information propagation between news articles. This corpus covers three domains and each dataset corresponds to one event type (FIFA World Cup, Earthquake, and Global Warming). The corpus is available to others for the evaluation of techniques for information spreading as it allows the analysis of cross-lingual news articles published by different publishers located geographically in different places.

In the future, we plan to add more attributes to each dataset. For instance, for now, we only know the publisher of a news article but in the future, we would like to include the publisher profile and the economic condition of a country from where the information is published. Also, we plan to apply and evaluate different techniques to analysis information propagation barriers.

6 ACKNOWLEDGEMENTS

This work was supported by the Slovenian Research Agency and the project leading to this publication has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 812997.

REFERENCES

[1] Janez Brank, Gregor Leban, and Marko Grobelnik. 2017. Annotating documents with relevant wikipedia concepts. In *Proceedings of Slovenian KDD Conference on Data Mining and Data Warehouses (SiKDD)*.

[2] Damon Centola. 2010. The spread of behavior in an online social network experiment. *science*, 329, 5996, 1194–1197.

[3] Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Covid-19: the first public coronavirus twitter dataset. *arXiv preprint arXiv:2003.07372*.

[4] David Liben-Nowell and Jon Kleinberg. 2008. Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the national academy of sciences*, 105, 12, 4633–4638.

[5] Kees Nieuwenhuis. 2007. Information systems for crisis response and management. In *International Workshop on Mobile Information Technology for Emergency Response*. Springer, 1–8.

[6] Everett M Rogers. 2010. *Diffusion of innovations*. Simon and Schuster.

[7] Sandeep Suntwal, Susan Brown, and Mark Patton. 2020. How does information spread? an exploratory study of true and fake news. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*.

[8] Satish V Ukkusuri, Xianyuan Zhan, Arif Mohaimin Sadri, and Qing Ye. 2014. Use of social media data to explore crisis informatics: study of 2013 oklahoma tornado. *Transportation Research Record*, 2459, 1, 110–118.

[9] Duncan J Watts and Peter Sheridan Dodds. 2007. Influentials, networks, and public opinion formation. *Journal of consumer research*, 34, 4, 441–458.

[10] Zilong Zhao, Jichang Zhao, Yukie Sano, Orr Levy, Hideki Takayasu, Misako Takayasu, Daqing Li, Junjie Wu, and Shlomo Havlin. 2020. Fake news propagates differently from real news even at early stages of spreading. *EPJ Data Science*, 9, 1, 7.

2.2 Propagation Analysis Across Economic, Time Zone, Geographic, Political, and Cultural Barriers

The details about the different types of barriers are explained in Chapter 1. To analyze the news spreading barriers we conduct a study. Our research hypothesis states that depending on the nature of an event, there will be variations in information spreading behavior across the observed barriers. We test our hypothesis on three types of events that are distinct: an earthquake within the natural disasters domain, the FIFA World Cup within the sports domain, and global warming within the climate change domain. This domain-specific dataset is explained in the Section 2.1. To aid understanding of the influence of different barriers on information spreading, this article sets four research questions:

- Q1: What are the properties (size and ratio) of cascading chains in events of different domains?
- Q2: Do the different information cascading chains have any relation to each other?
- Q3: How do economic, geographical, time zone, and cultural values influence news spreading in events of different domains?
- Q4: What is the correlation of news spreading of events in different domains to the political alignment of the news publishers?

This section presents the results of the paper titled *Analysis of information cascading and propagation barriers across distinctive news events* that was published in *Journal of Intelligent Information Systems* in 2021 [59]. This paper was authored by Abdul Sittar, Dunja Mladenić, and Marko Grobelnik. There are four main contributions of this study. Firstly, information cascading theory has been adopted to event-centric news analysis. Secondly, it provides new insights into the phenomenon of information propagation in news for different domains. Thirdly, it proved that with the incorporation of online sources, we can retrieve and enhance data related to information barriers. Lastly, it provides the analysis of the influence of multiple barriers on information propagation across distinctive news events.

The idea behind the above-mentioned research questions is to find the influences of different types of barriers on the three different domains. The properties of cascading chains can tell us the relation between the time and size of cascading chains. It further answers which events lasts over a longer period with large communities across different languages. It also explains the news events which spread more than other events. For instance, does news related to natural disasters spread more than sports or not? Analysis of barriers explains which barrier influences news spreading related to different barriers. For instance, does the news spreading influenced by the scope of events across different languages? Is the geographical size of a region have a direct relation with news spreading? Similarly, do the different cultures, economies, and time zone have any relation to the news spreading? Lastly, one of the research questions answers whether the political alignment of news publishers has any correlations while spreading the news about different events.

Journal of Intelligent Information Systems (2022) 58:119–152
https://doi.org/10.1007/s10844-021-00654-9

Analysis of information cascading and propagation barriers across distinctive news events

Abdul Sittar[1] · **Dunja Mladenić**[1] · **Marko Grobelnik**[1]

Received: 11 January 2021 / Revised: 30 June 2021 / Accepted: 30 June 2021 /
Published online: 31 August 2021

Abstract

News reporting, on events that occur in our society, can have different styles and structures, as well as different dynamics of news spreading over time. News publishers have the potential to spread their news and reach out to a large number of readers worldwide. In this paper we would like to understand how well they are doing it and which kind of obstacles the news may encounter when spreading. The news to be spread wider cross multiple barriers such as linguistic (the most evident one, as they get published in other natural languages), economic, geographical, political, time zone, and cultural barriers. Observing potential differences between spreading of news on different events published by multiple publishers can bring insights into what may influence the differences in the spreading patterns. There are multiple reasons, possibly many hidden, influencing the speed and geographical spread of news. This paper studies information cascading and propagation barriers, applying the proposed methodology on three distinctive kinds of events: Global Warming, earthquakes, and FIFA World Cup. Our findings suggest that 1) the scope of a specific event significantly effects the news spreading across languages, 2) geographical size of a news publisher's country is directly proportional to the number of publishers and articles reporting on the same information, 3) countries with shorter time-zone differences and similar cultures tend to propagate news between each other, 4) news related to Global Warming comes across economic barriers more smoothly than news related to FIFA World Cup and earthquakes and 5) events which may in some way involve political benefits are mostly published by those publishers which are not politically neutral.

Keywords Information spreading · Cultural barrier · Political barrier · Geographical barrier · Economic barrier · Time-zone barrier · Linguistic barrier

✉ Abdul Sittar
abdul.sittar@ijs.si

Dunja Mladenić
dunja.mladenic@ijs.si

Marko Grobelnik
marko.grobelnik@ijs.si

[1] Jozef Stefan Institute, Ljubljana, Slovenia

1 Introduction

News spreading is one of the most effective mechanisms for spreading information. Mainly there are two different ways of obtaining information: 1) observing or engaging in the event in person, for example, listening to a presidential speech, and 2) coming across information that is propagated from multiple channels or publishers which potentially influence the perception of the event they are reporting on and thus can consequently change the recipient's point of view.

Due to globalization, many events from different areas are internationally relevant. Representation of cross-lingual information about an event should be in a unique format and relevant context as this helps people to understand the entire story of current regional and international events that belong to diverse cultures. Information spreading via news is related to information cascading, where publishers decide to write on an event that is already published by another publisher. The result is subsequent news reporting on the same event, starting from the root news article to the last news article on the same event. The concept is commonly used in social media to find a set of subsequent re-shares starting from the root user (Hong et al., 2017).

By analyzing monolingual sets of news articles about an event, we can help in estimating the importance of the event for a specific language which further provides a basis for understanding to what extent an event is important for a country or a region. However, cross-lingual information cascading enables us to find the tendency and interest of each language group for a specific event. The decision on publishing news relies on some certain factors. News has a major impact on decision making for any country, while international news has an impact in terms of international relations and foreign policy as they can change public opinion, which leads towards impacting important decisions for democratic countries (Wu, 1998). In our contemporary society, international news about different events led us to investigate the reasons why news regarding specific events either spread or do not spread to certain geographic areas. Media focuses on specific foreign and regional events based on some certain factors. For instance, spreading of events may tilt toward developed countries such as United States, the United Kingdom, or Russia. Moreover, it may be due to geographical juxtaposition (latitude, longitude) of countries (Wilke et al., 2012). There is a great deal of negotiation between political actors and journalists in news production to enhance their influence on news coverage (Maurer & Beiler, 2018). Therefore political alignment of publishers can more or less impact their coverage of different events. We expect to observe this difference in reporting on sport events, natural disasters or climate changes. An important step which takes place prior to news spreading is news selection of foreign events. Multiple theoretical explanations for this have been presented in the pasts (Wu, 2007; Chang & Lee, 1992).

Two of the determinants for news coverage are economic conditions and association between countries (Chang & Lee, 1992). Cultural values and differences also impact information selection, analysis, and propagation. For instance, if two countries are culturally more similar, there are more chances that there will be heavier news flow between them (Wu, 2007).

This study is focused on three popular events in three different domains: sports, climate change, and natural disasters. The rationals behind selecting these domains is that they are expected to have different influence on the public and differ in information spreading. Natural disasters such as floods, earthquakes, and tsunami waves are unfortunate events caused

by the natural and geological processes of Earth. One would expect that news regarding natural disasters is mostly objective. Climate change, such as Global Warming and pollution, is a very controversial topic, with political interests of different actors. Thus the reporting is expected to be selective and biased. Sports news can be considered quite political in nature involving prediction and claims speculating about the results of a game.

This article makes the following contributions:

- Information cascading theory has been adopted to event-centric news analysis.
- We provide new insights into the phenomenon of information propagation in news for different domains.
- With the incorporation of online sources, we are able to retrieve and enhance data related to information barriers.
- We provide the analysis of the influence of multiple barriers on information propagation across distinctive news events.

The remainder of this paper is structured as follows: Section 2 provides related work on information spreading barriers, and an analysis of news events in different domains. In Section 3, we provide details about our data sets and data enhancement. After elaborating on the research methodology in Section 4, the experiments are described in Section 5. Section 6 provides a brief discussion about event-related findings and the corresponding results. Finally, Section 7 presents the conclusion and some ideas for future work.

1.1 Hypothesis and research questions

Our research hypothesis states that depending on the nature of an event, there will be variations in information spreading behavior across the observed barriers. We tested our hypothesis on three types of events which are distinct: an earthquake within the natural disaster domain, the FIFA World Cup within the sports domain, and Global Warming within the climate change domain. In order to aid understanding of the influence of different barriers on information spreading, this article set four research questions:

Q1: What are the properties (size and ratio) of cascading chains in events of different domains?

Q2: Do the different information cascading chains have any relation to each other?

Q3: How do economic, geographical, time zone, and cultural values influence news spreading in events of different domains?

Q4: What is the correlation of news spreading of events in different domains to the political alignment of the news publishers?

2 Related work

The concept of information spreading is a broad topic and has an enormous number of research dimensions. As this study focuses on information spreading and respective barriers, we review six different types of interconnected related works: information spreading and contrasting events, linguistic, economical, geographical and time zone differences as well as political and cultural barriers.

Information spreading and contrasting events As this topic is much known globally and pertinent to cross-cultural studies, various studies have been conducted in the past to understand various aspects of information spreading. Among them are understanding the key

features in sports-related information spreading (Alla et al., 2011), temporal aspects while modeling information spreading (Miritello et al., 2011), modeling information diffusion in online social networks (Kumar et al., 2020), and modeling the information dissemination under disaster (Cui et al., 2020). Moreover, our study focuses on modeling events in different domains (sports, natural disasters, and climate change) and understanding hidden patterns in the flow of information. Our preliminary results show that events have different spreading patterns depending on the domain; in particular, we notice the spreading of news related to natural disasters comes across fewer barriers than the news related to sports as well as news related to climate change.

Linguistic Barriers Much of the research on linguistic barriers has focused on understanding properties of information propagation such as speed, size, and structure. Some researchers focus on cross-lingual information diffusion to understand information cascading (Jin, 2017) in social networks. There are attractive options on social media such as share or retweet that have been used mostly to understand information cascading rather than to understand the semantic meaning of the text. Unlike within the social media domain, finding cross-lingual similarity in the context of news is one of the top priorities. There are numerous studies conducted on semantic textual similarity in the literature. The concept of measuring semantic textual similarity is based on estimating the semantic relatedness between two or multiple texts (Glavaš et al., 2018). Multiple techniques have been developed to estimate cross-lingual semantic similarity of texts achieving satisfactory results, such as word-to-word translation (Vulic & Moens, 2014), dictionary based translations (Krajewski et al., 2016) and word embeddings methods for semantic similarity (Şenel et al., 2017). These approaches are mostly evaluated in the context of plagiarism detection and discourse analysis. In our case, we are focusing on understanding whether a piece of news is discussing something related to an event or not. Therefore, we consider concept-based similarity more pertinent and calculate similarity based on Wikipedia concepts[1]. The existing research on information cascading focuses on social networks and relies on social media features (e.g., retweets, share). Our method presents a new approach to cascading based on news spreading. Firstly, we observe information flow in mono- and cross-lingual settings. Secondly, we go beyond information flow based on textual similarity and show the flow of news related to events in different domains and in different languages from the point of view of temporal elements (e.g., monthly spread of news). Moreover, language as a part of cultural values and beliefs, largely emerges to have a worldwide significant role in news selection regarding different domains and their propagation at different times. Therefore, cross-lingual news can help in understanding differences between high-resource languages and low-resource languages in the process of news creation and propagation.

Economical Barriers According to news flow theories, multiple determinants impact international news spreading. The economic power of a country is one of the factors that influence news spreading. Moreover, economic variations have different influence for different events (e.g., protests, conflicts, disasters) (Segev, 2015). The magnitude of economic interactivity between countries can also impact the news flow (Wu, 2007). Economic growth/income level shows the economic condition of a country. Multiple organizations are working on generating prosperity and welfare indexes on a yearly basis. Among them, "The

[1] http://wikifier.org/info.html

Legatum Prosperity Index" and "Human Development Index" are popular[2],[3]. These prosperity indexes are already used to compare and draw prosperity relations within a country or between the countries to understand different aspects such as education, business infrastructure, and technology (Büyüksarıkulak & Kahramanoğlu, 2019). Our intention here is to compare news spreading in relation to the income level in the country of the news publisher.

Geographical and Time Zone Barriers Geographical representation of entities and events has been utilized extensively in the past to detect local, global, and critical events (Quezada et al., 2015; Wei et al., 2020; Watanabe et al., 2011; Andrews et al., 2016). It can help us to observe the proximity effects on corresponding research questions. It has been said that countries with close distance share culture and language up to a certain extent which can further reveal interesting facts about shared tendencies in information spreading (Segev & Hills, 2014; Segev, 2015). Our motivation for using geographical locations is to analyze the impact of geographical proximity on news spreading in different domains. In addition, to represent relative time, the time zone is an alternative to geographical difference between countries (Dagon et al., 2006). The publishing time of news articles has a strong association with time zones; therefore, our analysis will also take it into account.

Political Barriers News agencies tend to follow the national context in which journalists operate. One of the related examples is the SARS epidemic study which found that cross-national contextual values such as political and economic situations impact the news selection (Camaj, 2010). It will be true to say that fake news is produced based on many factors and it is surrounded by a paramount factor that is political effect. A great amount of work regarding fake news dwells on different strategies, while few studies considered political alignment to have a compelling effect on news spreading (Bakshy et al., 2015; Maurer & Beiler, 2018). Maurer and Beiler (2018) strongly proved it to be a major strategy in news agencies to control the news and change accordingly due to the involvement of journalists and political actors. One can expect that news on climate change and sports is more prone to be altered due to the political alignment of publishers than news on natural disasters. Our study incorporates political alignment of news publishers.

Cultural Barriers Countries that share a common culture are expected to have heavier news flow between them when reporting on similar events (Wu, 2007). There are many quantitative studies that found demographic, psychological, socio-cultural, source, system, and content-related aspects (Al-Samarraie et al., 2017). There are many models that have tried to explain cultural differences between societies. Hofstede's national culture dimensions (HNCD) have been widely used and cited in different disciplines (Khosrowjerdi et al., 2020; He & Lee, 2020). It has also been criticized by many researchers for reducing culture into dimensions. Originally, HNCD were comprised of four dimensions: power distance, individualism, uncertainty avoidance by individuals, and masculinity vs. femininity. It then further extended by two dimensions: long-term vs. short-term orientation and indulgent versus restrained. Scores for all dimensions for different countries have been presented in Table 8. Following is a description of these dimensions.

[2] http://hdr.undp.org/en/content/human-development-index-hdi
[3] https://www.prosperity.com/

Power distance index: shows the extent to which the power inequalities in society agree with each other, such as children and parents, youths and elders, students and teachers. A higher degree of the index indicates that power is distributed transparently, whereas a lower degree of the index is a sign that people question authority.

Uncertainty avoidance by individuals: relates to the degree to which a society is tolerant toward unknown, unusual, and novel situations. Higher score signifies that people prefer to choose stiff codes of behaviors, and rules etc., whereas a lower score indicates that society prefers to impose fewer regulations.

Non-individualistic cultures: Individualism is characterized by loose ties between individuals and a focus on privacy and personal integrity. Higher score means that society, people are more integrated into groups. On the contrary, a lower score shows an emphasis on individualism.

Masculinity vs. Femininity: This describes whether a society is dominated by masculine culture; in this case, the society is assertive and competitive with a high level of gender inequality. Alternatively, in the case where the society is oriented more towards femininity, there prevails a modest and caring values. Higher score means society is more dominated by masculinity and vice versa.

Long-term orientation: These societies are likely to emphasize savings and hardworking behaviors preparing for potential critical events in the future. On the contrary, short-term oriented societies do not focus on a futuristic approach. A higher score indicates that a society is more future oriented whereas lower score signifies that a society places a greater emphasis on ancestors and honors traditions.

Indulgence vs. Restraint: HNCD claims that indulgent societies focus more on personal fulfillment and jolly behavior but restrained societies focus on having personal wishes and happiness controlled by social norms. In this dimension, higher index indicates a higher degree of freedom in fulfilling the human desires in a society. On the other hand, lower index indicates that a society controls the gratification of needs.

There are fewer studies that take cultural values into account while recognizing the flow of news related to different events. Our study considers cultural values to be a significant factor in the spread of cross-lingual news across different countries. For illustration of this effect, we show the score of each dimension for 67 countries in Table 8.

3 Data description and pre-processing

This section presents the dataset we have collected for the purpose of our research, definitions of the basic terms (Information Propagated, Unsure, and Information not propagated), and enhancements of the dataset.

3.1 Data description

We have created a corpus that consist of 7773 news articles published between 2015 – 2020 in one of the five selected languages (English, Portuguese, German, Spanish and Slovenian) (Sittar et al., 2020). The cross-lingual news articles were represented as vectors of

Wikipedia concepts obtained by annotating articles using the Wikifier Service[4]. As in bag-of-words representation, we calculate tf-idf weight for each element of the vector, which in our case were Wikipedia concepts. To identify information propagation, we measure similarity between each article and all the articles published later than the article using the feed-forward mechanism. To measure similarity between two articles, we calculate cosine similarity of their vectors (similarity scores varies between 0 and 1. 0 means minimum and 1 means maximum similarity). The number of final pairs was 7817. We have published this in a paper entitled "A Dataset for Information Spreading over the News"[5] created with the aim to analyze propagation of information over news articles. All the articles in our dataset belong to one of the three kinds of events: (1) Global Warming, (2) FIFA World Cup, and (3) Earthquake (see Table 1). Cross-lingual similarity between the news articles was calculated based on Wikipedia concepts, as each article was represented by the associated Wikipedia concepts (Brank et al., 2017) as provided by the Event Registry systems (Leban et al., 2014), which we have utilised for collecting the articles. Table 1 shows statistics for each dataset. Based on the similarity score, pairs of articles are classified into one of the following classes:

Information Propagated: These articles are likely to discuss a similar event.

Unsure: There is uncertainty whether the information is propagating or not.

Information not Propagated: These articles are likely to involve a discussion about different events.

3.2 Dataset enhancement

To understand the effect of multiple barriers, we have enriched the dataset using information related to each barrier under observation, obtaining the data from different sources as described in the rest of this section. For most of the barriers (Economic, Geographical, Time Zone, Political, and Cultural), we have utilised information connected with the country of the publisher of the news articles.

3.2.1 Economical data

Deciphering information about economical barriers across different countries depends upon the economic profile of countries. We collected economic profiles of all the countries using the latest prosperity ranking 2019[6] and income levels (High, Upper Middle, Lower Middle, and Low-income level) using World Bank Country data[7]. Data along with the economic profiles of the countries of the news publishers can be found on the Zenodo repository (version 1.0.2)[8].

[4] http://wikifier.org/info.html
[5] https://zenodo.org/record/3950065
[6] https://www.prosperity.com/rankings
[7] https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups
[8] https://zenodo.org/record/4117411

Table 1 Statistics about dataset showing for each event type, the number of articles per language

Dataset	Event type	Articles per language					Total articles
1	FIFA World Cup	Eng	Spa	Ger	Slv	Por	2682
		983	762	711	10	216	
2	Earthquake	941	999	937	19	251	3147
3	Global Warming	996	298	545	8	97	1944

3.2.2 Linguistic data

For the linguistic barrier, the dataset already encompasses information on the natural language in which the article is written. As the articles are represented in a language neutral way - by Wikipedia concepts, we are able to compare articles in different languages and in the case of high similarity, we assume that the information is spreading from older to newer articles. Data regarding the linguistic analysis for all three events can be found on the Zenodo repository (version 1.0.2).

3.2.3 Political data

To understand the influence of political actors, we have obtained profiles of publishers from the Wikipedia info-box[9]. However, the profile of some of the publishers did not exist on Wikipedia (see Fig. 1), as a result of which the number of articles was reduced after excluding those with missing publishers' profiles. Table 2 shows the total number of publishers with profiles and a reduced list of articles. The purpose of this publishers' profile was to understand the political alignment of publishers for event-specific news articles. Data regarding political barrier analysis for all three events can be found on the Zenodo repository (version 1.0.2).

3.2.4 Geographical data

For geographical analysis of events, there was a need for the physical location of the publishers' headquarters or the geographical location of the country to which the publisher belongs. Publishers' profiles (see Section 3.2.3) help us to determine the name of the headquarters, which is then used to obtain the location (latitude/longitude) of a publisher's country. We grouped the newspapers according to their locations to see the distribution of news publishers over world map. We also grouped the news articles for each country to see their distribution over world map. Data regarding geographical barrier for all three events can be found in Zenodo repository.

3.2.5 Time zone data

Having obtained the location of a publisher from its headquarters (see Section 3.2.4) and knowing that the time zone is associated with the geographical location, we fetched the general time zone of all the countries for all the publishers. In case of a country having

[9] https://en.wikipedia.org/wiki/Help:Infobox

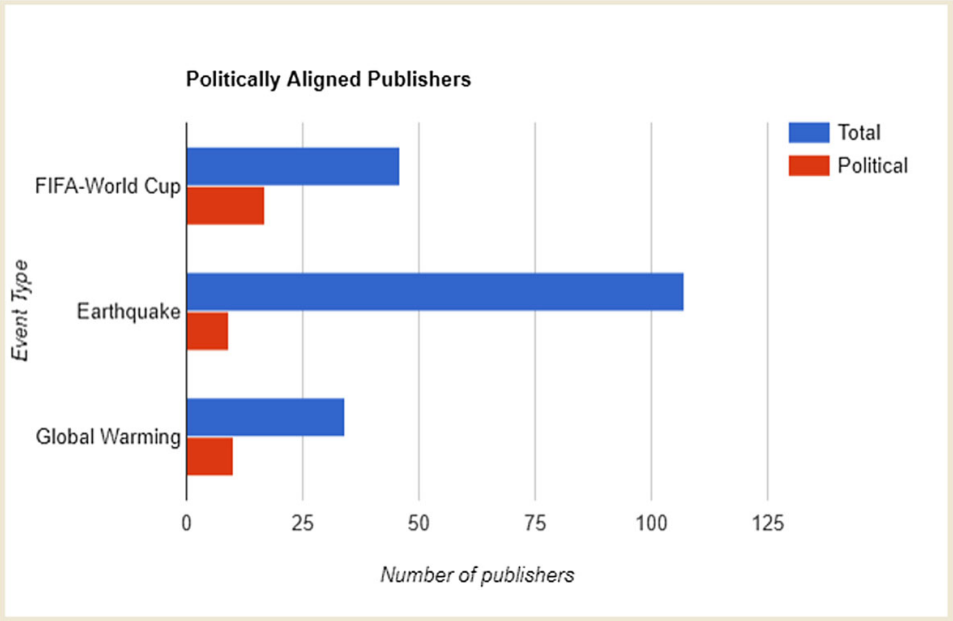


Fig. 1 Number of publishers with and without political alignment

multiple time zones, we selected one of them randomly. Data regarding time zone barrier for all three events can be found in the Zenodo repository.

3.2.6 Cultural data

To study the cultural barrier, we collected six dimensions showing cultural values of each country as suggested in Khosrowjerdi et al. (2020). Cultural dimensions have been assembled by an IBM study on different international populations and by different researchers (Hoftede et al., 2010). We extracted the list of countries along with these dimensions (shown in Table 8) from http://geerthofstede.com/research-and-vsm/dimension-datamatrix/. Table 8 shows the values of cultural dimensions. As only a few countries were missing from this list, we have excluded them from our data sets. Data regarding cultural barrier for all three events can be found in Zenodo repository.

4 Methodology

The presented research focuses on analysis of information propagation in news across different barriers in different domains. To this end, we propose a novel methodology consisting of several steps, as shown in Fig. 2.

Table 2 Statistics on available Wikipedia profiles for publishers and the number of corresponding news articles in our dataset

Domain	Publishers' profiles	Total articles
FIFA World Cup	515	324
Earthquake	341	406
Global Warming	399	226

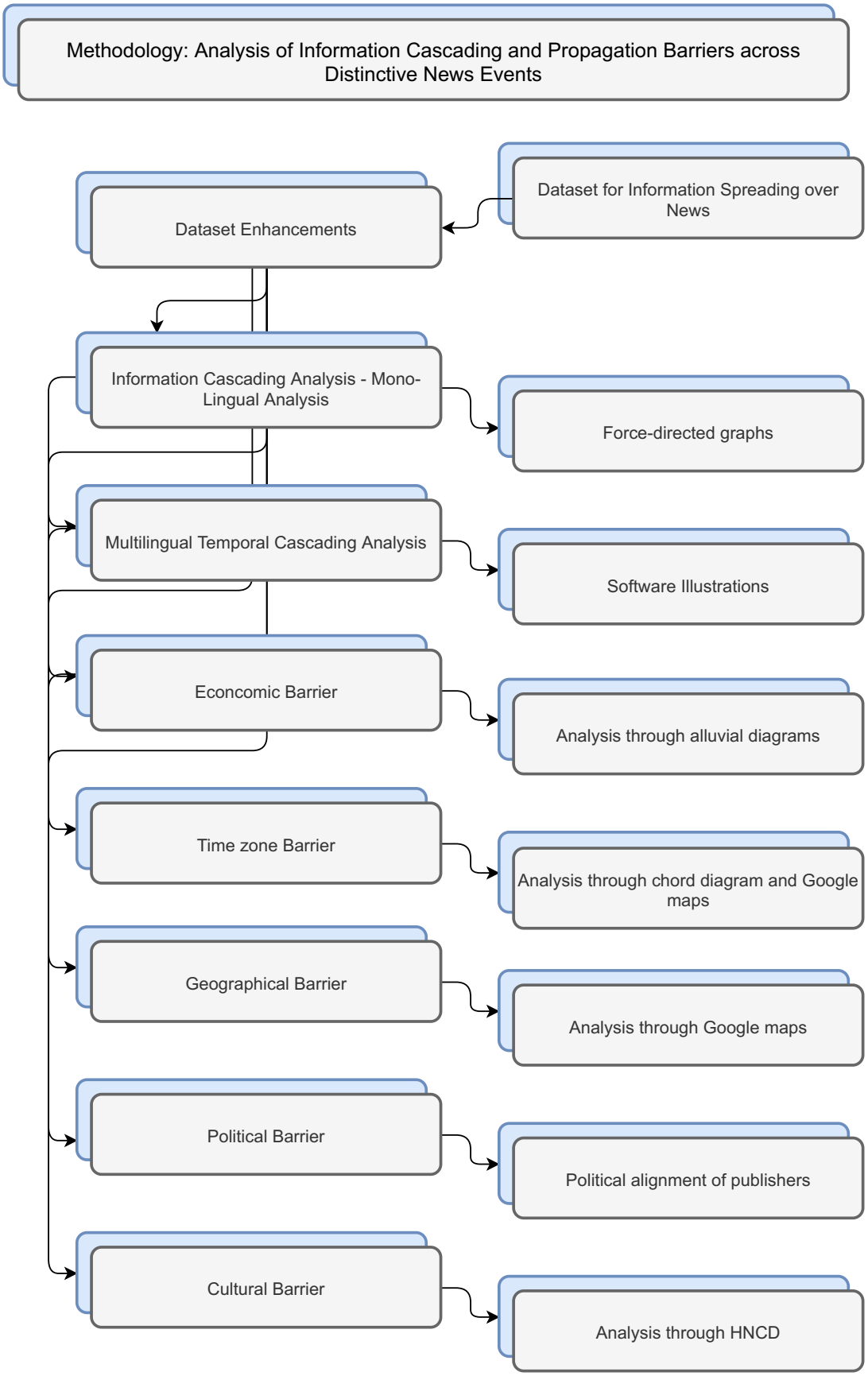


Fig. 2 Methodology to analyze information cascading and information propagation barriers. The dataset containing pairs of news articles exhibiting information spreading is the initial input. The data is enhanced to include background information on different barriers; analysis is performed to gain insights into information cascading and information spreading across different barriers

In the first step, we perform data enhancement to incorporate information related to different barriers as described in Section 3.2. Then we provide a distinct visualization for each of the observed barriers to help in the understanding of the direction and intensity of information spreading. The rationals behind each of the visualization methods are provided in the next part of this section. We perform different type of analysis mentioned in the proposed methodology on three types of events that are distinct in nature. We select sports, climate change, and natural disasters as three different domains with the rationale that each of these domains has a different influence on the general public. We utilise an existing dataset that is the result of our previous work (Sittar et al., 2020). For the purpose of the presented research, we perform some necessary enhancements of the original dataset in order to expand the focus beyond linguistic barriers.

The main aim of our study is to analyze multiple influences on the news spreading in three types of events (earthquakes, FIFA World Cup, and Global Warming) belonging to three different domains. We focus on information cascading and cross-lingual information spreading across geographical, economical, time zone, political and cultural barriers. For linguistic cascading and cross-lingual information spreading, the existing dataset already encompasses the needed information. For all other barriers, we collect additional information as described in Section 3.

For analysis of information spreading in mono-lingual settings, we perform network analysis using force-directed graphs (see Fig. 4). Force-directed graphs help to visualize connections between objects in a network and use to uncover relationship between the objects. Pairs of news articles with cosine similarity are passed as input to generate multiple chains of news articles as output (Similarity scores varies between 0 and 1, where 0 indicates no similarity and 1 indicates maximum similarity). For instance, if there exists two pairs with the news articles (A, B) and (B, C) respectively, then they will produce a chain/community. In the proposed visualization, we use a different color for each language. First we construct a network for each language for each event and then identify the chains/communities within a language. To detect the communities in each network, we use the Girvan-Newman algorithm based on Edge-Betweenness modularity. This algorithm takes a graph/network as input and provides possible communities along with different nodes. To analyze the communities, we looked into the text of news articles with communities of different size. The detailed implementation of the mono-lingual analysis using force directed graphs is available on Github[10].

For multi-lingual temporal information cascading, a visualization has been developed using Processing IDE to portray the temporal spread of a list of news articles about each event (see Fig. 6). Processing IDE[11] is a flexible software sketchbook that is used for prototyping new interfaces and services. It has been used in research laboratories of famous companies like Google and Intel for prototyping interfaces and services. It has also been used to visualize the data. For example, Yahoo! and Nokia used it for visualization and New York Times Company R&D Lab used it to visualize the way their news stories travel through social media. We used this software to develop a prototype to visualize the way the news propagates[12]. Similar to network analysis, we pass the pair of news articles as input and time as an additional feature where time provides the month of publishing. This approach enables us to identify the most influential languages about an event and results in

[10] https://github.com/abdulsittar/Mono-lingual-Analysis
[11] https://processing.org/overview/
[12] https://nytlabs.com/projects/cascade.html

an overview of the cascading structure. For instance, we have pairs of cross-lingual news articles along with the cosine similarity (see 3.1). This visualisation tool uses these pairs as input, draws spirals/circles of articles published within the same month and links the news articles that propagate information from one month to another. This enables finding how much information propagates across languages and over time. Figure 3 shows the temporal spreading of information in multi-lingual settings. The time interval captured by our datasets is 2015 – 2020 inclusive (6 years = 72 months). Each dot represents a news article. Each spiral/circle of dots indicates articles that were published within the same month. The connection between the two dots indicates that one article is spreading news to other article. We show two types of connections: within a month and within the following month. If two dots are connected within the same spiral/circle, this means that these two articles are propagating news from one article to other and were published within the same month. If a connection is made from one circle to another circle, it means that the two articles are spreading news and published in two consecutive months (see Fig. 3). The color of each dot represents the language of an article (Find the video file of prototype on Github[13]. The main purpose of the proposed multi-lingual temporal spreading visualisation is to show the overall prospect of information spreading on different events in different languages with time.

To understand the effect of economical barriers on information propagation, we perform the analysis based on economical categories of different countries of news publishers using the alluvial diagrams for each event (see Fig. 7). Alluvial diagrams are basically flow diagrams that help to discover change in large complex networks. We use the category of each news article (High-Income, Lower-Middle-Income, Upper-Middle-Income, or Low-Income) as input and generate an alluvial diagram for each event. This enables us to count the propagation among different types of economies and find associations between categories for each event.

To capture the similarities and differences across different countries regarding information propagation that is caused by different time zones, we construct a chord diagram for each event (see Fig. 9). A chord diagram is a graphical method to display the interrelationship between data points in a matrix. We pass the UTC time zone and UTC time difference for a pair of news articles as input and generate a chord diagram as output. This provides information regarding important time zones for each event.

To visualize information spreading paths from one country to another, we utilize Google maps (see Figs. 8, and 10). It enables to visualize entities on geographical landscapes. As input to the visualization service we provide the list of country names along with the number of articles originating from the country. On the generated maps, we draw links between the countries based on the intensity of information propagation. It enables to understand the geographical impact on news spreading for events. The map visualization further enables to obtain insights about economic, time zone and cultural relations among countries based on information propagation.

In the political barrier, we categorize the news publishers based on their political alignment (see Table 6). This enables to find the inclination of the news publishers altogether toward the events in different domains based on their political alignment. This also helps to know the political alignment of all the publishers that spread news for instance related to sports or natural disasters.

[13] https://github.com/abdulsittar/ProcesssingSketch/blob/master/FIFAWorldCup.mp4

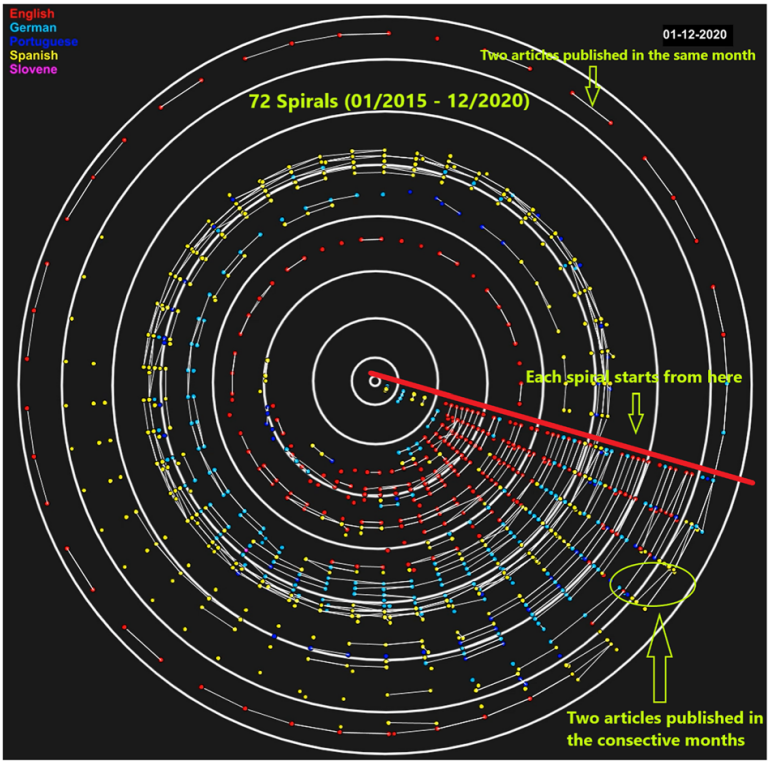


Fig. 3 Visual depiction of multi-lingual temporal propagation for FIFA World Cup enlarged from Fig. 6

Our proposed methodology uses two ways to approach the analysis: using visualization tools providing diagrams and maps to understand the influence of different barriers. This helps us in answering the last two research questions (see Section 1.1); the second way for approaching the analysis, which is new, provides a mechanism with which we are able to analyze the information cascading into mono-lingual settings and multilingual temporal cascading. This mechanism takes time into account and enables us to visualize the temporal spreading of information through news articles. This furthermore assists us in obtaining insight into the first two research questions).

5 Experiments

5.1 Propagation analysis through cascading

5.1.1 Mono-lingual analysis

For each event (i.e. Global Warming, earthquakes, and FIFA World Cup), we generated and analyzed networks of similar news articles. The network graph representing these three events is shown below (see Fig. 4). Nodes represent the articles and the color of the node uniquely identifies the language. Although the connected articles are clearly visible, the diagram is too dense to understand fully. Thus, as the next step in the analysis, we fetched important communities (small sub-networks with the largest path from one article to another) in order to analyze the flow of information. A community consists of nodes and edges and basically segregates a group of similar nodes from dissimilar groups (Raghavan et al., 2007). Figure 4 illustrates communities of the news articles that are spreading

Fig. 4 Overview of longest cascading chains in different languages related to events such as cascading chains in English, English and German for Earthquake, Global Warming and FIFA World Cup

information from one article to another for all three types of events. We used the Girvan-Newman algorithm based on Edge-Betweenness modularity for detecting communities in networks (Estrada, 2011). To analyze the communities we looked into the text of news articles within the small and large communities. Normally communities with a small number of articles (two or three) contain articles that are copy of each other or are reporting about the same event and the time difference between the articles is small (usually less than 1 hour). For instance, we randomly selected a community/chain of three news articles in English language related to FIFA World Cup domain.

Each of these news articles reports about anonymous attacks on 2022 Qatar World Cup and published within 10 minutes by three different publishers. A community with more articles could involve different discussions than a few articles, therefore we focused on large communities/chains.

5.1.2 Mono-lingual propagation - network analysis

For the FIFA World cup, total communities were 147. Each language English, German, Spanish, Portuguese, and Slovene had 47, 67, 26, 6, 0 chains respectively. For each of these languages, the largest communities consisted of 8, 24, 8, 5, 0 news articles whereas small communities consisted of 3 news articles (See Fig. 5). Time difference in the largest community (a chain of 24 news articles in German language) was of almost 3 months (31/08/2016 - 29/11/2016). In this community we see that it is related to the 2018 FIFA World Cup. Generally, by reading the articles manually, we found that most news articles report on the top three matches (Portugal vs. Spain, Egypt vs. Uruguay, and Morocco vs. Iran) minute-wise. In a group of 24 news articles, more than half of the articles were an exact copy of each other and the remaining articles were different only with regard to the varying amount of commentary text. For example, two articles were an exact copy of each other but the second one contained commentary up to the last 9 minutes of the game. Similarly, the other two

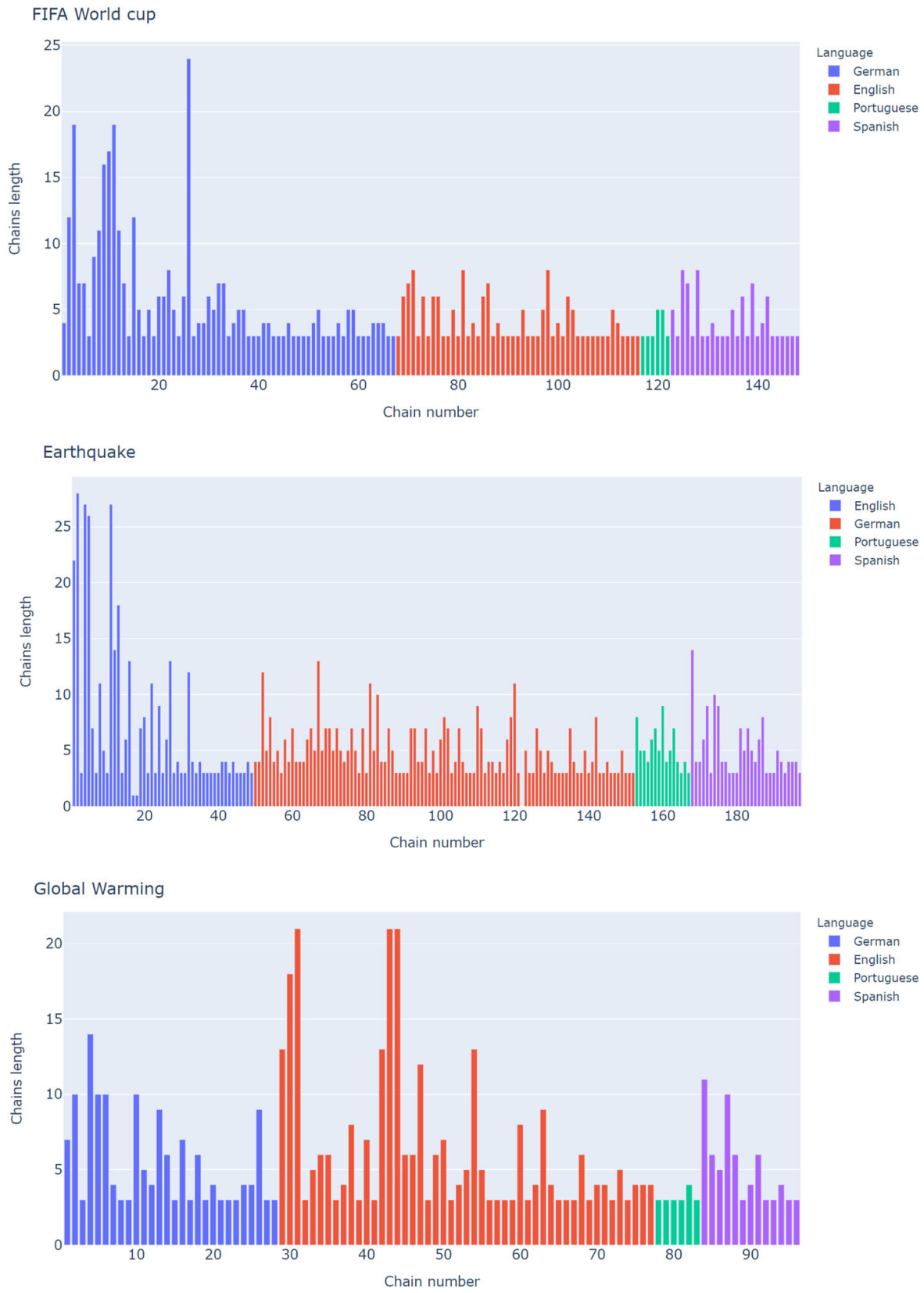


Fig. 5 Overview of length of chains for Earthquake, Global Warming and FIFA World Cup in different languages

articles were spreading the same information but the second one also contained extra text praising Cristiano Ronaldo.

On the other hand, with regard to earthquakes, there were 196 total chains. Each language English, German, Spanish, Portuguese, and Slovene had 49, 101, 31, 15, 0 chains respectively. For each of these languages, the largest chain consisted of 28, 12, 10, 9, 0 news articles whereas small communities consisted of 3 news articles (See Fig. 5). Time difference in the largest chain (a chain of 28 news articles in English language) was of almost 2.5 years (30/12/2017 - 29/04/2020). We see that the discussion within news articles was quite diversified. In general, we have seen that many of the news articles were related to COVID-19 but showed some pertinence to earthquakes. For instance, an article discussed that unlike cyclones or earthquakes, there are no natural disaster constraints in the COVID-19 pandemic situation. Similarly other articles enlightened related aspects such as NCCA (National Commission for Culture and Arts - Philippines) efforts to fight the spread of COVID-19, victims of earthquake, description about community-based disaster management plans in floods, earthquakes, and COVID-19 by German panchayats; the conversation about taking steps for COVID-19 as Tokyo has shockproof buildings for earthquakes; analysis of the budget of the Japanese government for massive earthquakes and tsunamis in earlier years; appraisal of Cuba's medical aid to other countries in time of earthquake and COVID-19 pandemics; and questions comparing which disaster (earthquake, bomb, tornado) could cause death toll. Secondly, out of a group of 28 articles, 8 news articles were found to have a discussion which was about Nasdaq insurance models and risk modeling services. Nasdaq is currently focusing on natural catastrophes with a model spanning to earthquakes, hurricanes, floods, and a number of other perils. Thirdly, four articles were related to earthquakes that occurred in California, New Zealand and Pakistan. Finally, a small number of articles involved unrelated discourses but were connected with earthquakes. For example, the first article was about the strange behavior of cats before an earthquake, second explains an earthquake felt by an English football team, and third article provides the detail about the drop in sales of Japanese auto manufacturer after a massive earthquake.

For Global Warming, there were 95 total chains. Each language English, German, Spanish, Portuguese, and Slovene had 48, 28, 13, 6, 0 chains respectively. For each of these languages, the largest chain consisted of 21, 14, 11, 4, 0 news articles whereas small communities consisted of 3 news articles (See Fig. 5). Time difference in the largest chain (a chain of 21 news articles in English language) was of almost 1 day (30/03/2018 - 31/03/2018). Contrary to the earthquakes, the largest community of news articles regarding climate changes had numerous discussions. Generally, every article was related to Global Warming but involved a variety of discussions ranging from the effect of climate change on COVID-19 to the design of buildings which are suitable for various climate changes. Very few articles were fully similar, with every item of news portraying a different point of view, described from a different perspective. Only three news articles were exactly similar that explain Green recovery is necessary to revive the global economy, whereas three articles comprised of three topics: first climate strike, limiting Global Warming by Donald Trump, and a speech about African actions to combat Global Warming in the United States. Since COVID-19 was affected by climate change, few articles appeared to have WHO related declarations and advice such as UN Chief Antonio Guerres address G20 regarding the fact that major developing and emerging economies together account for 80 percent global emissions, the fact that online streaming sites would not be allowed to upload a video that does not agree with UN intergovernmental panels position regarding exaggerated claims about global warming, and the fact that WHO-NASA affirmed that 2019 was the warmest year ever. The longest communities show the scope of an event for a specific language. Since English is currently the most widely used international language and events such as Global Warming and earthquakes incorporate news internationally, English appeared to have long cascading chains for both events. FIFA World Cup, being held mostly in Europe, has long cascading chains in the German language.

5.1.3 Multi-lingual temporal spreading

We observe information spreading via news articles that are published over some time period possibly crossing language barriers. We have built a visualization prototype to show this spreading over the period of six years focusing on information spreading from one language to another language (see Fig. 3)[14].

We have manually selected the articles which are spreading information to some other articles. Among them, we have chosen one chain for each event randomly to check the discourse of articles making the chain (see Fig. 3). Tables 3, 4 and 5 show the titles of the chains of articles with temporal information in each event. To understand the whole discourse, we manually read these chains.

Within the FIFA World Cup domain (see Fig. 6), in the chain of five articles (see Table 3), the first article reports on the FIFA World Cup player Norman Hunter, who was in the hospital due to COVID-19. This article also praised his past victories. The next article published on the same date that shows only upcoming matches schedules, and quiz about statistics of the matches and the names of famous players. This article was related to the FIFA World Cup but is entirely different from the first article. However, contrary to this, the remaining three articles were an exact copy of the first article. According to the dataset, the publishing time varied 3, 5, and 24 hours for the next three articles.

In case of climate changes (see Fig. 6), the chain of five news articles (see Table 4) which were published within 3 days appeared to have both similar and different types of discussions, for example the first and second article reports a general description of the Global Warming phenomenon and how the situation has worsened. Mainly both articles referred to the current analysis that is clear and removes any contradictions about the future threat of Global Warming. There was almost 13 hours of difference in publishing time. The next two articles were seen to have a discussion about Canadian jackets but a single word Global Warming appeared in that context. The publishing time difference were of approximately 3 days. The last article in this chain was published after 5 hours after the second last article. It explains the criticism of the measures that were taken by authorities in order to make climate policy and mentions a protest of students that was recorded and signed by professors to emphasize the need of change for future generations.

Finally, the chain of articles (see Fig. 6) present information spreading regarding an earthquake (see Table 5). It involves two types of discussions where the first article is written about a music festival that explains everything unrelated to earthquakes whereas the other three articles are reporting on an earthquake in Athens. The three articles had publishing time differences of about 10 minutes.

Looking at visualizations of multi-lingual temporal spreading of information, the significant findings are as follows:

1. Almost each year, FIFA World Cup had intensive information propagated in English and Spanish.
2. In case of Global Warming, looking at news articles in the period 2017 – 2020, information was propagated mostly in the English language.
3. For earthquakes, Spanish had the most number of articles compared to the other languages.

[14] https://github.com/abdulsittar/ProcesssingSketch

Table 3 An example of discourse along with publishing time taken from visual propagation (Fig. 6) related to FIFA World Cup

Publishing Time	Title
2019/11/30 - 18:57:00	Botond Barath and Johnny Russell learn potential group stage opponents at UEFA Euro 2020
2019/11/30 - 19:24:00	Euro 2020 Draw: Seeding and Schedule of Dates for Group Fixtures
2019/11/30 - 22:41:00	Roberto Mancini Says Italy 'Are Not the Favourites' After Euro 2020 Draw
2019/11/30 - 23:29:00	St. Petersburg fully ready to host Euro-2020 games – authorities - Russia News Now
2019/12/31 - 20:21:00	Spurned by Neighbors- Qatar Aims for Self-Sufficiency

Table 4 An example of discourse along with publishing time taken from visual propagation (Fig. 6) related to Global Warming

Publishing Time	Title
2019/01/10 - 19:31:00	Die Ozeane heizen sich schneller auf als gedacht
2019/01/11 - 07:46:00	Die Ozeane heizen sich schneller auf als gedacht - derStandard.at
2019/01/11 - 18:21:00	Gans dicke, Land Rover zum Anziehen
2019/01/14 - 08:15:00	Land Rover zum Anziehen,Land Rover zum Anziehen,Süddeutsche Zeitung
2019/01/14 - 13:57:00	Offener Brief: HNEE-Studierende fordern schnellstmöglichen Kohleausstieg

Table 5 An example of discourse along with publishing time taken from visual propagation (Fig. 6) related to Earthquake

Publishing time	Title
2019/06/04 - 17:54:00	Dub Inc e Horace Andy atuam em julho no festival Musa na praia de Carcavelos
2019/07/19 - 13:13:00	Terremoto magnitude 5.3 abala Atenas - ISTOÉ Independente
2019/07/19 - 13:26:00	Terremoto magnitude 5.3 abala Atenas - ISTOÉ DINHEIRO
2019/07/19 - 13:34:00	Terremoto de magnitude 5.3 abala Atenas,Terremoto de magnitude 5.3 abala Atenas - Alô Limeira!

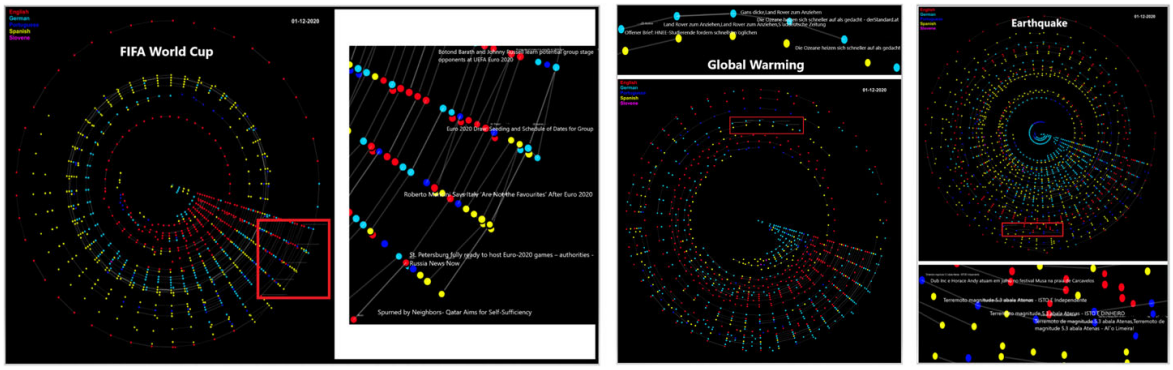


Fig. 6 Visual depiction of multi-lingual temporal propagation related to FIFA World Cup, Global Warming, and earthquake events

5.2 Propagation analysis across other barriers

5.2.1 Economical barriers

Propagation of the number of articles from one country to another country and the economic condition of countries has been analyzed. Figure 7 indicates the propagation across countries having different income levels (High-Income, Upper-Middle-Income, Lower-Middle-Income, and Low-Income). The number on the left side and on the right side of income level represents the total number of articles of all those countries that belong to a specific income level. The number in the middle shows the total amount of articles that has propagated the news between two different income levels. We represent each income level with a color to depict the difference between them. The streams of each income level show the propagation across different income levels.

For Global Warming, we observe that news from high-income countries propagated mostly to other high-income countries (blue area on the bottom graph in Fig. 7), with an exception of one news article which propagated to a low-income country (Nigeria). An interesting spreading involves low-income countries, e.g., Iran and Nigeria, where news articles from these countries propagated the news to high-income countries.

For the FIFA World Cup (see Fig. 7), the most frequent countries which appeared to have interesting facts were Germany, Spain, the United Kingdom, and India. News propagated from Germany and Spain to other European countries that have minor economical differences. However, Spain also had some news regarding the FIFA World Cup which propagated news to lower-ranked countries such as India, Bangladesh, Brazil, and South Africa. Most of the articles propagated the news from the United Kingdom to all other economically lower-ranked countries except one article from Germany. The United States appeared to have a mixed combination of articles propagating from and to the United States, Asian and European countries.

For earthquakes, some European countries such as Switzerland, Germany, Austria, France, Portugal and Brazil appeared to have a significant amount of propagation. Apart from many articles that propagated news to other European countries, one article appeared to propagate from Europe to Australia. Overall, there is information crossing economic barriers but we can observe that half of the news articles (yellow and green lines in Fig. 7) related to all events did not cross economical barriers.

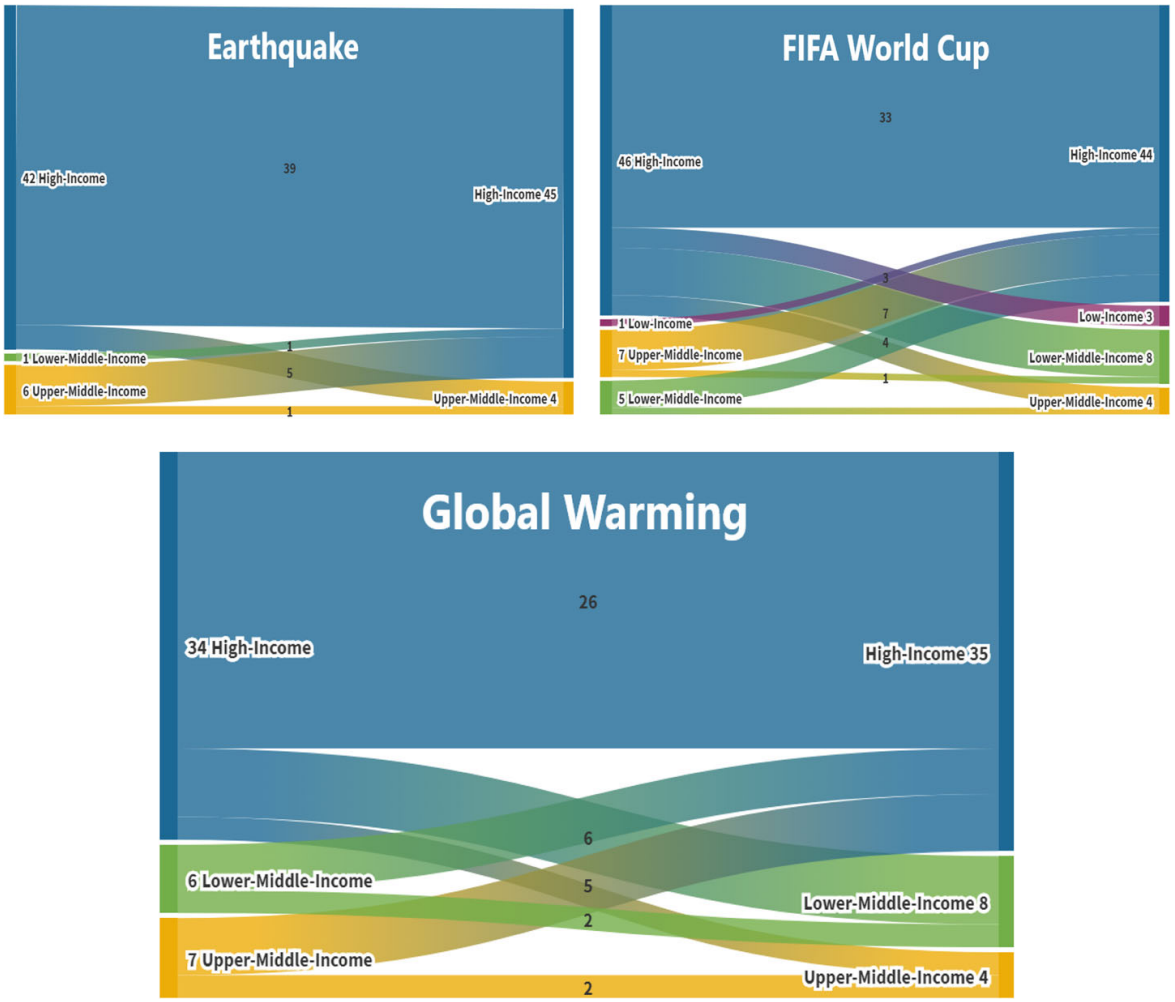


Fig. 7 Illustration of news spreading across different income levels (from top blue to bottom yellow: High-income, Lower-Middle-Income, Upper-Middle-Income and Low-income) related to earthquakes, FIFA World Cup and Global Warming events, respectively

5.2.2 Time zone barriers

To analyze spreading over geo-location, we mapped news articles on Google maps and drew links indicating the total number of propagations among countries to visualize the distance (see Fig. 8). Similar statistics have been shown using a chord diagram - the first three diagrams show propagation among time zones with specific time zone names, the next three with time zone values (see Fig. 9).

We observed several interesting and contrasting effects of different time zone on information propagation on all events. For earthquakes European countries such as Portugal, United Kingdom, Germany, Switzerland, and Brazil surfaced as the most popular in spreading news articles to other countries such as Taiwan, Canada, United States, Australia, and Israel. The difference in time zones among these countries lies between 3 - 13 hours (see Fig. 9). When we looked at FIFA World Cup news articles, the results were somewhat different. Mostly news propagated from the United Kingdom, United States, Spain, and India. The destined countries were mostly European such as Spain, Brazil, United Kingdom, as well as Asian countries such as Bangladesh, India, and Pakistan. The difference in time zone varied. Most of the articles propagated news to countries which had a time difference of 6 or 4 hours.

Fig. 8 Illustration of news propagation on Google Maps across different time zones related to earthquakes, FIFA World Cup and Global Warming

Lastly, we observed interesting factors related to Global Warming. India, Canada, and the United States appeared to have a larger time difference and more articles that propagated news to other countries. Countries that updated/received news from these countries were Saudi Arabia, Nigeria, the Philippines, Belgium, Brazil, Pakistan, Germany, and Switzerland.

5.2.3 Geographical barriers

To analyze the geographical impact on news propagation regarding distinctive events, we examined the distribution of publishers and articles over geographical distances among countries. For geographical influence, the distribution of publishers and articles has been illustrated on maps where green, yellow and red color depict the number of instances greater than 10, 50, and 100 respectively (see Fig. 10). In simple words, red, yellow, and green show the significant, medium, and smaller number of articles that are spreading among countries. Initially, our dataset had news articles published over different time periods. When we put them in temporal order based on propagation, the number of articles was reduced. The geographical distribution of publishers related to different events can show the significance of a country. Similarly, we can compare the distribution of news articles with an area or a country. Firstly, we categorized the articles' and publishers' distribution as high (green color), medium (yellow color), low (red color), and none (grey color). The low category indicates less than or equal to 10 publishers and articles whereas medium and high indicated less than or equal to 50 and 100 publishers, respectively. News publishers related to Global Warming are mostly from the United States, United Kingdom, and United Arab Emirates (145, 59, and 175 respectively) as can be seen in green (high category) in Fig. 10. All the countries (Ireland, Nigeria, Spain, Australia) lie in the medium category except ten countries with low category. There is only one country, the United States, which fell into the high category, with

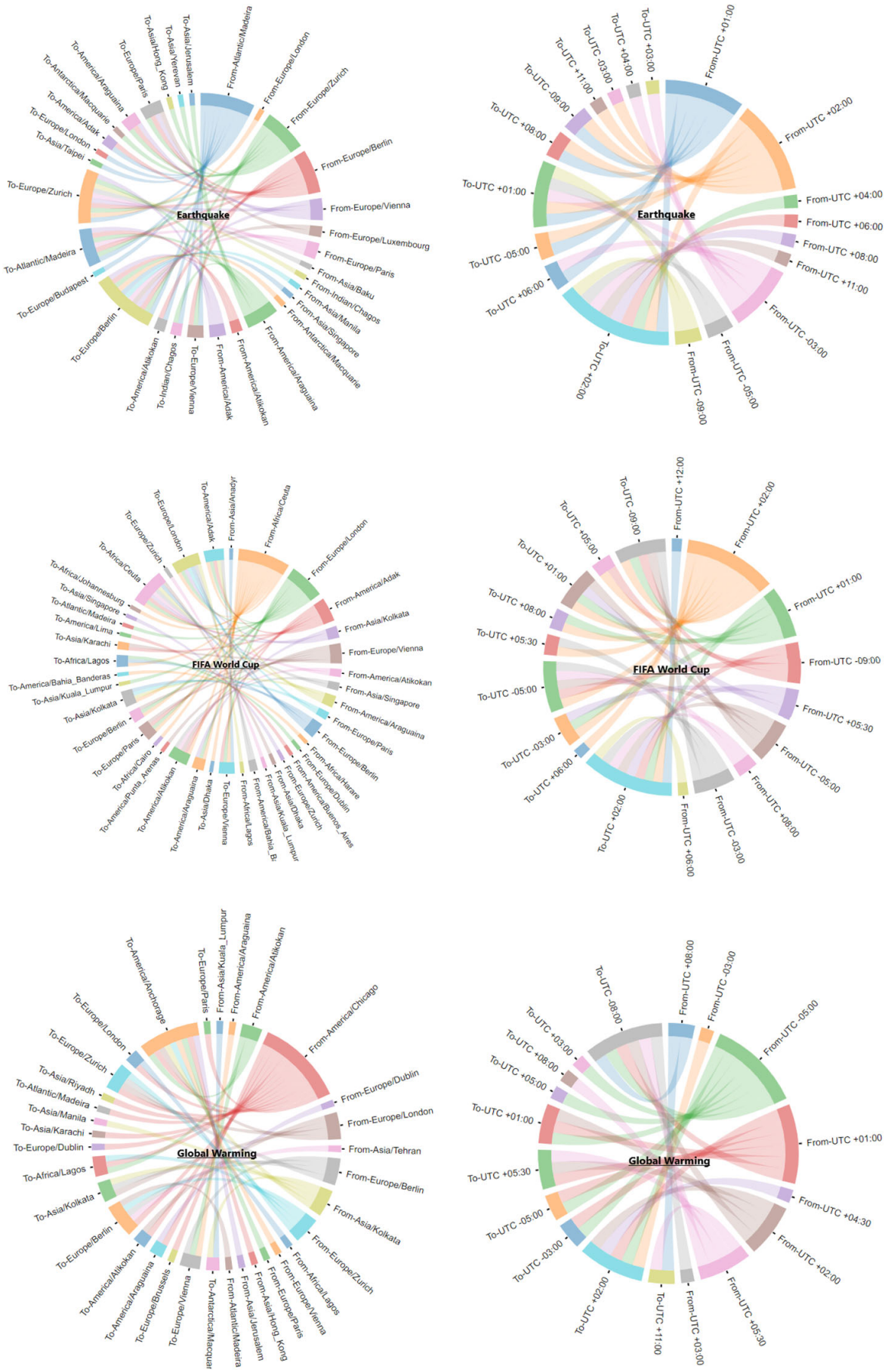


Fig. 9 Propagation depiction across different time zones illustrated with both a difference among time zone locations and UTC time

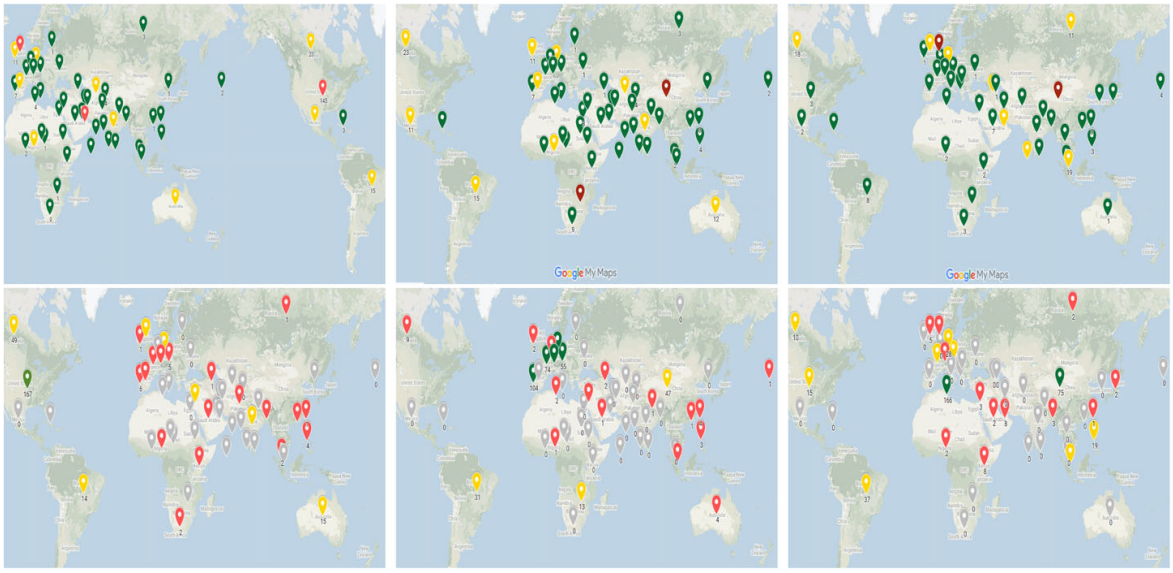


Fig. 10 An illustration of publishers' (first row) and articles' (second row) distribution over Google Maps for Global Warming, earthquakes, and FIFA World Cup events from left to right, respectively. Red, yellow and green colors show the significant, medium, and smaller amount of articles or news publishers

167 articles when considering the articles' distribution. 35 countries out of 55 happened to have less than two articles, whereas only 8 countries existed with the double-figure values.

Looking at the geographical distribution for the earthquake domain, we see that only two countries - the United Kingdom and the United States, had a high number of publishers (see Fig. 10). For low and medium categories, there were random distributions over the map, however the low category comprises of as many as double than those in the medium category. Figure 10 shows the distribution of news articles related to earthquakes. Most prominent pins are of grey color which means that they did not propagate news. There were five countries nearby pointing to the high category. These countries were Austria, France, Portugal, Switzerland, and Germany with 55, 74, 104, 185, and 181 articles, respectively. Countries in the medium category (the United Kingdom, the United States, and Brazil) also have huge differences geographically. Figure 10 previewed the publisher's distribution regarding the FIFA World Cup. Mostly news publishers belonged to UK and US with a count of 134 and 109. Eight countries (Australia, British Indian Ocean Territory, Canada, Germany, Pakistan, Portugal, Switzerland, and UAE) stood in the medium category with a count of 13, 26, 18, 17, 11, 19, 11, 19, respectively. Finally, related to the FIFA World Cup, most articles were published in Spain (166) and U.S. (75). Austria, Brazil, France, Nigeria, Taiwan, Canada, Germany, Portugal, had news articles in the medium category (with a count of 11, 37, 13, 19, 15, 10, 28, 16, news articles respectively). Overall, our studies depicted that countries which had large geographical area have more publishers for earthquakes than the FIFA World Cup and Global Warming.

5.2.4 Political barriers

Primarily, each publisher resides in one of the sixteen classes as displayed in Table 6. According to the table, events related to earthquakes were only published by those publishers which were politically neutral, progressive, and impartial. Publishers having anti-communist, pluralism, and new-left political ideals were more toward spreading news related to Global Warming. The other 10 categories show a presence in all events (see

Fig. 11). Since there is lack of data about political alignments of other publishers, therefore it is difficult to infer more interesting relations between political alignments and events.

5.2.5 Cultural barriers

Since the representation of culture has already been described with 6 dimensions (see Table 2), we found that most of the countries which appeared to spread news related to all events through their cultural dimensions were different from each other. For instance, Argentina is the only country that appeared to propagate news related to the FIFA World Cup. Argentina has a moderate score in each culture dimension in comparison to its long term orientation. It has a score as low as 20. We divide the list of cultural values into four categories: low (< 30), upper lower ($>= 30\& < 55$),upper higher ($>= 55\& < 80$), and high ($>= 80$) and compare the cultural values among countries and observe the spreading patterns in each event. Figure 12 illustrates information propagation between the countries with different power distances (PDI). We can see that countries with low category only propagate news to those countries that stand in the upper-lower and upper-higher categories (indicated with green lines), whereas countries with high category only propagate news to those with upper-higher category (indicated with red lines).

A list of the countries which draw attention to Global Warming include India, United States, Germany, Switzerland, Canada, Portugal, the United Kingdom, and Brazil whereas the list of countries underlined for the earthquake event are United States, Switzerland, Germany, United States, Portugal, Australia, Canada, Brazil, Israel, Taiwan, Armenia, Hong Kong, and the Philippines. Furthermore, for the FIFA World Cup, countries propagate more news articles were Singapore, United States, Russia, Brazil, United States, India, Spain, United States to Canada, Pakistan, Austria, India, Spain, Brazil, Canada, Austria, and France. For all three events, some of the countries share culture and also propagate

Table 6 Proclivity of news publishers with specific political class toward different events

No.	Classes of political alignment	Event type
1	Anti-communist	Global Warming
2	Catholic	Global Warming, Earthquake
3	Centrism	Global Warming, FIFA World Cup, Earthquake
4	Conservative	FIFA World Cup, Global Warming
5	Independent	FIFA World Cup, Global Warming
6	Liberalism	FIFA World Cup, Global Warming, Earthquake
7	New Left	Global Warming
8	Pluralism	Global Warming
9	Social Liberalism	FIFA World Cup, Global Warming, Earthquake
10	Left Wing	FIFA World Cup, Global Warming
11	Center Right	FIFA World Cup
12	Moderate	FIFA World Cup
13	Progressive	FIFA World Cup
14	Impartiality	Earthquake
15	Progressive	Earthquake
16	Neutral	Earthquake

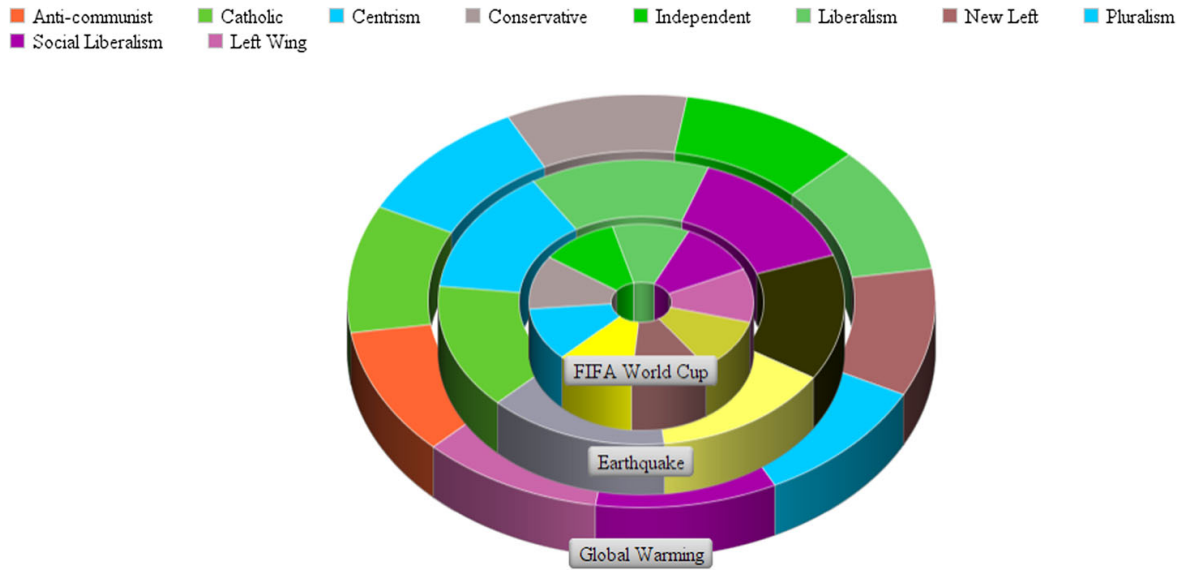


Fig. 11 Different events with political classes of news publishers

news. Table 7 shows the list of countries which propagate news and share one of the cultural dimensions (Table 8).

In short, for earthquakes, articles propagate news from Switzerland, Germany, France, Austria. We can see that they have more or less similar culture. There is a fewer number of articles that propagate news toward those countries which are quite different culture-wise such as Switzerland to Australia, Austria to Brazil, and Germany to Canada. For the FIFA World Cup, more articles propagate news to those countries which belong to different culture such as UK to Spain, US to Spain. For Global Warming, similar to the earthquake events, articles propagate news to those countries which are similar in culture.

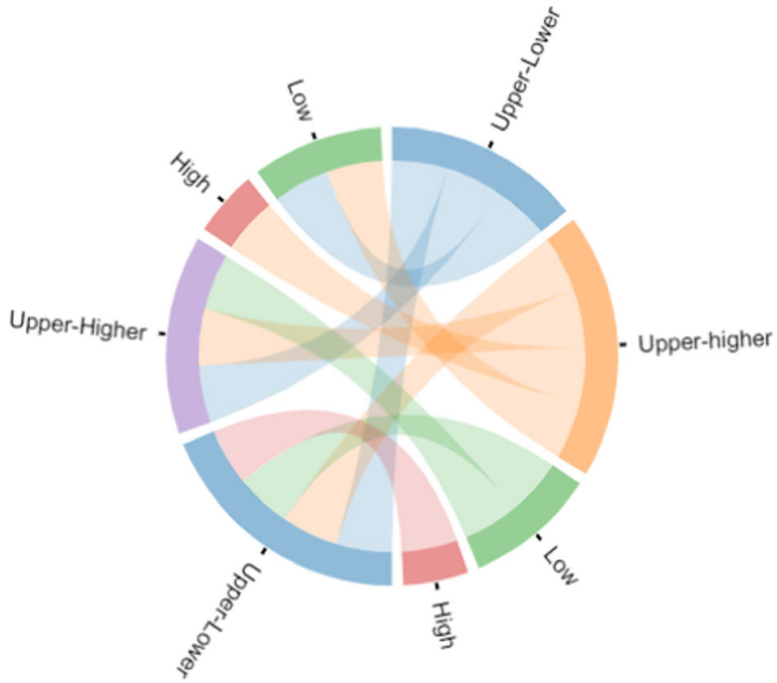


Fig. 12 News propagation visualization for one cultural dimension (PDI) in Global Warming enlarged from Fig. 13

Table 7 Table illustrates the list of countries that either share culture or not and propagates news articles

Event type	News propagation	Dultural dimension	Share culture
Global_Warming	India to the U.S., the U.S. to Brazil.	-	No
Global_Warming	the U.S. to France, Portugal to Germany.	-	No
Earthquake	Portugal to the U.S., Canada to Germany, Brazil to Israel.		No
Earthquake	Brazil to Armenia, Portugal to Canada.	-	No
Earthquake	Switzerland to Hong Kong, Brazil to Switzerland.	-	No
Earthquake	Brazil to Germany, and the Philippines to Germany.	-	No
Earthquake	The U.S. to Switzerland, Germany to the U.S.	IDV-MAS,IDV-IVR, MAS-IVR, IDV-MAS.	Yes
Earthquake	Switzerland to Australia, Australia to Germany.	IDV-IVR,MAS-IVR, IDV-MAS, MAS-UAI.	Yes
Earthquake	Portugal to Taiwan.	MAS-IVR.	Yes
FIFA World Cup	Singapore to Canada, Brazil to India, Spain to Canada.	IDV-IVR, PDI-MAS, PDI-LTO,˜IDV-MAS, IDV-LTO, MAS-LTO.	Yes
FIFA World Cup	The U.S. to Austria, Singapore to Spain, and Brazil to Spain.	MAS-IVR, LTO-IVR, IDV-IV, PDI-MAS, PDI-LTO,IDV-MAS.	Yes
FIFA World Cup	the U.S. to Pakistan, Russia to Austria, the U.S. to Spain.	PDI-MAS, MAS-IVR, MAS-IV, IDV-MAS, PDI-LTO,IDV-MAS.	Yes
FIFA World Cup	the U.S. to France, Austria to Canada, and the U.K. to the U.S.	PDI-MAS, MAS-IV, MAS-IV, LTO-IVR, UAI-IVR,UAI-IVR, MAS-LTO.	Yes

To conclude discussion on the cultural barrier, we can say news related to earthquakes has crossed cultural barriers. For instance, the maximum number of countries with a different score in each dimension propagate news to other countries. Only two cultural dimensions (individualism, indulgent versus restraint) were observed as barriers, as the news seems to propagate mostly to countries which are more or less similar in these two dimensions (see the second and the last circle in Fig. 14). For the FIFA World Cup, we observe that half of the countries with a similar score in dimension of individualism propagate news to each other (see Fig. 15). Similarly, half of the countries appeared to spread news concerning Global Warming topic, with a similar score within the individualism dimension (see Fig. 13).

6 Results and discussion

Experiments of the proposed methodology on three types of events have brought some insights regarding information propagation barriers in relation to the type of observed events. The results of mono-lingual information cascading indicate that news related to

Table 8 Score of each cultural dimension for different countries (For the interpretation of values see Section 2)

Country	Power distance (PDI)	Indivi dualism (IDV)	Mascu linity vs Femi-ninity (MAS)	Uncer tainty avoid-ance by individ-ualism (UAI)	Long and short term ori-entation (LTO)	Indul gence vs Restraint (IVR)
Africa East	64	27	41	52	32	40
Africa West	77	20	46	54	9	78
Arab countries	80	38	53	68	23	34
Argentina	49	46	56	86	20	62
Australia	38	90	61	51	21	71
Austria	11	55	79	70	60	63
Bangladesh	80	20	55	60	47	20
Belgium	65	75	54	94	82	57
Brazil	69	38	49	76	44	59
Bulgaria	70	30	40	85	69	16
Canada	39	80	52	48	36	68
Chile	63	23	28	86	31	68
China	80	20	66	30	87	24
Colombia	67	13	64	80	13	83
Croatia	73	33	40	80	58	33
Czech Rep	57	58	57	74	70	29
Denmark	18	74	16	23	35	70
El Salvador	66	19	40	94	20	89
Estonia	40	60	30	60	82	16
Egypt	70	25	45	80	7	4
Finland	33	63	26	59	38	57
France	68	71	43	86	63	48
Germany	35	67	66	65	83	40
Great Britain	35	89	66	35	51	69
Greece	60	35	57	112	45	50
Hong Kong	68	25	57	29	61	17
Hungary	46	80	88	82	58	31
India	77	48	56	40	51	26
Indonesia	78	14	46	48	62	38
Iran	58	41	43	59	14	40
Ireland	28	70	68	35	24	65
Italy	50	76	70	75	61	30
Japan	54	46	95	92	88	42
Korea South	60	18	39	85	100	29
Latvia	44	70	9	63	69	13
Lithuania	42	60	19	65	82	16
Luxembourg	40	60	50	70	64	56
Malaysia	104	26	50	36	41	57

Table 8 (continued)

Country	Power distance (PDI)	Indivi dualism (IDV)	Mascu linity vs Femi-ninity (MAS)	Uncer tainty avoid-ance by individ-ualism (UAI)	Long and short term ori-entation (LTO)	Indul gence vs Restraint (IVR)
Malta	56	59	47	96	47	66
Mexico	81	30	69	82	24	97
Morocco	70	46	53	68	14	25
Netherlands	38	80	14	53	67	68
New Zealand	22	79	58	49	33	75
Norway	31	69	8	50	35	55
Pakistan	55	14	50	70	50	0
Peru	64	16	42	87	25	46
Philippines	94	32	64	44	27	42
Poland	68	60	64	93	38	29
Portugal	63	27	31	104	28	33
Romania	90	30	42	90	52	20
Russia	93	39	36	95	81	20
Serbia	86	25	43	92	52	28
Singapore	74	20	48	8	72	46
Slovak Rep	104	52	110	51	77	28
Slovenia	71	27	19	88	49	48
Spain	57	51	42	86	48	44
Sweden	31	71	5	29	53	78
Switzerland	34	68	70	58	74	66
Taiwan	58	17	45	69	93	49
Thailand	64	20	34	64	32	45
Trinidad and Tobago	47	16	58	55	13	80
Turkey	66	37	45	85	46	49
U.S..	40	91	62	46	26	68
Uruguay	61	36	38	100	26	53
Venezuela	81	12	73	76	16	100
Vietnam	70	20	40	30	57	35

sports and natural disasters propagated smoothly based on the fact that the topic of discourse was contiguous, whereas for Global Warming the conversation was more divergent (see Fig. 4). In other words, after looking into the text within chains/communities of articles, we found that news articles related to Global Warming discuss more about other topics rather than only Global Warming whereas chains of the other two topics are more focused on the relevant description (see Section 5.1.2). The concept of understanding information propagation through information cascading has been addressed only in the context of social networks. We are focused to use the same cascading concept and structure over news articles. Although social media is a attractive and effective way of information spreading, a

large number of people still rely on newspapers and have a habit to follow print or broadcast mainstream media. While performing cascading experiments, we mainly came across two challenges: 1) Social networks are based on a well-defined and structured architecture whereas in the case of news, such structures are unavailable, and 2) In many cases social networks provide feature of translation that makes a bit easy to find the textual similarity whereas for news articles such features are still not mature enough.

Analysis of multi-lingual temporal spreading suggests that news related to natural disasters exhibits relevant and longer cascading chains of articles indicating smoother information propagation than in the domain of sports and climate changes. We conclude two types of results at the end of this analysis. Firstly, influence of language out of five languages has been identified for each distinctive domain (English language appeared to have more influence for the FIFA World Cup and Global Warming events, whereas Spanish appeared to have more influence for the FIFA World Cup and earthquakes). Secondly, we looked into the text of chains of news articles similar to the mono-lingual analysis and found that news articles related to natural disasters and FIFA World Cup include the relevant communication in long/short chains when compared to the topic of Global Warming (see Fig. 6). Overall, the results of our linguistic analysis correlated with our research hypothesis. Visualizing the temporal information spreading across different languages is one of the challenges in visual analytics. We built a prototype to help understanding temporal spreading though, there are still many improvements required. Firstly, visualization tool is not fine grained at hour, days, week levels. Secondly, 3D tool with zooming functionality can improve this visualization. The analysis of an economic barrier that has been performed over Google Maps using economic rankings of publishers' countries has not established compelling outcomes until we adopted another method. Using the economic ranking of countries, we observed information propagation from economically strong countries to economically weaker countries and vice versa. In all three distinctive domains, news related to Global Warming did not reach economically weaker countries, such as Iran and Nigeria. Using the income-level of countries, we analyzed the spread of news through alluvial diagrams and found that more information was spreading among economically stronger countries than in economically weaker countries (see Fig. 7). Economic barrier is normally more valuable to detect when there is a lot of economic interactivity between two countries and vice versa (Wu, 2007). Other than this, selection of events (conflicting events, popular events) also matters for economic barrier (Segev, 2015). For instance, in our three events, Global Warming is more suitable event for this barrier than FIFA World Cup and Earthquakes.

Our time zone analysis over Google maps suggests a harmony of spreading information related to natural disasters and sports (see Figs. 8, and 9). For natural disasters, European countries appeared to be the origin of the information spreading. In the case of sports, our results were inverse; the origin of the news were mostly countries with larger time zone differences but spreaders were close with respect to time zones (European countries). For climate change, the time zone barrier does not show any significant findings.

Geographical analysis portrays that a high number of publishers and news articles are from the countries that have large geographic areas (see Fig. 10). Apart from this, we did not find any significant differences in process of information spreading among the three distinctive domains. As a result, we could not confirm our research hypothesis on different spreading patterns that were generated for geographical barrier based on three distinctive domains.

Cultural differences also do not demonstrate a correlation with our research hypothesis (see Figs. 13, 14, and 15). Nevertheless, our analysis has shown an interesting observation

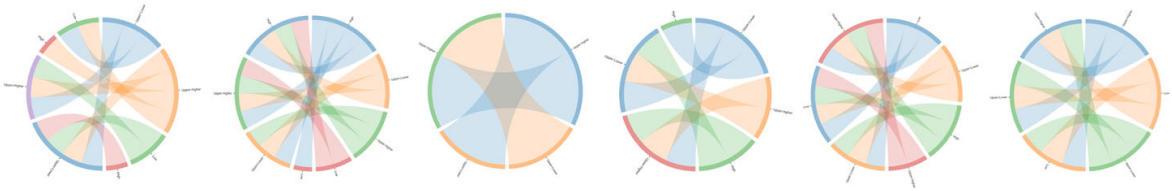


Fig. 13 News propagation depiction across all cultural dimensions (from left to right: PDI, IDV, MAS, UAI, LTO and IVR) related to Global Warming

that Argentina is the country which culturally supports sports activities more than other countries. In fact, we observe that some articles propagate news from a source country to the destination country however it also appeared that countries with common culture propagate news between each other. As a separate result of each domain, we see that news articles related to either Global Warming or earthquakes propagate news to countries that share cultural characteristics whereas news articles related to FIFA World Cup propagate news to those countries with entirely different culture.

Political alignment is important strategy that can be used to control the coverage of news in news agencies. Analysis of political involvement suggests that information on natural disasters spread smoothly, as earthquake-related news was mostly published by publishers that have neutral, progressive, and impartial political alignment (see Table 6). For Global Warming and FIFA World Cup we were not able to summarize the results due to lack of information. Coverage of these both events is not performed by a particular type of publishers (see Table 6).

Overall, our analytical findings suggest that most of the barriers (linguistic, economical, time zone, and political) by and large support our hypothesis. The generalizability of the results is to some extent limited by all the steps including computation and representation. By finding associations we are recognizing that the same information have presence across the barriers and we are not speculating on the cause of that. For instance, if two articles in different languages are similar that we conclude the information cross the linguistic barrier. We notice that the reasons for crossing that can be different and we are not investigating them. Firstly, by changing the cross-lingual similarity method, the results could be clearer and efficient. Secondly, with the availability of publishers' profiles along with the accurate location of their headquarters, the results would be more accurate. Additionally, more experiments on different types of events would enable more precise and robust comparison of information barriers.

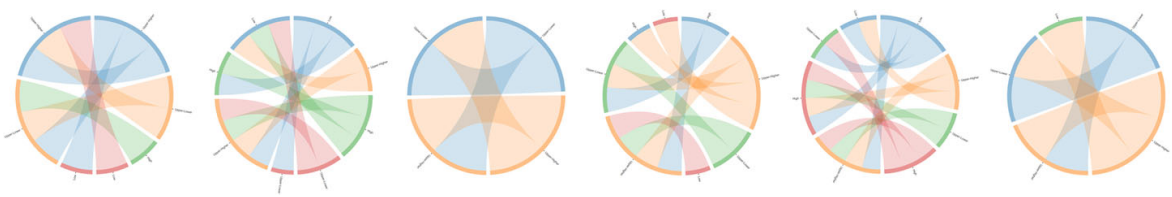


Fig. 14 News propagation depiction across all cultural dimensions (from left to right: PDI, IDV, MAS, UAI, LTO and IVR) related to earthquakes

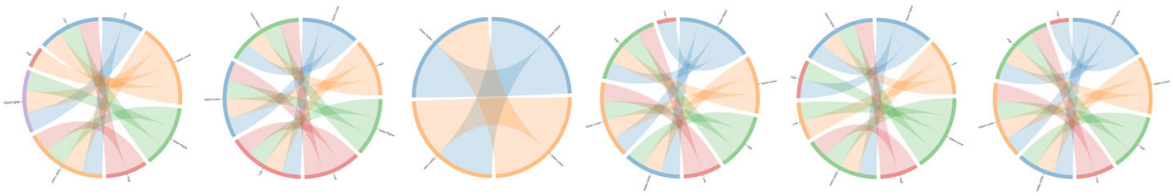


Fig. 15 News propagation depiction across all cultural dimensions (from left to right: PDI, IDV, MAS, UAI, LTO and IVR) related to FIFA World Cup

7 Conclusions

In this paper, we focused on the analysis of information spreading barriers by observing different aspects of news spreading in a global setting. Our motivation was primarily to understand the multilingual information cascading regarding different types of events within news articles, as it is not only valuable for journalists but it is also beneficial in a pragmatic sense for those who wish to follow globally influential-events (e.g. football and basketball in Europe) and investigate the influence of multiple barriers (e.g. economical, geographical, time zone, political and cultural) across different types of events. We firstly characterized the concept of information cascading on news articles in different languages and then find the total and largest cascading chains across distinct kind of events: Global Warming, FIFA World Cup and earthquakes. We also performed analysis to detect the influence of multiple barriers on event-centric news spreading.

In order to answer the first research question of this study - What are the properties (ratio and size) and values of cascading chains in events of different domains? - we identified the number of total communities showing information spreading as 196, 147, and 95 in the earthquake, FIFA World Cup, and Global Warming, respectively. Similarly, the largest cascading chains were in the earthquakes, FIFA World Cup, and Global Warming in English (28 articles), German (24 articles), and English (21 articles), respectively.

However, regarding the second research question - Do the different information cascading chains have any relations with each other? - There is a strong relation between the size and time of cascading chains and information spreading as the time duration of the longest chain of the earthquake was of 2.5 years and 3 months and single day for the FIFA World Cup and Global Warming respectively. Overall, it shows that more news propagated earthquakes than FIFA World and Global Warming.

In order to answer the third research question - Do the economic, geographical, time zone, and cultural values influence event-centric news spreading?- We observed all barriers having a certain effect on news propagation related to events. Firstly, it appears that spreading across languages was influenced by the scope of the event, the geographical size of an area directly related to the amount of news published from this area, places having the same culture publish similar news, news spreads firstly towards areas with adjacent time zones and economic barriers show more news spread upwards like low-income countries to high-income countries. Secondly, a comparison among the events shows that news related to FIFA World Cup propagated toward countries with shorter time zone differences and news propagated between countries with larger time zone differences in other events, economic values indicate that news related to Global Warming did not cross the economic barrier, countries with larger areas have more publishers but European countries have a large number of news articles related to earthquakes, and news related to earthquakes cross cultural barriers effectively than FIFA World Cup and Global Warming.

Finally, the answer to the last research question - What is the correlation of news spreading among events of different domains and political alignment of news publishers? - Was easier to predict based on the simple political alignment of publishers and it suggested that news related to an earthquake event is propagated mostly by those publishers which were politically neutral. Overall, our findings suggest that news published in news articles propagates to a greater degree across languages for natural disasters (earthquake events) than climate change (Global Warming) and sports (FIFA World Cup). In our experiments, we observed more cascading chains as well as longer cascading chains for earthquakes. If we look at the temporal difference between news articles, it shows that there is a larger time difference in the longest cascading chain of earthquake events (which is almost 2.5 years compared to 3 months for FIFA World Cup and 1 day for Global Warming). Our analysis of an economical barrier suggest that for Global Warming events, more news has propagated among economically stronger countries than to economically weaker countries. We found strong evidence that the news related to Global Warming has not crossed the economic barrier to reach Iran and Nigeria, which have currently considered as economically weaker countries. When we looked for information spreading in the domain of natural disaster, cultural barriers were important and difficult to cross. When we moved on to climate change events, linguistic and cultural barrier were difficult to cross, while for sports events, time zone barrier shows to be difficult to cross.

As the result of our research, we have provided a new publicly available dataset to help in understanding information spreading within three domains: natural disasters, climate change and sports. Apart from the data sets, a reusable visualization has been developed to show the real-time spreading of events using cross-lingual news articles. To this end, we utilized Google Maps, alluvial and chord diagrams to look at the spreading of information across the world with economic, cultural, time zone, and geographical distance between larger entities, such as countries. Moreover, we have considered the political involvement of publishers while analyzing the spreading patterns within different domains.

Acknowledgements This work was supported by the Slovenian Research Agency under the project J2-1736 Causalify and co-financed by the Republic of Slovenia and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 812997.

Data availability The datasets generated during and/or analysed during the current study are available on the Zenodo repository[15].

References

Al-Samarraie, H., Eldenfria, A., & Dawoud, H. (2017). The impact of personality traits on users' information-seeking behavior. *Information Processing & Management*, *53*(1), 237–247.

[15] https://zenodo.org/record/4117411

Alla, S., Sullivan, S. J., McCrory, P., & Hale, L. (2011). Spreading the word on sports concussion: citation analysis of summary and agreement, position and consensus statements on sports concussion. *British Journal of Sports Medicine*, *45*(2), 132–135.

Andrews, S., Gibson, H., Domdouzis, K., & Akhgar, B. (2016). Creating corroborated crisis reports from social media data through formal concept analysis. *Journal of Intelligent Information Systems*, *47*(2), 287–312.

Bakshy, E., Messing, S., & Adamic, L.A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, *348*(6239), 1130–1132.

Brank, J., Leban, G., & Grobelnik, M. (2017). Annotating documents with relevant wikipedia concepts. *Proceedings of SiKDD*.

Büyüksarıkulak, A. M., & Kahramanoğlu, A. (2019). The prosperity index and its relationship with economic growth: Case of turkey. *Journal of Entrepreneurship, Business and Economics*, *7*(2), 1–30.

Camaj, L. (2010). Media framing through stages of a political discourse: International news agencies' coverage of kosovo's status negotiations. *International Communication Gazette*, *72*(7), 635–653.

Chang, T.-K., & Lee, J.-W. (1992). Factors affecting gatekeepers' selection of foreign news: A national survey of newspaper editors. *Journalism Quarterly*, *69*(3), 554–561.

Cui, Y., Ni, S., Shen, S., & Wang, Z. (2020). Modeling the dynamics of information dissemination under disaster. *Physica A: Statistical Mechanics and its Applications*, *537*, 122822.

Dagon, D., Zou, C. C., & Lee, W. (2006). Modeling botnet propagation using time zones. In *NDSS*, (Vol. 6 pp. 2–13).

Estrada, E. (2011). Community detection based on network communicability. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *21*(1), 016103.

Glavaš, G., Franco-Salvador, M., Ponzetto, S. P., & Rosso, P. (2018). A resource-light method for cross-lingual semantic textual similarity. *Knowledge-Based Systems*, *143*, 1–9.

He, M., & Lee, J. (2020). Social culture and innovation diffusion: a theoretically founded agent-based model. *Journal of Evolutionary Economics*, 1–41.

Hoftede, G., Hofstede, G. J., & Minkov, M. (2010). *Cultures and organizations: software of the mind: intercultural cooperation and its importance for survival*. McGraw-Hill.

Hong, X., Yu, Z., Tang, M., & Xian, Y. (2017). Cross-lingual event-centered news clustering based on elements semantic correlations of different news. *Multimedia Tools and Applications*, *76*(23), 25129–25143.

Jin, H. (2017). Detection and characterization of influential cross-lingual information diffusion on social networks. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 741–745).

Khosrowjerdi, M., Sundqvist, A., & Byström, K. (2020). Cultural patterns of information source use: A global study of 47 countries. *Journal of the Association for Information Science and Technology*, *71*(6), 711–724.

Krajewski, R., Rybinski, H., & Kozlowski, M. (2016). A novel method for dictionary translation. *Journal of Intelligent Information Systems*, *47*(3), 491–514.

Kumar, S., Saini, M., Goel, M., & Panda, B.S. (2020). Modeling information diffusion in online social networks using a modified forest-fire model. *Journal of Intelligent Information Systems*, 1–23.

Leban, G., Fortuna, B., Brank, J., & Grobelnik, M. (2014). Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 107–110).

Maurer, P., & Beiler, M. (2018). Networking and political alignment as strategies to control the news: Interaction between journalists and politicians. *Journalism Studies*, *19*(14), 2024–2041.

Miritello, G., Moro, E., & Lara, R. (2011). Dynamical strength of social ties in information spreading. *Physical Review E*, *83*(4), 045102.

Quezada, M., Pe na-Araya, V., & Poblete, B. (2015). Location-aware model for news events in social media. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 935–938).

Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, *76*(3), 036106.

Segev, E. (2015). Visible and invisible countries: News flow theory revised. *Journalism*, *16*(3), 412–428.

Segev, E., & Hills, T. (2014). When news and memory come apart: A cross-national comparison of countries' mentions. *International Communication Gazette*, *76*(1), 67–85.

Şenel, L. K., Yücesoy, V., Koç, A., & Çukur, T. (2017). Measuring cross-lingual semantic similarity across european languages. In *2017 40th International Conference on Telecommunications and Signal Processing (TSP)*, IEEE (pp. 359–363).

Sittar, A., Mladenić, D., & Erjavec, T. (2020). A dataset for information spreading over the news. In *Proceedings of the 23th International Multiconference Information Society SiKDD*, (Vol. C pp. 5–8).

Vulic, I., & Moens, M.-F. (2014). Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)* (pp. 349–362). East Stroudsburg: ACL.

Watanabe, K., Ochi, M., Okabe, M., & Onai, R. (2011). Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 2541–2544).

Wei, H., Sankaranarayanan, J., & Samet, H. (2020). Enhancing local live tweet stream to detect news. *GeoInformatica*, 1–31.

Wilke, J., Heimprecht, C., & Cohen, A. (2012). The geography of foreign news on television: A comparative study of 17 countries. *International Communication Gazette*, *74*(4), 301–322.

Wu, H. D. (2007). A brave new world for international news? exploring the determinants of the coverage of foreign news on us websites. *International Communication Gazette*, *69*(6), 539–551.

Wu, H. D. (1998). Investigating the determinants of international news flow: A meta-analysis. *Gazette (Leiden, Netherlands)*, *60*(6), 493–512.

Chapter 3

News Reporting Differences Across the Barriers

There are many types of cultures such as societal, organizational, and business cultures, etc. The hidden nature of cultural behavior causes some difficulties in measurement and definition. To cope with difficulties, researchers have developed measurements that measure culture on a general scale to compare differences among cultures and management styles. These results can be used to find similarities within a region and differences in other regions. Hofstede's national cultural dimensions (HNCD) have been widely used and cited in different disciplines. These measurements are the outcome of a factor analysis at the country level using a comprehensive survey instrument, with the goal of identifying systematic differences in national cultures. We present the classification results of culture from text such as news events and demonstrate its usefulness in categorizing news events from different categories (society, business, health, recreation, science, shopping, sports, arts, computers, games, and home) across different geographical locations.

Few studies have analyzed the relation of news outlets to political and economic activities. Generally, news reporting about different events is inclined toward certain characteristics of newspapers. It has also appeared that local newspapers have a relatively distinctive content emphasis. Filla [18] investigates the political participation by the local news outlets in elections and finds the relationship between the political participation and availability of local news outlets. We investigate the relationship between world events as reported in newspapers and the characteristics of the newspapers in terms of political alignment and economic conditions.

There is a significant increment in international news events. The importance of understanding the differences in news spreading has increased for researchers and professionals in many disciplines such as digital humanities, media studies, and journalism. The examples of most prominent recent events are the migration crisis in Europe, Brexit, and COVID-19. Furthermore, many factors influence the news selection, reporting, and spreading, such as cultural, political, economic, geographical, and linguistic. Nowadays, social scientists and psychologists are interested to know how the occurrences of events can influence the everyday life of people in the world.

The role of content is an essential research topic in news spreading. The content refers to the type of language that is used in the news. It is used to convey meaning and it can impact social and psychological constructs such as social relationships, emotion, and social hierarchy. Features that could classify reporting across different regions can be adapted to classify the news. News reporting differences can be reflected through one's speech, writing, images, etc.

The rest of the chapter presents the methodology and results of four related studies.

Firstly it shows the results of the classification of cultures on top of different types of news events across different geographical locations (see Section 3.1). It also presents the results of an investigation of world events as reported in newspapers and the characteristics of the newspapers in terms of political alignment and economic conditions (see Section 3.2). Secondly, it presents an enhanced topic modeling approach along with an analysis of COVID-19 news across different economic and political contexts. Thirdly, it presents results of clustering news reporting and a comparison of bag-of-words and stylistic features (see Section 3.3).

3.1 Classification of News Events

The news events have a cultural influence of varying intensity depending upon the domain. The influence of cultures is investigated by classifying cross-culture privacy, attitude prediction, and influence on business in previous studies. Since this thesis is analyzing the news spreading barriers where two of the barriers are cultural and geographical barriers, performing the classification of culture given news events belonging to different domains and geographic locations, was the start of profiling the news spreading barriers. This section presents a study that is investigating the classification of cultures across different locations (117 countries) using different classical machine learning algorithms. The used news events were generated by the Event Registry. The selected news events are published by the top ten daily read newspapers (Asahi, Chinadaily, Dawn, NYTimes, SMH, TheGuardian, Timesofindia, Washingtonpost, WSJ, Zaman). The news events belong to society, business, health, recreation, science, shopping, sports, arts, computers, games, and home.

For this multi-class classification task, we investigate the performance of classification methods such as SVM, Decision Tree, kNN, Naive Bayes, Logistic Regression, and pre-trained Glove embeddings. We input character and word n-grams varying the number of grams from 5K to 20K. The promising results were achieved using Logistic Regression, kNN, and Decision Tree. These results were presented at Proceedings of the 24th International Multiconference Information Society SiKDD in Slovenia, Ljubljana, in 2021 [60] with title *Classification of Cross-cultural News Events* by Abdul Sittar, and Dunja Mladenić. In this paper, a novel perspective of aligning news events across different cultures through categorising countries and news events is explored. To perform the categorisation, a cross-cultural annotated dataset is created where news articles belong to different categories including business, science, sports, and health. Lastly, experimental comparison of several classification models is conducted adopting different sets of features (word ngrams, character ngrams, and glove embeddings).

Classification of Cross-cultural News Events

Abdul Sittar*
abdul.sittar@ijs.si
Jožef Stefan Institute and Jožef Stefan International Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

Dunja Mladenić
dunja.mladenic@ijs.si
Jožef Stefan Institute and Jožef Stefan International Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

ABSTRACT

We present a methodology to support the analysis of culture from text such as news events and demonstrate its usefulness on categorising news events from different categories (society, business, health, recreation, science, shopping, sports, arts, computers, games and home) across different geographical locations (different places in 117 countries). We group countries based on the culture that they follow and then filter the news events based on their content category. The news events are automatically labelled with the help of Hofstede's cultural dimensions. We present combinations of events across different categories and check the performances of different classification methods. We also presents experimental comparison of different number of features in order to find a suitable set to represent the culture.

KEYWORDS

cultural barrier, news events, text classification

1 INTRODUCTION

Culture is defined as a collective programming of the mind which distinguishes the members of one group or category of people from another [9]. It has a huge impact on the lives of people and in result it influences events that involve cross-cultural stakeholders. News spreading is one of the most effective mechanisms for spreading information across the borders. The news to be spread wider cross multiple barriers such as linguistic, economic, geographical, political, time zone, and cultural barriers. Due to rapidly growing number of events with significant international impact, cross-cultural analytics gain increased importance for professionals and researchers in many disciplines, including digital humanities, media studies, and journalism. The most recent examples of such events include COVID-19 and Brexit [1]. There are few determinants that have significant influence on the process of information selection, analysis and propagation. These include cultural values and differences, economic conditions and association between countries. For instance, if two countries are culturally more similar, there are more chances that there will be a heavier news flow between them [10], [3]. In this paper, we focus on classification of news events across different cultures. We select some of the most read daily newspapers and collect information using Event Registry about the news they have published. Event Registry is a system which analyzes news articles, identifies groups of articles that describe the same event and represent them as a single event [7]. The description of the meta data of an event is shown in the Table 1. The main scientific contributions of this paper are the following:

(1) A novel perspective of aligning news events across different cultures through categorising countries and news events.
(2) A cross-cultural automatically annotated dataset in several different domains (Business, Science, Sports, Health etc.).
(3) Experimental comparison of several classification models adopting different set of features (character ngrams, GLOVE embeddings and word ngrams).

Table 1: The description of the meta data of an event.

Attributes	Description
title	title of the event
summary	summary of the event
source	event reported by a news source
categories	list of DMOZ categories
location	location of the event

2 RELATED WORK

In this section, we review the related literature about the influence of culture, its representation and classification in different fields.

Countries that share a common culture are expected to have heavier news flows between them when reporting on similar events [10]. There are many quantitative studies that found demographic, psychological, socio-cultural, source, system, and content-related aspects [2].

Cross-cultural research and understanding the cultural influences in different fields have competitive advantages. The goal of researching the impact of culture might be to draw conclusions in which way the cultural factors influence a specific corporate action. There are many type of cultures such as societal, organizational, and business culture etc [8].

The hidden nature of cultural behavior causes some difficulties in measurement and defining these. To cope with difficulties, researchers have developed measurements that measure culture on a general scale to compare differences among cultures and management styles. These results can be used to find similarities within a region and differences to other regions. There are many models that have tried to explain cultural differences between societies. Hofstede's national culture dimensions (HNCD) have been widely used and cited in different disciplines [6, 5]. Hofstede's dimensions are the result of a factor analysis at the level of country means of comprehensive survey instrument, aimed at identifying systematic differences in national cultural. Their purpose is to measure culture in countries, societies, sub-groups, and organizations; they are not meant to be regarded as psychological traits.

There is a plethora of research studies that were conducted to understand the cultural influences such as cross-culture privacy and

Information Society 2021, 4–8 October 2021, Ljubljana, Slovenia

attitude prediction, and cultural influences on today's business. [4] explores how culture affects the technological, organizational, and environmental determinants of machine learning adoption by conducting a comparative case study between Germany and US. Rather than looking at the influence of cultural differences within one domain, we intend to understand association between news events belonging to different domains (society, business, health, recreation, science, shopping, sports, arts, computers, games and home) and different cultures (117 countries from all the continents). We conduct this research to find an appropriate representation and classification of culture across different domains.

3 DATA DESCRIPTION

3.1 Dataset Statistics

We choose the top 10 daily read newspapers in the world in 2020 [1] and collect the events reported by these newspapers using Event Registry [7] over the time period of 2016-2020. Approximately 8000 events belongs to each newspaper with exception of "Zaman" that has only 900 events. Figure 1 shows the number of events reported by the selected newspapers on a yearly basis. This dataset can be found on the Zenodo repository (version 1.0.0) [2]

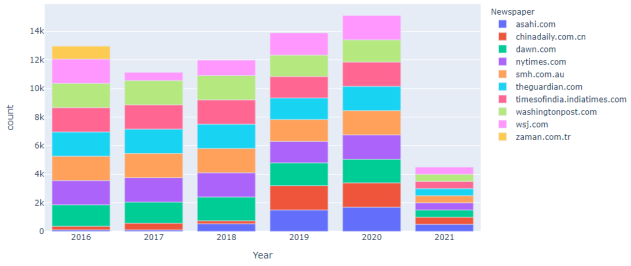


Figure 1: Each color in a bar represents the total number of events per year by a daily newspaper and a complete bar shows the total number of events per year by all the newspapers.

The attributes of an event with description are displayed in Table 1. Few attributes are self-explanatory such as title, summary, date, and source. DMOZ-categories are used to represent topics of the content. The DMOZ project is a hierarchical collection of web page links organized by subject matters [3]. Event Registry use top 3 levels of DMoz taxonomy which amount to about 50,000 categories [4].

4 MATERIAL AND METHODS

4.1 Problem Definition

There are two main parts of the problem that we are addressing. The first part is to label the examples by assigning a culture C to a news event E using its location L. The second part is a multi-class classification task where we predict the culture C of a news event E using its summary description S and its content category G as provided by the Event Registry. This task can be formulated as:

$$C = f(S, G)$$

C donates the culture of the news event, f is the learning function, S donates summary of a news event and G donates category of a news event (see Table 1).

4.2 Methodology

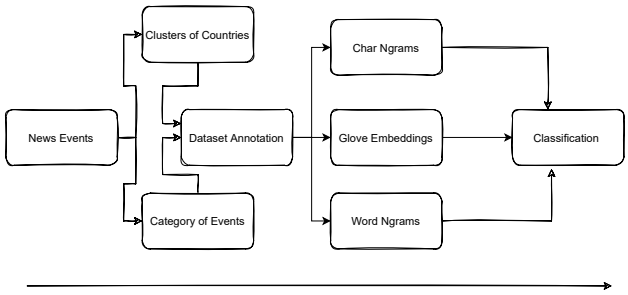


Figure 2: Classification of cross-cultural news events.

4.2.1 Data labeling. Each news event has information about the type of categories to which it belongs and the location where it happened (see Table 1). Each event has many categories and each category has a weight reflecting its relevance for the event. We only keep the most relevant categories and group the news events based on their categories. For each group of events, we estimate the cultural characteristic of each event through the country of the place where the event occurred. We cluster the countries based on their culture. We utilize the Hofstede's national culture dimensions (HNCD) to represent the culture of a country. We take average of cultural dimensions and call it average cultural score. Based on this score, we find optimal number of clusters using popular clustering algorithm k-means (see Figure 4). Finally, we label each news event with one of the six cultural clusters.

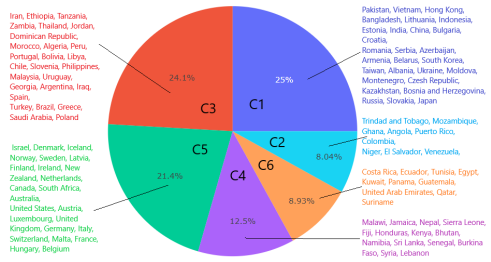


Figure 3: The pie chart depicts the percentage of the news events that occurred in six different clusters (each cluster consists of a list of countries with similar culture).

4.2.2 Data representation. Each news event in Event Registry has associated categories with it along with a weight (see Table 1), we take the top categories based on their weight. In case of multiple categories with equal weight, we sort them alphabetically and keep the first one. We represent each news event by a short summary S and a set of content categories G.

[1] https://www.trendrr.net/
[2] https://zenodo.org/record/5225053
[3] https://dmoz-odp.org/
[4] https://eventregistry.org/documentation?tab=terminology

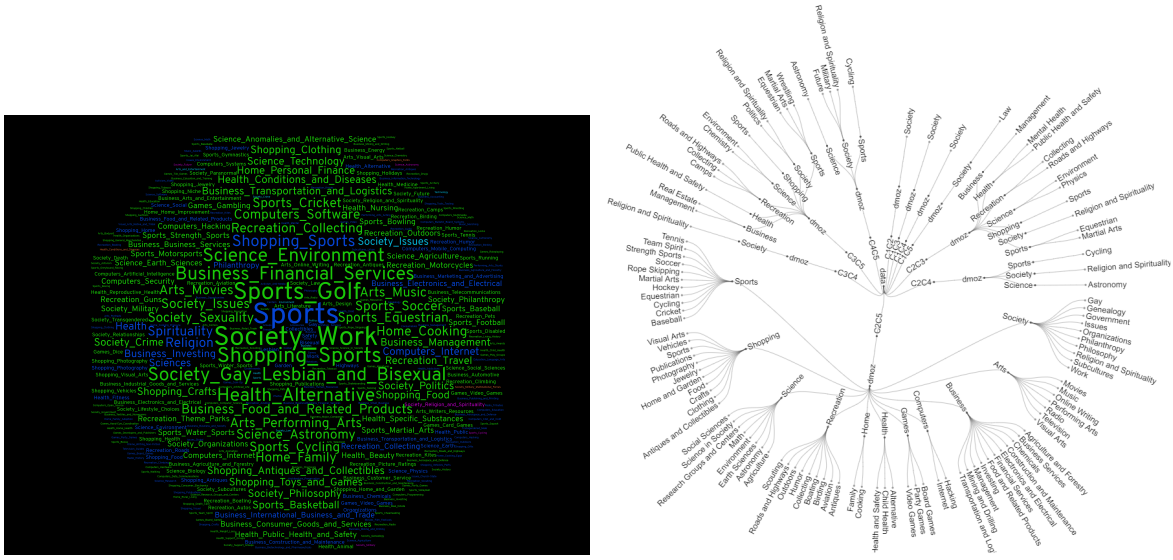


Figure 4: In word cloud, the color of each word shows cluster to whom it belongs (see Figure 3). Radial dendrograms illustrate the shared categories of news events between the pair of six clusters.

4.2.3 Data Modeling. For multi-class classification task, we use simple classification models (SVM, Decision Tree, KNN, Naive Bayes, Logistic Regression) as well as neural network. For simple classification models, we input character and word ngrams varying the number of ngrams and compare the results. We also use pre-trained Glove embeddings.

5 EXPERIMENTAL EVALUATION

5.1 Evaluation Metric

For multi-class classification task, we use following most commonly used evaluation measures: accuracy, precision, recall, and F1 score.

6 RESULTS AND ANALYSIS

6.1 Annotation Results

The results of annotation are six clusters where almost 50% news events belong to the two clusters (shown with red and blue colors) and remaining 50% belong to the other four clusters 3. Looking in each group, we find that clusters do not lies in a specific geographic area or a continent. Rather all the countries in a cluster belong to the different continents. Similarly, these clusters do not have all the countries that are economically rich or poor.

There are more categories in green and red colors in the word cloud (see Figure 4) which represent to the cluster with that colors. Radial dendrograms in Figure 4 present the shared categories between the clusters. In the figure, root of the tree is data and then there are ten pair of clusters that share the same categories. The objective of this whole process was to keep news events according to the category to whom they belongs. Moreover, we can only observe the cultural differences when we have same type of news events from different places.

6.2 Classification Results

Fro the experimental results we can see that the best performance is achieved by Logistic Regression, kNN and Decision Tree. The performance of SVM varies depending on the number of selected features: the highest F1-score is achieved with the top 10K or 20K word ngrams using 1 to 3 word ngrams (see Figure 5). Looking at the character ngrams, the highest F1-score is achieved when we select the top 15K characters for all the tested algorithms except Naive Bayes which declines in performance with the growing set of features. Based on these settings, we achieve the highest accuracy (0.85) using Logistic Regression. Using Glove embeddings, we experiment with and without using the category of event. The highest F1-score with and without the category is 0.80 and 0.79 respectively.

7 CONCLUSIONS AND FUTURE WORK

For researchers and professionals, it is very important to analyse the cross-cultural differences in different disciplines. As the international impact is increasing and international events are becoming popular, the need to develop some automatic methods is significantly increasing and leaving a blank space. We conducted experiments on news events related to different fields to have a broader look on data and machine learning methods. Further research would be helpful in examining the impact of specific socio-cultural factors on news events. In this research work, we estimate the culture of a specific place by its country, use basic features and simple classification models. To continue this work further, we would like to improve feature set such as by including part of speech tagging (POS) as well as other state of the art embeddings.

ACKNOWLEDGMENTS

The research described in this paper was supported by the Slovenian research agency under the project J2-1736 Causalify and by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 812997.

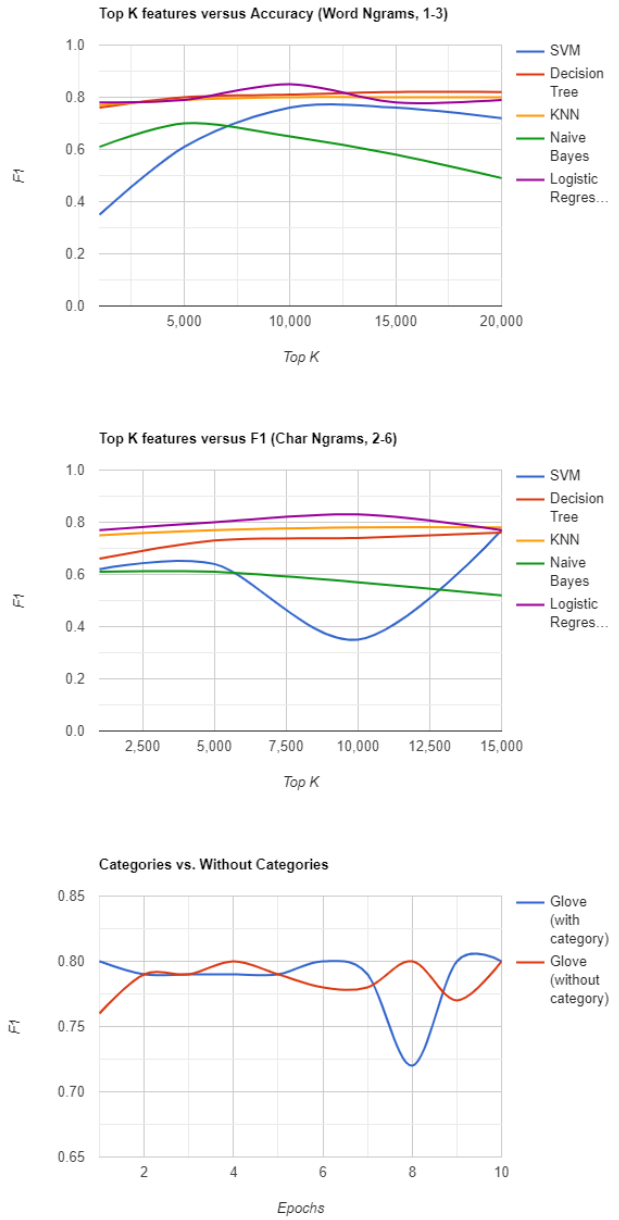


Figure 5: First two line charts illustrate the variations in F1 score by simple classification models after varying the number of features. The first line chart depicts the results of word ngrams whereas the second one shows the results for character ngrams. The last line graph presents comparison between Glove embeddings (with and without category feature).

REFERENCES

[1] Sara Abdollahi, Simon Gottschalk, and Elena Demidova. 2020. Eventkg+ click: a dataset of language-specific event-centric user interaction traces. *arXiv preprint arXiv:2010.12370*.

[2] Hosam Al-Samarraie, Atef Eldenfria, and Husameddin Dawoud. 2017. The impact of personality traits on users' information-seeking behavior. *Information Processing & Management*, 53, 1, 237–247.

[3] Tsan-Kuo Chang and Jae-Won Lee. 1992. Factors affecting gatekeepers' selection of foreign news: a national survey of newspaper editors. *Journalism Quarterly*, 69, 3, 554–561.

[4] Verena Eitle and Peter Buxmann. 2020. Cultural differences in machine learning adoption: an international comparison between germany and the united states.

[5] Meihan He and Jongsu Lee. 2020. Social culture and innovation diffusion: a theoretically founded agent-based model. *Journal of Evolutionary Economics*, 1–41.

[6] Mahmood Khosrowjerdi, Anneli Sundqvist, and Katriina Byström. 2020. Cultural patterns of information source use: a global study of 47 countries. *Journal of the Association for Information Science and Technology*, 71, 6, 711–724.

[7] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, 107–110.

[8] Björn Preuss. 2017. Text mining and machine learning to capture cultural data. Technical report. working paper, 2. doi: 10.13140/RG. 2.2. 30937.42080.

[9] Giselle Rampersad and Turki Althiyabi. 2020. Fake news: acceptance by demographics and culture on social media. *Journal of Information Technology & Politics*, 17, 1, 1–11.

[10] H Denis Wu. 2007. A brave new world for international news? exploring the determinants of the coverage of foreign news on us websites. *International Communication Gazette*, 69, 6, 539–551.

3.1.1 News Reporting Differences across Political and Economic Contexts

There is a great deal of negotiation between political actors and journalists in news production to enhance their influence on news coverage. Similarly, the political alignment of publishers can more or less impact the coverage of different events. Economic conditions and associations between countries are other determinants for news coverage. Few studies have worked on finding the relation of news outlets to political and economic activities. News reporting about different events is inclined toward certain characteristics of newspapers. Similar to the previous section, we can call this study a subpart of the main problem of profiling the news spreading barriers. In this study, we focus on political and economic factors in news reporting. We proposed a methodology for clustering the daily read newspapers based on political and economic characteristics using different similarity mechanisms. We focused on finding representation of the news events for two characteristics (political and economic) separately that could be able to cluster the input of text in their original groups.

The best results are achieved for an economic attribute using the most frequent Wikipedia concepts, then with Jaccard similarity, and then lastly with Euclidean distance. For the political attribute, the best performing mechanism is Euclidean distance and Jaccard similarity. Based on the overall accuracy of the economic attribute, we can say that the representation with Wikipedia-concepts is better than DMOZ categories for capturing economic characteristics of the newspaper, while the opposite is true for capturing political attributes. Also, dynamic time warping is more suitable for economic characteristics but not for political characteristics.

These results were presented at the Central European Conference on Information and Intelligent Systems in 2021 [61] with the title *How are the economic conditions and political alignment of a newspaper reflected in the events they report on?* by Abdul Sittar, and Dunja Mladenić. In this paper, a novel methodology is evaluated on real-world news data using representation of world events by a set of concepts and content categories testing different similarity measures between newspapers using dynamic time warping with cosine similarity or with Euclidean distance on trend lines of concepts and categories, jaccard similarity between concepts and categories. This methodology includes selection, representation and clustering of the newspapers to analyze relationship between the reported events and characteristics of the newspaper.

How Are the Economic Conditions and Political Alignment of a Newspaper Reflected in the Events They Report On?

Abdul Sittar, Dunja Mladenic
Jozef Stefan Institute and Jozef Stefan International Postgraduate School
Department for Artificial Intelligence
Jamova Cesta 39, Ljubljana, Slovenia
{Abdul.sittar, dunja.mladenic}@ijs.si

Abstract. *The paper investigates relationship between world events as reported in newspapers and characteristics of the newspapers in terms of political alignment and economic conditions. We propose a novel methodology that includes selection, representation and clustering of the newspapers to analyse relationship of the events and characteristics of the newspaper. We represent world events by a set of concepts and content categories of the news articles reporting on the event. Each newspaper is represented by a set of events they reported about over several years. We investigate different similarity measures between the newspapers to see whether the newspapers with the same characteristics are reporting on similar events over a given time span.*
The results indicate: 1) the representation of the news events with the Wikipedia-concepts and DMOZ-categories appears an appropriate way to understand relationships between the newspapers, 2) economic conditions of the country of the newspaper publisher reflect better in Wikipedia-concepts than when using representation with DMOZ-categories, whereas for identifying politically aligned groups of newspapers, DMOZ-categories stand out more suitable, 3) for capturing economic groups, clustering using the Dynamic Time Warping similarity between the trend lines of newspapers is better aligned with the ground truth groups than others tested similarities, whereas for capturing political group, Jaccard distance using the frequent terms and Euclidean distance between the trend lines turn up more useful.

Keywords. Information Propagation Barriers, Political Alignment, Economic Conditions, Dynamic Time Warping (DTW), Euclidean Similarity, Jaccard Similarity.

1 Introduction

Economic strength and political situation of a country have a strongest association with news prominence. One of the assumptions of many news-flow related studies is that external variables such as economic power and political events in a country can define the scope of fame around the world. In fact, there are also various internal factors such as the structure of international telecommunications, the presence of news agencies, as well as the editorial practices and traditions that effect the news prominence of a country (Segev, 2015). According to news flow theories, multiple determinants impact international news spreading. The economic power of a country is one of the factors that influence news spreading. One of the parameters to represent the economic condition of a country is the economic growth/income level. Moreover, it was noted that the magnitude of economic interactivity between countries can also impact the news flow (Wu, 2007).

The idea of event-centric news spreading disclosed internationally and become popular due to globalization (Hong et al., 2017). Global events become famous and catch the attentions in all corner of the world. News agencies or news publishers play their role in this process. Varying nature of living styles, cultures, economic conditions, time zone, and geographical juxtaposition of countries present a significant role in the process of reporting on news events (Wei et al., 2020, Quezada et al., 2015, p. 935-938). The news to be spread wider cross multiple barriers such as linguistic, economic, geographical, political, time zone, and cultural barriers. News publisher have different characteristics such as the reporting languages, their political alignment, economic conditions, cultural values, time-zone, and geographical position of their headquarters. In this paper, we focus on two characteristics of news publishers, namely political alignment and economic situation. We select ten most read daily newspapers in the world in 2020 and collect information using Event Registry about the news they have published. Event Registry is a system which analyses news articles and identify groups of articles that describe the same event and represent them as single event (Leban et al., 2014,

p. 107-110). The description of the meta data of an event is shown in the Table 1.

Following are the main scientific contributions of this paper:

- A novel methodology that includes selection, representation and clustering of the newspapers to analyze relationship between the reported events and characteristics of the newspaper.
- Evaluation of the proposed methodology on real-world news data using representation of world events by a set of concepts and content categories testing different similarity measures between newspapers: Dynamic Time Warping (DTW) with cosine similarity or with Euclidean distance on trend lines of concepts and categories, Jaccard similarity between the concepts and categories.
- We show that what type of features are more suitable for clustering newspapers based on economic and political characteristics.

Table 1. The description of the meta data of an event

Attributes	Description
Uri	A unique event identifier
Title	Title of the event
Summary	Summary of the event
Event date	Date of the event in yyyy-mm-dd format
Source	Event reported by a news source
Total articles	Total articles reporting about the event
Categories	List of DMOZ (Directory Mozilla) categories
Concepts	List of Wikipedia concepts

2 Related Work

International news about different events led us to investigate the reasons why news regarding specific events either spread or do not spread to certain geographic areas. Media focuses on specific foreign and regional events based on some certain factors. For instance, spreading of events may tilt toward developed countries such as United States, the United Kingdom, or Russia. Moreover, it may be due to geographical juxtaposition (latitude, longitude) of countries (Wilke et al., 2012). There is a great deal of negotiation between political actors and journalists in news production to enhance their influence on news coverage. It will be true to say that fake news is produced based on many factors and it is surrounded by a paramount factor that is political effect (Martens et al., 2018). Therefore, political alignment of publishers can more or less impact their coverage of different events. Two of the determinants for news coverage are economic conditions and association between countries. These factors also impact information selection, analysis, and propagation (chang et al., 1992).

There are few studies that have worked on finding the relation of news outlets on political and economic activities. Generally, news reporting about different events (elections etc.) is inclined towards certain characteristics of newspapers. As there is a tendency to support underground and indirectly, the research interest in the reporting characteristics of each newspaper has begun (Jo et al., 2018). The study of information flows between media sources from different countries explores the dynamics underlying transnational communication spaces. Castells sees the emergent Euro-state not only as a political-economic zone but, by virtue of privileging its network character, also as a specific kind of communicative space (veltri , et al., 2012). It has also appeared that local newspapers have a relatively distinctive content emphasis (lin et al., 2001). Filla investigates the political participation by the local news outlets in elections and find the relationship between the political participation and availability of local news outlets (Filla, 2010, p. 679-692). Another study was conducted to find the correlation at the outlet level between public trust and experts' evaluation. It had compared the evaluation of accuracy of news outlets and trust scores (Schulz, et al., 2020). News agencies tend to follow the national context in which journalists operate. One of the related examples is the SARS epidemic study which found that cross-national contextual values such as political and economic situations impact the news selection (Camaj, 2020, p. 635-653). A great amount of work regarding fake news dwells on different strategies, while few studies considered political alignment to have a compelling effect on news spreading (Bakshy et al., 2015, p. 1130-1132). (Maurer, 2018, p. 2024-2041) strongly proved it to be a major strategy in news agencies to control the news and change accordingly due to the involvement of journalists and political actors.

Although the previous work involves relationship between outlets, and political activities or public interests, our work focused directly on studying and confirming the political alignment and economic conditions with news outlets. There should be a representation and computational method to reflect the political and economic characteristics. The objective is to find a representation able to cluster the input of text in their original groups. Previously a newspaper corpus having five knowledge fields (Human Sciences, Biological sciences, Social Sciences, Religion and Thought, Exact Sciences) in Brazilian Portuguese was used to verify whether an automated clustering process could create the correct clusters of newspapers (Afonso, et al., 2014). We utilize more than 65,000 news events published by top ten newspapers across different countries in English. In addition to that we consider the temporal information of events while finding similarities between newspapers. There are two

approaches (hierarchical and non-hierarchical clustering) to cluster the text. We focus on a non-hierarchical text clustering method. Non-hierarchical text clustering is applied when the goal is to produce text clusters which do not fit in specific knowledge hierarchy (Afonso, et al., 2014).

3 Data Description

3.1 Data Statistics

We choose the top 10 daily read newspapers in the world in 2020 (https://www.trendrr.net/) and collect the events reported by these newspapers using Event Registry over the time period of 2016-2020. Approximately 8000 events belong to each newspaper except "Zaman" (only 900 events) (see Table 1). Figure 1 shows the number of events reported by the selected newspapers on a yearly basis.

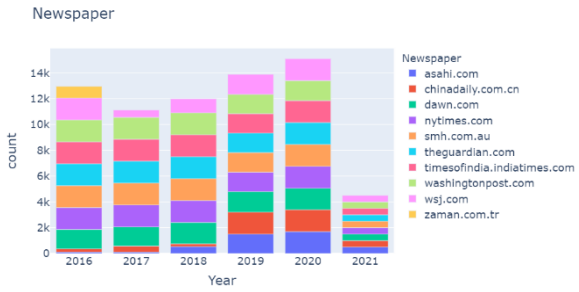


Figure 1. Each colour in a bar represents the total number of events per year by a daily newspaper and a complete bar shows the total number of events per year by all the newspapers

The attributes of an event with description have been displayed in the Table 1. Few attributes are self-explanatory such as Uri, title, summary, date, source, and total news articles. Concepts are the annotation for events. Concepts can represent entities (locations, people, organizations) or non-entities (things such as personal computer, toy). In Event Registry Wikipedia's URLs are used as concept URIs. DMOZ-categories represent what topic the content is about. The DMOZ project is a hierarchical collection of web page links organized by subject matters\footnote{https://dmoz-odp.org/}. Event Registry use top 3 levels of taxonomy which amount to about 50,000 categories (https://eventregistry.org/documentation?tab=terminology).

Each newspaper leans to a different political alignment in the political spectrum. We estimate the political alignment of a newspapers through the political alignment of its publisher and the economic conditions through the country of headquarter of its publisher (see Figure 2). We fetched the headquarter and the political alignment of each newspaper from Wikipedia Info-box. Each newspaper has its headquarters in different countries varying the economic ranking and the income levels (see Table 2).

There are four main income levels of the economies: Low-Income Economies (\$1035 or less), Lower-Middle Income Economies (\$1036 to \$4045), Upper-Middle Income Economies (\$ 4046 to \$12,535), and High-Income Economies (\$12,535 or more) (https://datahelpdesk.worldbank.org/knowledgebase/articles/906519). The overall ranking (from 1 to 149) of each country bases on 12 features: Safety and Security, Personal Freedom, Governance, Social Capital, Investment Environment, Enterprise Conditions, Market Access and Infrastructure, Economic Quality, Living Conditions, Health, Education, and Natural Environment. Table 2 shows the political alignment of each newspaper and economic conditions (ranking, income-level) of the headquarters of the newspapers (https://www.prosperity.com/rankings).

3.2 Similarity Measures

We propose to estimate similarity between the newspapers by looking at events they are reporting about over a period of time. In one case we consider trend lines of concepts or content categories characterising the events and apply Dynamic Time Warping or Euclidean distance. In the other case we ignore the time dimension and simply take the union of all the concepts or categories from events over the years.

3.2.1 Dynamic Time Warping (DTW)

Dynamic Time Warping is a method for calculating the similarity between two time series which can occur at different times or speeds. Its ability to warp time axis and find optimal alignment between time series has made it very popular. DTW has been used

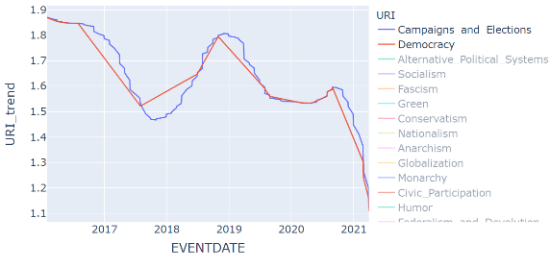


Figure 3. This line chart shows the trend lines of two Wikipedia-concepts (Campaigns and Elections, and Democracy) over time period of six years

in several disciplines such as: Speech recognition, gesture recognition, data mining, robotics, manufacturing and medicine. DTW aligns two time series in the way some distance measure is minimized (usually Euclidean or Cosine distance is used). Optimal alignment (minimum distance warp path) is obtained

by allowing assignment of multiple successive values of one time series to a single value of the other time series and therefore DTW can also be calculated on time series of different lengths (Strle, et al., 2009).

Table 2. Political alignment of the daily newspapers and the economic conditions (Prosperous ranking, Income-level) of the country of publisher's headquarters

Daily Newspapers	Headquarters	Total Events	Political Alignment	Economic Conditions
The Guardian	London	8804	Centre-left	13, High-income
The Wall Street Journal	New York	7094	Conservative	18, High-income
The New York Times	New York	8802	Liberal	18, High-income
The Washington Post	Washington	8657	Democrat	18, High-income
China Daily	Beijing	4833	Democratic Centralism	54, Upper-Middle-income
The Times of India	Bombay	8804	Centre-right	101, Low-income
The Sydney Morning Herald	Sydney	8829	Centre-left	16, High-income
The Asahi Shimbun	Osaka	4452	Centre-left, Liberalism	19, High-income
Dawn	Karachi	8389	Liberal, Centrist and Progressive	138, Low-income
Zaman	Istanbul	900	Pro-Government	94, Lower-Middle

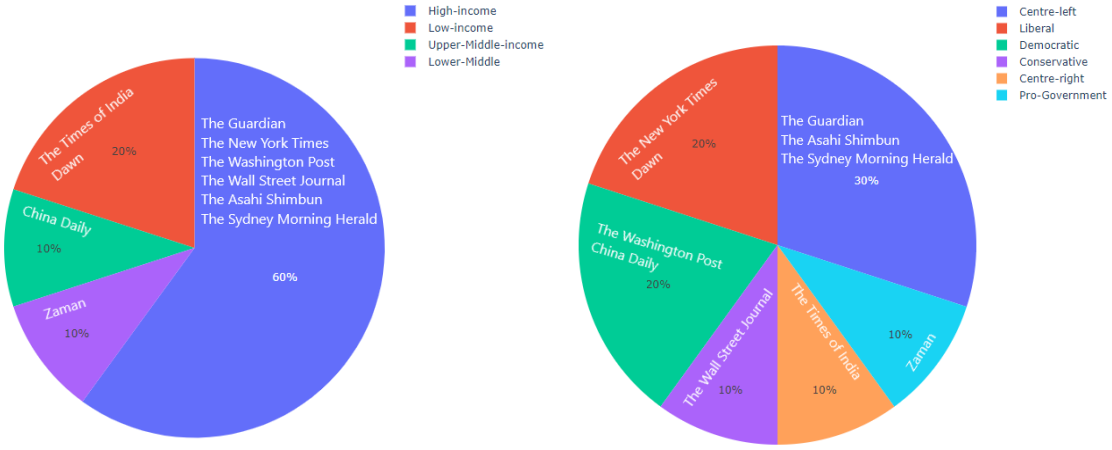


Figure 2. The pie chart on the top depicts the percentage of four categories of income levels associated with the daily newspapers. And the second pie chart illustrates the percentage of categories of political alignment associated with the daily newspapers (see Table 2)

3.2.2 Euclidean Distance

Euclidean distance between two points is the length of a line segment between two points (also called Pythagorean theorem as shown below).

$$d(A,B) = \sqrt{\sum_{i=0}^{n}(A_i - B_i)^2}$$

3.2.3 Jaccard Distance

Measuring the Jaccard similarity coefficient between two data sets is the result of division between the number of features that are common to all and the number of properties as shown below (Niwattanakul, et al., 2013, p. 380-384).

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

4 Methodology

We present a novel methodology for clustering the daily read newspapers based on political and economic characteristics using different similarity mechanisms. It is based on the trend lines of Wikipedia-concepts, and simple count of DMOZ-categories (Directory Mozilla) and Wikipedia-concepts related to the news events (see Figure 4). We built hierarchy of clusters and also compare the generated clusters with ground truth values.

We extract the 100 most-frequent Wikipedia-concepts for all the news events reported by each newspaper. Afterward, we generate the trend lines using The Hodrick Prescott filter (Bhowmik et al., 2021, p. 7-17) for each Wikipedia-Concepts per each daily read newspaper. An example of trend lines between two concepts is shown in Figure 3. Having a bunch of trend lines, we calculate the Dynamic Time Warping (DTW) distance between them using the cosine similarity for each daily read newspaper separately. We filter out the Wikipedia-concepts if their distance does not lie in threshold value of 0.1 (0.0 means the trends lines are absolutely similar whereas maximum value varies depending on the difference between trend lines). At this stage, we have pair of those Wikipedia-concepts that have similar trends over time for each newspaper. To calculate the distance between two newspapers, we measure the overlap between corresponding pairs to the newspapers. Then we built hierarchy of clusters using popular algorithm of hierarchical clustering called dendrograms. At each step in dendrograms, the two clusters that are most similar are joined into a single new cluster. We chose top four and six hierarchies separately (see Figure 5) for economic and political characteristics respectively. We revise this process similarly for 150 and 200 most-frequent Wikipedia-concepts for all daily read newspapers.

We apply a second mechanism to calculate the similarity between the trend lines. We choose the same number of Wikipedia-concepts and set the same threshold value as we set in the previous method to filter out non-similar Wikipedia-concepts. The only difference is that we use aligned values between trend lines and compute the similarity using euclidean distance rather than DTW. As this method does not tackle the situation if two trend lines have different lengths over time, we cut out the extra line and only keep similar length of two trend lines (DTW do handle this situation). Further we built hierarchy of clusters in the similar way as we did in the previous mechanism.

Lastly, we extract two lists of the unique Wikipedia-concepts, and the DMOZ-categories, and two lists of all the Wikipedia-concepts, and all the DMOZ-categories. Then we compute the Jaccard similarity between the daily read newspapers based on these counting and built hierarchy of clusters and generate four and six clusters for economic and political characteristics respectively.

The GitHub repository containing the scripts is available at https://github.com/cleopatra-itn/Trends_Clustering.

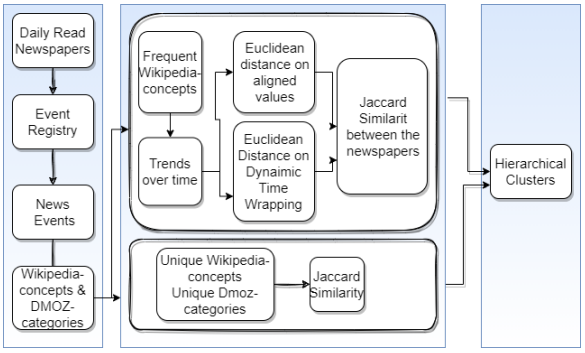


Figure 4: Methodology to create hierarchical clusters

5 Experimental Evaluation

5.1 Evaluation Metric

To provide insights in how the economic conditions and political alignment of newspapers may be reflected in the events they report on, we first generate hierarchical clustering of the newspapers using the proposed methodology. Then we cut the hierarchy in a way to obtain as many clusters to match the predefined number of economic groups or political groups. Then we manually compare the generated clusters with our ground truth clusters (see Figure 2) by calculating accuracy on the economic condition and on the political alignment.

Accuracy: Accuracy is the most intuitive performance measure and best to use when we have symmetric data set where values of false positive and false negatives are almost same.

We consider each newspaper as one instance. To calculate the number of correctly grouped instances, we compare the output clusters with our original clusters (see Figure 2) of economic and political characteristics. For instance, in case of economic characteristics, if output cluster consist of a cluster with the single newspaper “Zaman”, it means one instance is correctly grouped because there exists a cluster with same newspaper. Further, if output cluster contains one of the four clusters with the two newspapers “China Daily” and “Dawn”, it means one instance is correctly grouped and one is incorrectly grouped, because there is a cluster with the two newspapers “The times of India” and “Dawn”. Therefore, one is correctly grouped and one is incorrectly grouped. We use algorithm 1 to compare the output clusters with the original groups.

Algorithm 1 Count correctly grouped instances

Input: $outputCluster, originalCluster$
Output: (Number of correctly grouped instances)

```
Sum ← 0
for i ← 1 to len(outputCluster) do
    m ← 0
    tc ← outputCluster[i]
    for j ← 1 to len(originalCluster) do
        p ← set(tc) and set(originalCluster[j])
        if len( p ) > 0 then
            m ← len(p)
        end if
    end for
    Sum ← Sum + m
end for
return Sum
```

5.2 Results and Analysis

Figure 5 shows three hierarchies of clusters built upon three different similarity mechanisms. First two diagrams present the hierarchies of clusters with Dynamic Time Warping (DTW) and Euclidean distance. Whereas third diagram shows the hierarchy of clusters using Jaccard similarity between the unique Wikipedia-concepts. While we follow the same mechanism to choose the four and six hierarchies in all three diagrams, so we will explain only first diagram. Considering our original clusters (see Figure 2), we choose top four and six hierarchies for economic and political characteristics respectively. For example, using the first diagram in figure 5, we create the following four clusters to compare with our original economic cluster (see Figure 2):

- "Zaman"
- "Dawn" and "The Times of India"
- "The Asahi Shimbun"
- "The Guardian", "The Washington Post", "The Sydney Morning Herald", "The Wall Street Journal", "New York Times"

Similarly, we create the following six clusters also using the same figure 5 to compare with our original political cluster (see Figure 2):

- "Zaman"
- "Dawn" and "The Times of India"
- "The Asahi Shimbun"
- "The Sydney Morning Herald"
- "The Wall Street Journal", "New York Times"
- "The Guardian", "The Washington Post"

Table 3 shows the results in form of overall accuracy for both economic and political attributes. For economic attribute, firstly highest results are achieved using most-frequent Wikipedia-concepts (88.89%). Then second-best performing mechanism is Jaccard similarity with 80% accuracy, and then lastly Euclidean distance with 77.78% accuracy. For political attribute, the best performing mechanism is Euclidean distance and Jaccard similarity with 77.78%, and 70.0% accuracy.

Based on the overall accuracy for economic attribute, we can say that the representation with Wikipedia-concepts is better than DMOZ-categories for capturing economic characteristic of the newspaper, while the opposite is true for capturing political attribute. Furthermore, it can be noticed that similarity using DTW is more suitable for economic characteristic but not for political characteristic.

Table 3. Overall accuracy of each similarity measure for economic and political characteristics.

Features	Overall Accuracy (Economic)	Overall Accuracy (Political)
100 Most-frequent Concepts (DTW)	**88.89%**	40%
150 Most-frequent Concepts (DTW)	**88.89%**	44.44%
200 Most-frequent Concepts (DTW)	41.67%	62.5%
100 Most-frequent Concepts (Trends: Euclidean Distance)	44%	55.56%
150 Most-frequent Concepts (Trends: Euclidean Distance)	**77.78%**	55.56%
200 Most-frequent Concepts (Trends: Euclidean Distance)	**77.78%**	**77.78%**
All Categories (Jaccard Similarity)	50%	**70%**
Unique Categories (Jaccard Similarity)	50%	**70%**
All Concepts (Jaccard Similarity)	60%	**70%**
Unique Concepts (Jaccard Similarity)	**80%**	**70%**

6 Conclusions and Future Work

Newspapers have different characteristics such as political alignment, economic values, cultural differences, reporting languages and geographical differences. In this paper, we focused on to find representation of the news events for two characteristics (political and economic) separately that could be able to cluster the input of text in their original groups. We represent the news events with set of concepts and content categories separately, create hierarchical cluster and compare the output clusters

with original groups. Instead of just keywords, we also consider the trends of DMOZ-categories and Wikipedia-concepts. Moreover, we perform different similarity mechanisms (Dynamic Time Warming, Jaccard Distance, Euclidean Distance) before creating the clusters and see which mechanism is suitable for economic and political groups.

Our results (see Section 5.2) suggest that economic conditions of the country of the newspaper publisher reflect better with a set of concepts then content categories whereas content categories are more suitable for politically aligned groups of newspapers. Furthermore, results show that clustering using the Dynamic Time Warping similarity between trend lines of newspaper is better aligned with ground truth of economic groups and Jaccard distance using the frequent terms and Euclidean distance between the trend lines is more useful for ground truth of political groups.

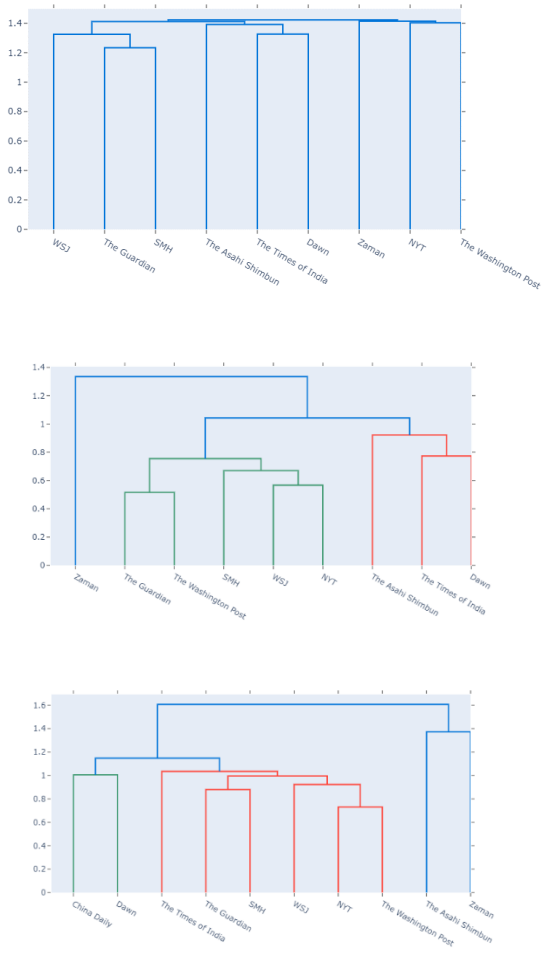


Figure 5. First two figures (from top to bottom) shows the hierarchical clustering with the DTW and Euclidean distance based on 100 most-frequent Wikipedia-concepts. Third figure presents the hierarchical clustering with Jaccard similarity between the unique Wikipedia-concepts.

Our research experiments only centered around economic and political attributes. In future, we would like to explore the other characteristics such as cultural, time-zone, geographical, and linguistic. We also have a plan to use advance tools of Natural language Processing (NLP) to detect the cultural differences in news events.

Acknowledgments

This work was supported by the Slovenian research agency under the project J2-1736 Causalify and co-financed by the Slovenian Research Agency and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 812997.

References

Afonso, Alexandre Ribeiro and Duque, Cl{\'a}udio Gottschalg (2014). Automated text clustering of newspaper and scientific texts in brazilian portuguese: analysis and comparison of methods. JISTEM-Journal of Information Systems and Technology Management, 11, 415--436.

Lin, Carolyn A and Jeffres, Leo W (2001). Comparing distinctions and similarities across websites of newspapers, radio stations, and television stations. Journalism \& Mass Communication Quarterly, 78(3), 555--57.

Veltri, Giuseppe Alessandro(2012). Information flows and centrality among elite European newspapers. European Journal of Communicationy, 27(4), 354--375.

Chang, Tsan-Kuo and Lee, Jae-Won(1992). Factors affecting gatekeepers' selection of foreign news: A national survey of newspaper editors. Journalism Quarterly, 69(3), 554--561.

Wilke, J{\"u}rgen and Heimprecht, Christine and Cohen, Akiba(2012). The geography of foreign news on television: A comparative study of 17 countries. International communication gazette, 74(4), 301--322.

Jo, HyunChae and Park, Cheolyong (2018). Analysis of reporting characteristics of newspapers in the 19th presidential election based on random forest. The Korean Data \& Information Science Society, 29(2), 367--375.

Wei, H., Sankaranarayanan, J., & Samet, H. (2020). Enhancing local live tweet stream to detect news. GeoInformatica, 1-31.

Hong, X., Yu, Z., Tang, M., & Xian, Y. (2017). Cross-lingual event-centered news clustering based on elements semantic correlations of different news.

Multimedia Tools and Applications, 76(23), 25129-25143.

Leban, G., Fortuna, B., Brank, J., & Grobelnik, M. (2014, April). Event registry: learning about world events from news. In Proceedings of the 23rd International Conference on World Wide Web (pp. 107-110).

Quezada, M., Pena-Araya, V., & Poblete, B. (2015, August). Location-aware model for news events in social media. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 935-938).

Bhowmik, D., & Poddar, S. (2021). Cyclical and seasonal patterns of India's GDP growth rate through the eyes of Hamilton and Hodrick Prescott Filter models. Asia-Pacific Journal of Management and Technology, 1(3), 7-17.

Heywood, A. (2017). Political ideologies: An introduction. Macmillan International Higher Education.

Strle, B., Mozina, M., & Bratko, I. (2009, June). Qualitative approximation to Dynamic Time Warping similarity between time series data. In Proceedings of the Workshop on Qualitative Reasoning.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013, March). Using of Jaccard coefficient for keywords similarity. In Proceedings of the international multiconference of engineers and computer scientists (Vol. 1, No. 6, pp. 380-384).

Segev, E. (2015). Visible and invisible countries: News flow theory revised. Journalism, 16(3), 412-428.

Wu, H. D. (2007). A brave new world for international news? Exploring the determinants of the coverage of foreign news on US websites. International Communication Gazette, 69(6), 539-551.

Filla, J., & Johnson, M. (2010). Local news outlets and political participation. Urban Affairs Review, 45(5), 679-692.

Schulz, A., Fletcher, R., & Popescu, M. (2020). Are news outlets viewed in the same way by experts and the public? A comparison across 23 European Countries. Reuters Institute for the Study of Journalism.

Camaj, L. (2010). Media framing through stages of a political discourse: International news agencies' coverage of Kosovo's status negotiations. International Communication Gazette, 72(7), 635-653.

Martens, B., Aguiar, L., Gomez-Herrera, E., & Mueller-Langer, F. (2018). The digital transformation of news media and the rise of disinformation and fake news.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. Science, 348(6239), 1130-1132.

Maurer, P., & Beiler, M. (2018). Networking and political alignment as strategies to control the news: Interaction between journalists and politicians. Journalism Studies, 19(14), 2024-2041

3.2 Enhanced Topic Modeling

Topic modeling is designed to analyze a corpus of text and extract latent themes or topics. The results are more coherent if the text is semantically similar. We propose and evaluate the enhanced Topic Modelling approach that uses LDA with a combination of 1-6 grams and articles' pooling based on queries to improve the quality of topics without modifying the basic structure of LDA. This section presents the results of the paper titled *Political and Economic Patterns in COVID-19 News: From Lockdown to Vaccination* that was published in IEEE Access in 2022 [62]. The authors of this paper are Abdul Sittar, Daniela Major, Caio Mello, Dunja Mladenić, and Marko Grobelnik.

We identify the main topics related to COVID-19 and construct five queries to pool news articles for each month in the period between January 2020 and May 2021. These five queries are: 1) Lab leak theory, 2) Efficacy of vaccines, 3) Lockdown policies and efficiency, 4) Seriousness of Coronavirus, and 5) Can masks protect against COVID-19? We evaluate the topic modeling approach based on the coherence score which shows that pooling of news articles can improve the quality of the topic. We saw that the coherent score of LDA topics was always higher for all the queries with pooling than without pooling.

3.2.1 Case Study COVID-19

With a rapidly growing number of events with significant international impact, understanding the news reporting has increased importance for researchers and professionals in many disciplines, including digital humanities, media studies, and journalism. Also, analysing the influencing factors on news reporting and spreading is an open research area. This work focused directly on studying insights in reporting differences in different political and economic contexts. We propose a methodology that can be adopted on news of different types of events to understand the effects of different political and economical context's on news reporting. We answer the following two research questions in this study: 1) How political and economic issues propagated over time during the pandemic across different socio-political and economic contexts 2) how different discussions evolved during different stages of COVID-19 epidemic?

The results showed that the extracted topics correspond to the chronological development of the pandemic and the measures that were under discussion in different countries and across different economies and political alignments. Also, the results across different political and economical conditions are consistent.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 15, 2022, accepted March 31, 2022, date of publication April 4, 2022, date of current version April 19, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3164692

Political and Economic Patterns in COVID-19 News: From Lockdown to Vaccination

ABDUL SITTAR[1], DANIELA MAJOR[2], CAIO MELLO[2], DUNJA MLADENIĆ[1], AND MARKO GROBELNIK[1]

[1]Department for Artificial Intelligence, Jozef Stefan Institute, 1000 Ljubljana, Slovenia
[2]School of Advanced Study, University of London, London WC1E 7HU, U.K.

Corresponding author: Abdul Sittar (abdul.sittar@ijs.si)

This wok was supported in part by the Slovenian Research Agency under Project J2-1736 Causalify, and in part by the European Union's Horizon 2020 Research and Innovation Program through the Marie Skłodowska-Curie under Grant 812997.

ABSTRACT The purpose of this study is to analyse COVID-19 related news published across different geographical places, in order to gain insights in reporting differences. The COVID-19 pandemic had a major outbreak in January 2020 and was followed by different preventive measures, lockdown, and finally by the process of vaccination. To date, more comprehensive analysis of news related to COVID-19 pandemic are missing, especially those which explain what aspects of this pandemic are being reported by newspapers inserted in different economies and belonging to different political alignments. Since LDA is often less coherent when there are news articles published across the world about an event and you look answers for specific queries. It is because of having semantically different content. To address this challenge, we performed pooling of news articles based on information retrieval using TF-IDF score in a data processing step and topic modeling using LDA with combination of 1 to 6 ngrams. We used VADER sentiment analyzer to analyze the differences in sentiments in news articles reported across different geographical places. The novelty of this study is to look at how COVID-19 pandemic was reported by the media, providing a comparison among countries in different political and economic contexts. Our findings suggest that the news reporting by newspapers with different political alignment support the reported content. Also, economic issues reported by newspapers depend on economy of the place where a newspaper resides.

INDEX TERMS COVID-19, economic issues, political issues, sentiment analysis, topic modeling.

I. INTRODUCTION

We can say that news coverage directly catalogs the occurrence of specific events and indicates the local as well as global opinions of stakeholders. As the epidemic took over the world, news coverage became a significant source of information that allowed populations to adapt their behaviours and fight off the disease. Because of this, COVID-19 has been extensively reported by Global and local media [1]. Analysing the multi-facet and spatio-temporal aspects of the news coverage of the pandemic is essential for a clear understanding of this event, as news contributes to the way people understand the world. It is a 'place of reference' [2] where people 'go' in search of stability. In order to understand the evolution of the coverage of this pandemic, we propose a methodology in which we employ variants of popular techniques of Natural Language Processing (NLP) such as Topic Modeling (TM) and Sentiment Analysis (SA). Topic models are designed to analyze a corpus of text and extract the latent themes or topics. The results are more coherent if the text is semantically similar. Therefore, the motivation behind our proposed method is to pool the news articles based on user queries to extract the most relevant latent themes without modifying the basic structure of LDA. Previous studies used pooling based on other parameters. Mehrotra *et al.* applied pooling based on hashtags on twitter datasets [3], while [4] proposed a scheme for pooling tweets into longer documents based on conversations. However, the problem of pooling based on user queries is not explored for news articles. To fulfil this gap, the present study aims to see the coherence score differences with and without pooling based on user queries. SA is used to check positive or negative sentiments within a text. Among famous sentiment analyzers (Ratio, TextBlob, NRC Emotion Lexicon, VADER), we use VADER to extract sentiments. The motivation behind finding the sentiments is to provide information on how the media

The associate editor coordinating the review of this manuscript and approving it for publication was Liviu-Adrian Cotfas.

approached each phase of the pandemic. The identification of sentiments discursively expressed in the narration of this event offer insights on how different outlets interpreted the different dimensions of the pandemic such as the lockdown, the protective measures, the research to develop vaccines and their distribution over time.

Numerous studies have investigated the impact of COVID-19 in different countries. Reference [5] investigates how different discussions evolved over time and the spatial analysis of tweets. Reference [6] addresses the diffusion of information about the COVID-19 using a large amount of data from popular social media networks. Reference [7] performs SA in the early stages of the COVID-19. These studies, however, do not conduct experiments on large timespans that may provide a better overview of the pandemic. Secondly, most studies employ Twitter content, and when news are used they are either limited to one or two newspapers or belong to the same country.

This study identifies how the discussions evolved over time in top newspapers belonging to three different continents (Europe, Asia, and North America) and nine different countries (UK, India, Ireland, Canada, U.S., Japan, Indonesia, Turkey, and Pakistan). It uses spatio-temporal TM and SA. TM will be used to determine the topics of discussion especially pertaining to different economic and political perspectives, while SA will be used to determine the change in sentiments over time. Identifying these topics and emotions could help newspapers to improve how they communicate information and provide data which would enable governmental bodies review their communication strategies.

The remainder of this paper is structured as follows: Section II provides related work on spatio-temporal analysis, SA and data description about COVID-19. After elaborating on the research methodology in Section III, the results are described in Section IV. Section V provides a brief discussion about findings and the corresponding results. Finally, Section VI presents the conclusion and some ideas for future work.

A. MOTIVATION

The motivation behind our work are stemmed from the following facts:

- With a rapidly growing number of events with significant international impact, understanding the differences in news reporting has increased importance for researchers and professionals in many disciplines, including digital humanities, media studies, and journalism. The most prominent recent examples of such events include the migration crisis in Europe, COVID-19 outbreak, and Brexit.
- There are many factors that influence the news selection, reporting, and spreading such as cultural, political, economic, geographical, and linguistic. Analyzing these factors in news spreading related to different international events is open research area. Since COVID-19 news has many conspiracies and fake information attached with it, and it has major effects on different economies, we take into account only two factors political and economic.
- Nowadays, social scientists and psychologist are interested to know how the COVID-19 pandemic has influenced the everyday life of people in the world. It has raised the issues including unemployment, stress and depression due to the lockdown, and inflation and so many other issues. In general, newspapers report overall situation of a specific area. Analyzing the sentiments in news articles can help to see the changes in sentiments over time and to make comparison of sentiments across different countries.

B. CONTRIBUTIONS

The following are the contributions of this study:

- We propose and evaluate the enhanced Topic Modelling approach that uses LDA with combination of 1-6 grams and articles' pooling based on queries to improve the quality of topics without modifying the basic structure of LDA.
- We propose a methodology to understand political and economic differences in news reporting using TM.
- We study the comparison of sentiments between newspapers across different geographical areas.

C. RESEARCH QUESTIONS

This paper is the first of its kind to analyse the news related to COVID-19 based on political alignment and the economic situation of different countries from January 1st, 2020 to May 31, 2021. Our hypothesis states that the topics present in news related to the COVID-19 pandemic vary according to the publisher's political alignments and the economic situation of a specific area. Moreover, the topics that have strong relation with each other will have similar trends over time. We used TM, and SA to answer the following research questions:

Q1: How political and economic issues propagated over time during the pandemic across different socio-political and economic contexts?

Q2: How different discussions evolved during different stages of the COVID-19 epidemic?

Q3: What are the patterns of emotional states during different stages of the pandemic across different countries?

In our first research question, we refer ''political issues propagated over time'' as political issues that have spread/reported over time. Similarly, ''economic issue propagated over time'' refers to the economic issues that have spread/reported over time. In our second research question, ''discussions'' refers to different topics that evolved during different stages of the COVID-19 epidemic. In our third research question, ''emotions'' refers to sentiments over time.

II. RELATED WORK

COVID-19 is a broad topic and has enormous number of research dimensions. As this study focuses on analyzing the

spatio-temporal political and economic patterns, and sentiment analysis, we review six different types of related works: data for COVID-19 analysis, spatio-temporal analysis, COVID-19 analysis using social media, characteristics of the newspapers, sentiment analysis, and topic modeling.

A. DATA FOR COVID-19 ANALYSIS

Analytical studies regarding COVID-19 have conducted research to portray its emerging effects in different fields using popular analytical methods and data from different sources such as Nigeria Centre for Disease Control (NCDC) [8], Facebook [9], News from the New York Times (United States of America) and Global Times (China) [10], Twitter [11], [12], News from China National Knowledge Infrastructure (CNKI) database [13]. The timeline for all this data coverage is only a few months (two, three, or the first few months of the pandemic). Reference [14] has studied general issues reported in news media. To the best of our knowledge there is a lack of studies that cover the complete phase of the pandemic from lockdown to the vaccination (from January 2020 to May 2021) by using news and that find the issues focusing on different political and economic contexts.

B. SPATIO-TEMPORAL ANALYSIS

Spatio-temporal analysis is used to uncover the relations between locations over time [15]. Reference [16] identify spatial effects and spatio-temporal patterns of the outbreak of COVID-19 in different regions of Italy. Reference [17] identifies seasonality in disease in Spain due to variations in temperature, humidity, and hours of sunshine. References [18], [19] study the spatio-temporal propagation of the first wave of the COVID-19 virus in China and compare it to the other global locations in terms of distance, population size, and human mobility and their scaling relations. To our knowledge, there is no study which focus on top read newspapers. And these newspapers belong to nine different countries which belong to three continents.

C. COVID-19 ANALYSIS USING SOCIAL MEDIA

Social media is one of the major sources to understand the societal and crowd response nowadays [12]. One of the reasons is its widespread growth. However, the consumption of news articles through news publishers is associated positively with higher trust whereas the information related to the social media is linked with lower trust [20]. Analyzing and linking emerging events to relevant social issues is an increasingly important task. Most of the focus has been on social networks when studying spatio-temporal effects of the COVID-19 Pandemic. News is also one of the most important sources containing detailed information [21]. There is a fundamental problem attached to news articles which relates to the unstructured and noisy nature of data.. Several studies employ news but are limited to one or two countries. Reference [22] measure depressiveness, and informativeness for various states in the US. Reference [10] take news from the New York Times (United States of America) and Global Times (China) to study the way the news used for political and ideological purposes. To our knowledge, there is space for research studies which conduct COVID-19 spatio-temporal analysis by utilizing news published around the world (including countries from different continents).

D. CHARACTERISTICS OF THE NEWSPAPERS

International news led us to investigate the reasons why specific event-centric news either spread or do not spread to certain geographic area. Based on some factors, media target specific foreign and regional events. For example, spreading of news related to specific events may tilt toward developed countries such as United Kingdom, U.S.A, or Russia. Furthermore, it may be due to geographical juxtaposition (latitude, longitude) of countries [23]. There is great deal of negotiation between political actors and journalists in news production to enhance their influence on news coverage. Fake news also produced based on many factors and surrounded by a dominant element that is political effect [24]. Thus, political alignment of news publishers also impact the coverage of different events. One of the determinants for news coverage is economic conditions that also impact information selection, analysis, and propagation [25]. Generally, news reporting about different events (elections etc.) is inclined towards certain characteristics of newspapers. As there is a inclination to back underground and indirectly, the research interest in the reporting characteristics of each newspaper has begun [26]. Filla investigates the political participation by the local news outlets in elections and find the relationship between the political participation and availability of local news outlets [27]. News agencies tend to follow the national context in which journalists operate. One of the related examples is the SARS epidemic study which found that cross-national contextual values such as political and economic situations impact the news selection [28]. A great amount of work regarding fake news dwells on different strategies, while few studies considered political alignment to have a compelling effect on news spreading [29]. With a rapidly growing number of events with significant international impact, understanding the news reporting has increased importance for researchers and professionals in many disciplines, including digital humanities, media studies, and journalism. The most prominent recent examples of such events include the migration crisis in Europe, COVID-19 outbreak, and Brexit. Although the previous work involves relationship between outlets and political activities or public interest [30], our work focused directly on studying insights in reporting differences in different political and economic contexts.

E. SENTIMENT ANALYSIS (SA)

SA is used to check different levels of positive or negative opinions within a text. It is useful to determine the emotional state expressed in news articles in response to the outbreak.

Reference [5] performed SA on tweets belonging to different countries and generated time-series plots to see whether the spikes in positive or negative sentiment can be associated with certain events. Another study develop a Recurrent Neural Network (RNN) for predicting emotions using tweets and compare the model with TextBlob [31]. Some studies find country wise sentiment analysis using R [32], NRC Emotion Lexicon [33]. Reference [12] highlight the concerns of Gulf countries' people to lessen the vaccine hesitancy. It uses three different methods (Ratio, TextBlob, and VADER) to extract the sentiments. Working with SA implies a range of challenges to obtain accurate results such as the analysis of negation, sarcasm and ambiguity [34]. These limitations were considered when making use of this algorithm. The technique is used as a exploratory tool to produce insights of the data rather than a conclusive method.

Reference [35] use a method that connects the polarity scores to emotional states with the use of Emotional Guidance Scales. This scale from -1 to 1 contains 11 emotions where each emotions change with increments of 0.2. -1 denotes the most depressed and fear feeling and +1 denotes emotion of being happy and joyful. This brings up the interesting problem of spatio-temporal SA from news articles. The objective is to discover sentiment on news articles on the COVID crisis at country level over temporal intervals of a month using a large collection of news articles from a period starting from January 2020 to May 2021.

F. TOPIC MODELING (TM)

Several studies used TM to determine the topics of discussion about COVID-19. The intention was to extract popular topics and understand evolution of different discussions [36]–[38]. LDA is used to infer topics from the collection of text-document. Some techniques used only frequent words whereas some use pooling to generate relevant topics and maintain coherence between topics [5]. To pool the relevant documents several mechanisms have been used. Unlike simple static TM in this modeling it is assumed that the data is partitioned on a time bases (e.g. hourly, daily, monthly, or yearly). In fact, the order and arrangement of documents reflects the evolving set of topics [39]. These pooling techniques are famous for social media where set of documents/tweets are partitioned based on hashtags, and authors, etc. [3]. Carmela combines peak detection and clustering techniques to identify topics. Space-time features are extracted from tweets and modeled as time series. Using these time series peaks are identified and clustered based on co-occurrence in the tweets [18]. Innovative approaches have been proposed to detect spatio-temporal topics. The problem of topic identification over spatio-temporal analysis is also considered as problem of stream clustering. Tweet-based clustering method used the content, structural and temporal information, Hashtag-Based clustering focused on hierarchical spatio-temporal techniques which explore different event with different time granularity [40]. Information retrieval is understood as an automatic process that respond to a user and returning a list of documents that should be relevant to the user query [41]. In this process, terms (which appear in the queries and set of documents) are ranked using some weighting techniques. One of the popular methods is TF-IDF weight. This is a statistical measure used to evaluate how important a word is in a document in a dataset [42]. Data provided by social media can be pooled based on meta data (e.g. hash tags in Twitter) but for news articles this type of pooling does not exist. Therefore, we pool the news articles based on user queries. LDA is used to discover underlying topics directly from the raw text features in the documents. When it is applied directly on raw texts then result in topics are uninformative and hard to interpret [4]. Pooling of raw text into groups appeared appealing with vast improvements in results [43]. However, these methods only applied on data which has structured form of schema such as Social Networks (SN). The information in SN has attractive options such as share or retweet and information makes cascading structure, which means one can easily group the information based on the users, hashtags, timestamp and location. Contrary to SN, there is no such structure exist in the case of news articles. Therefore, to overcome this issue [14] uses Top2Vec algorithm and Doc2Vec to place news articles close to other similar news articles. The limitation of this study is that it requires to input all the news articles at the same time while finding the topics. Although the methods used are efficient, the results would change if news articles grouped based on user queries. Reference [44] used topic modeling to depict the overall picture of Taiwan's pandemic where they divide the data into different stages based on temporal information. They divide the data into four stages based on the development of COVID-19. They have also used basic LDA technique and the quality of topics is compromised as the news articles are not filtered based on queries or general themes. Our study provides the basis for improvement of the results if we group the news articles based on queries.

III. METHODS

This study adopts an approach that integrates computational as well as qualitative techniques to explain perceptions and attitudes towards COVID-19. The present research focuses on spatio-temporal analysis of different discussions: news events and governmental decisions, cures and prevention measures, political and economic discourses, and emotions. Fig. 1 shows the experimental methodology adopted in this work. Several steps were performed sequentially in the workflow, including data collection, data preparation, filtering news articles, TM, topic analysis, analysis of political and economical issues, and SA. This methodology can be adopted on news of different type of events to understand the effects of different political and economical context's on news reporting. The following sections present the data collection and preprocessing, TM, topic visualization, and SA in detail.

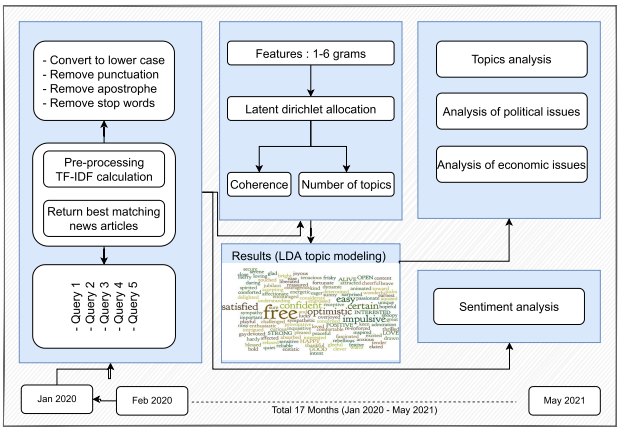


FIGURE 1. Workflow to identify patterns of different political and economic issues as well as SA. Data preprocessing for five different queries for each month (sequentially) is the initial step. TM using combination of 1-6 grams is the second step. The last step is to identify the most frequent political and economic issues. In case of SA, the body text of news article temporally pass to the sentiment analyzer.

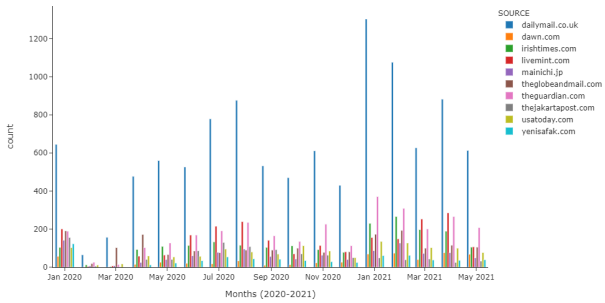


FIGURE 2. Total number of news articles published in each month (01/2020- 05/2021) by ten newspapers. Main purpose of the line graph is to show the variations in collected number of news articles reported by top ten newspapers during the set timeline.

A. DATA COLLECTION AND PREPROCESSING

The collected dataset of news articles is based on ten newspapers.[1] These newspapers were selected based on the pre-conditions:

- At least a few news articles must be published by a newspaper for each month (January 2020 to May 2021),
- Newspapers should belong to different political alignments,
- The newspapers should belong to countries from different economic backgrounds.

Following these preconditions we constructed a dataset which consists of 24,000 news articles published by ten newspapers. Figure 2 shows the number of articles published by each newspaper in each month. Figure 3 shows coverage of news articles per country, continent and political alignments. The data was collected by using Event Registry platform. This platform identifies events by collecting related articles written in different languages from tens of thousands of news sources [45]. Using Event Registry APIs,[2] we fetched news articles in English published by the selected newspapers. We use the following keywords to search news articles: COVID-19, Coronavirus, COVID pandemic, and COVID Outbreak. Each news article consists of a few attributes: title, body text, name of the news publisher, time of publishing. We estimate the political alignment of each newspaper using Wikipedia info-box as already done in [15]. We perform a number of preprocessing steps: conversion to lower case, removal of punctuation marks and removal of stop-words. It is important to mention that there were other newspapers (see the link[3]) that were unable to follow the preconditions. Some newspapers stood in same position in political spectrum. Some newspapers belonged to countries with same economic backgrounds. For some of the newspapers, we were unable to find news articles from Event Registry platform for few of the months in the set time period (from January 2020 to May 2021). Therefore the newspapers reduced and as a result the news articles also reduced to few thousand (24,000) news articles. Another significant information that we can see in Fig. 3 is the varying number of news articles by all newspapers. For example, in case of UK based newspaper Daily Mail, we see there is large collection of news articles in each month, whereas for U.S.A, there is only one newspaper with approximately five percent of news articles. This is because of our preconditions. If we remove this potential bias then the main goal of this methodology would not be achieved that is understanding news reporting across different political and economic contexts.

B. TOPIC MODELING (TM)

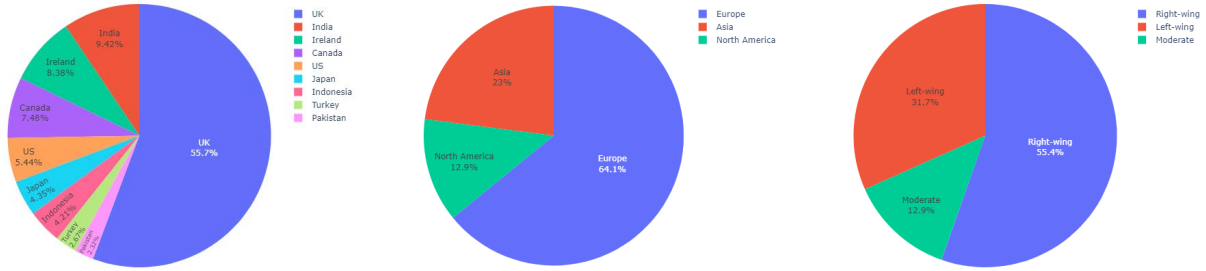
We identify main topics related to COVID-19 and construct five queries to pool news articles for each month in the period between January 2020 and May 2021. These five queries are: 1) Lab leak theory, 2) Efficiency of vaccines, 3) Lockdown policies and efficiency, 4) Seriousness of Coronavirus, and 5) Can masks protect against COVID-19.

Previous studies conducted research on most critical topics related to COVID-19 pandemic ([46]–[50]). We construct queries following these famous topics. Since these are the most researched topics related to COVID-19, we select them to see the reporting differences across difference newspapers. To pool relevant news articles, we perform filtering based on each query for each month by calculating TF-IDF score of unique words for all news articles. During the filtering process, for individual newspaper and for each query on the time period of one month, we take top matching news articles (ten percent if total news articles are greater than hundred, ten if total news articles are less then hundred and greater than ten, all news articles if total news articles are less than 10). We perform the same preprocessing steps on each query, as we did on the news articles: conversion to lower case,

[1] https://www.4imn.com/top200/

[2] https://github.com/EventRegistry/event-registry-python/blob/master/eventregistry/examples/QueryArticlesExamples.py

[3] https://www.4imn.com/top200/



a). Percentage of news articles by nine countries.

b). Percentage of news articles by three continents.

c). Classification of newspapers with respect to political spectrum.

FIGURE 3. Data distribution is not equal in all three cases (country wise, continent wise, and with respect to political spectrum). Statistics about newspapers w.r.t. their geographical location of headquarters and political alignments.

removal of punctuation marks and removal of stop-words. Then for each token from the query we calculate TF-IDF score for each news article and then sort the news articles based on the sum of TF-IDF score of these tokens. Then we take into account the top news articles. For instance, we have a query "Can mask protect against corona virus?." We perform preprocessing such as conversion to lower case, removal of punctuation marks and stop-words. The final tokens for this query are four e.g., "mask," "protect," "corona," "virus." We sum the TF-IDF scores for these tokens in each of the news articles. Based on the outcome, we sort the news articles and take top news articles as relevant for the query. A method that is used for finding the abstract topic in a large collection of documents is called TM. LDA is a thematic probabilistic modeling algorithm that takes into account both words and documents while capturing the topics. We call a set of topics coherent if they support each other and cover most of the facts in a set of documents. There are many coherence measures (C_v, C_p,C_uci, C_umass, C_npmi, C_a) and the way they are calculated is different. We use only C_c measure. It is based on sliding window. It basically performs one-set segmentation of the top words and an indirect confirmation measure that finds normalized point-wise mutual information (NPMI) and cosine similarity. We compare coherence score of each individual query with query and without queries along with simple uni-gram representation for each newspaper. We see that content is more coherent with queries (see Figure 4). Also, we find an optimal number of topics for all queries and for all newspaper (see Figure 4). We use a combination of 1 to 6 ngrams when performing TM on news articles.

C. TOPIC VISUALIZATION

We filter political and economic topics for each query manually. For each type of political alignment we put together all the filtered topics and show them in word clouds. Figure 5 shows the word clouds for different political alignments along with all the queries. Figure 6 shows the word clouds for different economic levels along with all queries. There were different number of filtered news articles for each newspaper that are used to create Figures 5, 6 and 7. More specifically, for each query, there were 615, 308, 227, 209, 355, 488, 131, 439, 112, 2125 news article filtered for Dailmail.co.uk, Dawn.com, Irishtimes.com, livemint.com, mainichi,jp, theglobeandmail.com, theguardian.com, thejakartapost.com, usatoday.com, yenisafak.com respectively. For each query the number of filtered news articles was the same within the same newspaper, but the news articles were different following the process of filtering that take queries into account. To observe the topics which had similar trend over time, we use topics that have been filtered for different political alignments and economic levels. These trends are based on frequency of these topics over time. Afterward, we generate the trend lines using the Hodrick Prescott filter [51].

D. SENTIMENT ANALYSIS (SA)

Sentiment in news is expressed in a variety of forms, from interviewers quotes to journalists choices to use one term instead of another. This becomes clear when words such as fear are used in the headlines "Coronavirus Spreads Fear." There was a choice of using "fear," which is a word that carries negative sentiment in Vader dictionary. The sentiment of each news articles was classified using VADER Sentiment Analyzer. It is a rule-based SA tool and a lexicon which is used to express sentiments in social media [52]. In the analysis, we only take body text of news articles into account. Figures 8, 9, and 10 depict the comparison of sentiments across three continents (Asia, Europe, and North America), utilizing the news articles. The granularity of sentiment is limited to days.

IV. RESULTS

The aim of proposing the enhanced TM approach was to improve the quality of topics without modifying the basic structure of standard LDA. For that we evaluate the model based on coherence score which shows that pooling of news article can improve the quality of topic (see Figure 4) but it also help to answer customized research questions (see Section IV-A, IV-B). Figure 4 shows the comparison of coherent score of LDA with pooling and without pooling.

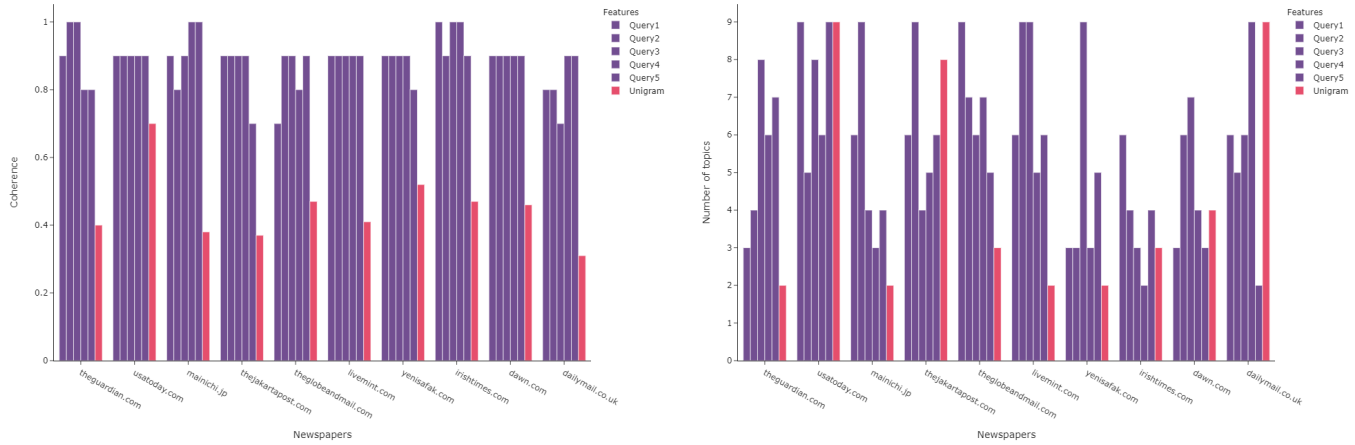


FIGURE 4. Coherence score increases and number of topics also varies with pooling. Comparison of pooling with simple uni-grams based on C_V Coherence measure (left-hand side) and the number of topics (right hand-side). The left bar chart compares the coherence measure score between pooling with combination of 1-6 grams and simple uni-grams without pooling. Coherence score is always high in case of pooling than without pooling. The right bar chart compares the best number of topics using 1-6 grams with pooling and without pooling along with simple uni-grams. Number of topics also vary for each newspaper for all queries.

There are six bars for each newspaper. Red line shows the coherent score without pooling and other lines show coherent score with pooling for our five queries (see Section III-D). We can see that coherent score of LDA topics is always higher than 0.7 for all queries in case of pooling, whereas in case of without pooling and without queries, the coherent score of LDA topics is always lower than 0.7 for each newspaper. The second bar chart in figure 4 shows the variations in number of LDA topics with pooling and without pooling. There are six bars for each newspaper. Red line shows the number of topics for without pooling and other lines show coherent score with pooling for our five queries (see Section III-D). It shows that for seven newspapers the number of topics increased with pooling for all the queries whereas only three newspapers have higher number of topics without pooling. Overall, It shows that with pooling for each query the coherent score of LDA topics is high and number of LDA topics also increased. The aim of this spatio-temporal analysis is to answer the three research questions regarding COVID-19, that we have described in Section I-C. For the first research question related to spreading of political and economic issues, we observe the frequency of words in news articles and visualize it using word clouds containing the frequent topics across different political alignments and different levels of economies. Analysing the frequent topics across different political alignments can help us to find correlations between topics and political alignments. Therefore it might be possible to see whether political alignment is associated with particular effect on news spreading related to COVID-19 or not. Analyzing the frequent topics across different levels of economic prosperity can help us to see the correlations between certain topics and different levels of economic prosperity.[4] Therefore, it might be possible to see whether a country's economic situation has a certain effect on news spreading related to COVID-19 or not. For the second research question related to the discussions development over time, we compare trends of different political and economic terms and present as line graphs (see Figure 7). Looking at trends of words over time can help us understand how major political and economic topics evolved together. As a result we can see if there is any semantic relation between those topics or not. To answer the third research question we created a line graph for each newspaper based on sentiment score (see Figures 8, 10, 9). Sentiments over different geographical areas help to compare and rank the emotional states expressed through news articles. Moreover, looking at the overall stance of reporting by a newspaper during a specific time period requires an explanation about why at this time the reporting is positive or negative. Therefore we also identified a list of common and most frequent topics shown in Table 3.

[4]https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

A. POLITICAL ISSUES:

Figure 5 shows the word clouds of main and common topics discussed in newspapers related with different political alignments. The findings have been summarized below:

1) LAB LEAK THEORY

Left wing including liberal views have been found when discussing the origin and emergence of virus, such as seafood and laboratories whereas moderate and right wing newspapers appeared to focus more on conspiracy and misinformation about virus. The conservative Islamic newspaper discuss symptoms and experiments to combat the virus.

2) EFFICACY OF VACCINES

The newspapers representing five different political views haves shown different points of view on vaccine efficacy. Left wing newspapers talk about the after-effects of vaccination

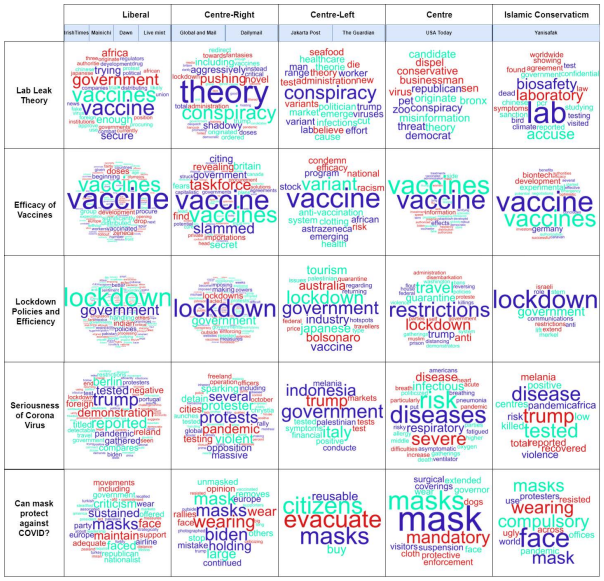


FIGURE 5. Word clouds showing the keywords appearing most frequently related to five queries in relation to different political alignments.

and raise questions about the efficacy of vaccines. Liberal newspapers forecast the rollout of vaccination and write about different vaccines along with age-group vaccination and their efficacy level. They also mention shipment of stocks and their distribution.

3) LOCKDOWN POLICIES AND EFFICIENCY

In the case of lockdown policies, left wing newspapers are found discussing restrictions and the effects of lockdown on different fields such as tourism and industry, whereas right wing newspapers are found discuss anti-lockdown protests.

4) SERIOUSNESS OF COVID-19 VIRUS

In the case of the seriousness of COVID-19, left wing newspapers published news discussing symptom identification and the financial consequences of the pandemic, whereas liberal newspapers mention the challenges of this disease, such as worldwide confinement. They also mention the required equipment to fight off the disease such as ventilators. Moderate newspapers, on the other hand, appeared to have entirely focused on the disease symptoms, such as respiratory and heart problems and the need for oxygen. Right wing newspapers reflect topics related to the issues faced by the general public as a consequence of lockdown, including shut down places, anti-lockdown rallies and pandemic forecastings.

5) CAN MASK PROTECT AGAINST COVID?

When it comes to mask protection, left wing newspapers discuss reusable masks. Liberal newspapers discuss mandatory face masks requirements and social distancing at airports. Moderate newspapers talk about a variety of masks, such as cloth masks. Right wing newspapers publish news about changes in policy in Europe when it comes to face masks. For example, how vaccination changes the mandatory requirements for mask wearing. Islamic conservative newspapers resist compulsory mask wearing.

B. ECONOMIC ISSUES:

Figure 6 shows the word clouds of main and common topics discussed in newspapers belonging to different economies. Each topic has been discussed in different economies with different results. The findings have been summarized below:

1) LAB LEAK THEORY

High income countries mention the topics theory, practice, pitfalls, scheme in connection to the lab leak theory, which might indicate that they engage in discussions about conspiracy theories. As middle income countries benefit from trade and investments made by high income countries topics such as funds, trade and investors appeared with high frequency. Generally, lower income countries struggle financially so the content published and discussed in the news is more related to funding, pays, pricing, therefore topics such as farmers, banking, and pricing appeared as most frequent.

2) EFFICACY OF VACCINES

The terms that appeared amongst low income countries are vaccine unions, cost and potential marketing. It shows that low income countries are potentially more concerned about affording the vaccine. In the case of middle income countries, we see economy, trade, vaccine roll-out and administration costs were more prevalent topics, which shows, as previously said, that middle income countries are more concerned with trade. As high income countries such as U.S and China were more involved in vaccine production, we see topics related to vaccines such as number of doses, stocking the vaccine and large task forces.

3) LOCKDOWN POLICIES AND EFFICIENCY

We see that all countries were discussing issues that occurred as a result of lockdown, such as inflation, damage to the economy and the tourism sector. We see three different types of discussions for low income, middle income and high income countries respectively: 1) lockdown effects on labour, remittances, oil prices, 2) job market crisis, fragile economy and lack of trade, and 3) quarantine issues, protests and hotspots lockdown issues.

4) SERIOUSNESS OF COVID-19 VIRUS

Terms that appeared in newspapers from high income countries show that these countries declare COVID-19 as a Global pandemic. Newspapers from middle income countries report that COVID-19 has reduced business activities (less travellers), which has resulted in unemployment. They also report shortages of medical staff as a result of contact with COVID patients. On the other hand, newspapers from low income countries report that COVID-19 has severally effected the remittances, earnings of families which results in difficulty accessing essential goods.

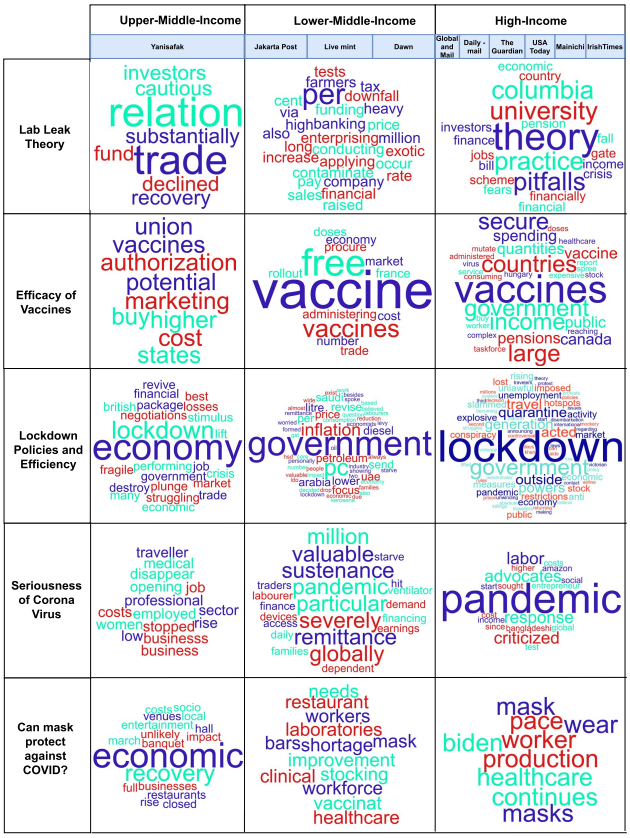


FIGURE 6. Word clouds showing the keywords appearing most frequently related to five queries in relation to different economies.

5) CAN MASK PROTECT AGAINST COVID?

Low income countries' newspapers discuss the usage of masks at places like restaurants, bars, and laboratories. They also discuss mask-shortage issues. News from middle income countries present keywords such as entertainment halls, banquet halls, and restaurants. On the other hand, news from high income countries mention the benefits of mask use.

C. DISCUSSIONS EVOLVED DURING DIFFERENT STAGES OF COVID-19:

The results showed that the extracted topics correspond to the chronological development of the pandemic and the measures that, according to our research queries, were under discussion in different countries and across different economies and political alignments. Figure 7 shows the trends of different economic and political topics. We have identified nine different groups of topics that in our data had similar trends over time: 1) Restrictions (lockdown), oxygen, ventilators, Allergies, Reusable masks, 2) Protests, rally, 3) Pneumonia, dose, rollout, 4) Effects, origin, emerge, lab, misinformation, breath, and identify, 5) Unemployment, scheme, labor, bars, restaurants, protests, 6) Trade, investors, price, stock, 7) Farmers, earning, banking, hot spot, 8) Travellers, rollout, and 9) Practice, fund, marketing, hall. By observing the trends of the discussion topics, we see that some topics are present to some extent over time, such as restrictions and ventilators. There are some other topics that where very frequent

TABLE 1. This table shows the predominant sentiments (Negative, Positive, Fluctuation) for each newspaper during different quarters (2020–2021).

Newspapers	1/2020 - 4/2020	5/2020 - 8/2020	9/2020 - 12/2020	1/2021 - 5/2021
Dawn.com	Negative	Negative	Positive	Fluctuation
Thejakartapost.com	Negative	Fluctuation	Positive	Fluctuation
Livemint.com	Negative	Positive	Positive	Positive
Mainichi.jp	Negative	Fluctuation	Positive	Fluctuation
Yenisafak.com	Fluctuation	Negative	Fluctuation	Fluctuation
Theglobeandmail.com	Negative	Fluctuation	Positive	Positive
USAToday.com	Negative	Positive	Positive	Positive
Dailymail.co.uk	Negative	Positive	Positive	Positive
Theguardian.com	Negative	Positive	Positive	Positive
Irishtimes.com	Negative	Positive	Positive	Positive

in news at the beginning of 2020 and lost popularity over time, such as investors and pneumonia, while others gained popularity, such as unemployment and protests. A closer look in Figure 7 shows a few groups of topics whose frequency is dropping over time. For instance, groups 3), 4), 6), 8) which cover topics such as rollout, effects, misinformation, breath, trade, price, travellers. On the other hand, group 2) and 7) with topics such as protests, rally, earnings, banking have increasing popularity over the observed time period.

D. SENTIMENT SCORE DURING DIFFERENT STAGES OF COVID-19:

The data on COVID-19 as reported by the European Centre for Disease Prevention and Control shows that Europe is affected more than Asia and North America (respectively) in terms of total positive cases and deaths per million and total tests per thousand. Overall, from January 2020 to April 2020, the pattern of sentiment in news about the outbreak of COVID-19 disease was the same in Asia, Europe and North America (see Figures 8, 10, 9). However, from May 2020 to July 2021, sentiment score was low in Europe and North America whereas in Asia it varied on the positive side. Onward until December 2020, sentiment score was negative for Europe only. From January 2021 to May 2021, the sentiment score was negative in Europe and Asia whereas it remained positive in North America.

Figure 8 shows the comparison of sentiments across Asia, utilizing the news articles. Each line graph shows the SA of news reported by top five newspapers (Dawn, The Jakarttapost, Livemint, Mainichi, Yanisafak), covering a period from the January 2020 to May 2021. In Asia, the sentiment score increased from negative to positive in articles published by Dawn, The Jakarttapost, Livemint, Mainichi newspapers, while in the case of Yanisafak it stayed negative or fluctuated. In the first quarter (January 2020 - April 2020), the reported sentiment was mostly negative for all the newspapers except Yanisafak which fluctuated between positive and negative (see Table 1). In the second quarter (May 2020 - August 2020), the reported sentiment was approximately positive for Livemint, negative for Dawn and Yanisafak and there were fluctuations for almost all the newspapers. In the third quarter (September 2020 - December 2020), the reported sentiment

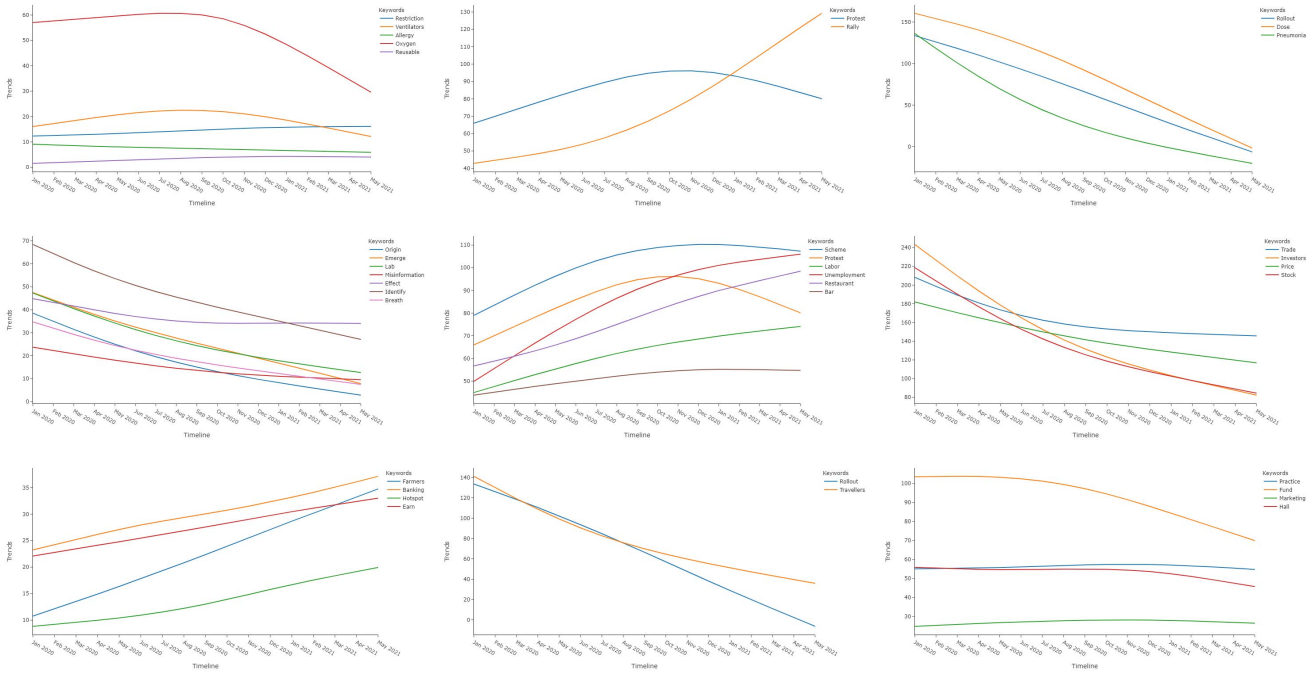


FIGURE 7. Graphs showing nine groups of topics with similar trends (x-axis = time, y-axis = trends). From top left 1) Restrictions, next to it 2) Protests and 3) Pneumonia, in the second row 4) Effects, 5) Unemployment, 6) Trade, in the bottom row 7) Farmers, 8) Travellers, 9) Practice.

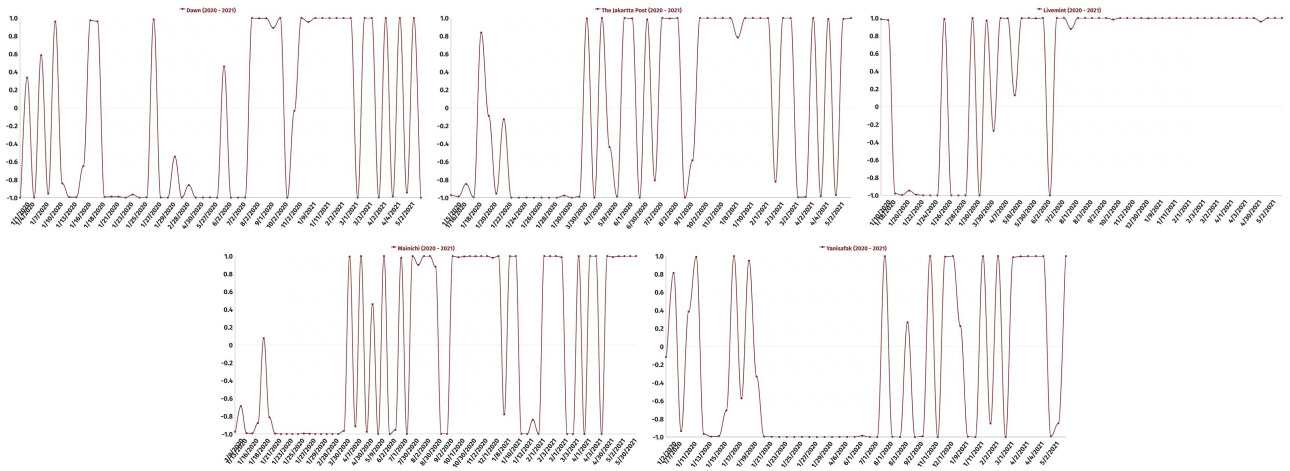


FIGURE 8. Comparison of sentiment score on scale -1 to 1 over months in news reported by top newspapers in Asia related to COVID-19. Each dot represents the sentiment of news articles published on specific date. The shown sentiment score varies: first quarter (January-April 2020) mostly negative, second quarter (May-August 2020) mostly positive, third quarter (Sep-Dec 2020) mostly positive and, fourth and last quarter (Jan-May 2021) fluctuation between positive and negative.

was approximately positive for all the newspapers except Yanisafak which fluctuated between positive and negative. In the first quarter of 2021 (January 2021 - May 2021), there were fluctuations for mostly all the newspapers.

In North America, the sentiment score began with negative scores and then changed to positive score dramatically for news reported by all the newspapers in North America (Theglobemail and USAToday). We located the reasons for this change by looking the most frequent words (see Table 3). In fact, in the beginning, there were more negative words such as outbreak, killed, risk, and coronavirus whereas after 4 months there were less negative words such as virus, and pandemic. At the start of the period (January 2020 - April 2020), the reported sentiment was mostly negative for both newspapers. In the second quarter (May 2020 - August 2020), the reported sentiment was approximately positive for USAToday and fluctuated between negative and positive for Theglobemail. In the third quarter of 2020 (September 2020 - December 2020) and in the first quarter of 2021 (January 2021 - May 2021), the reported sentiment was approximately positive for both the newspapers.

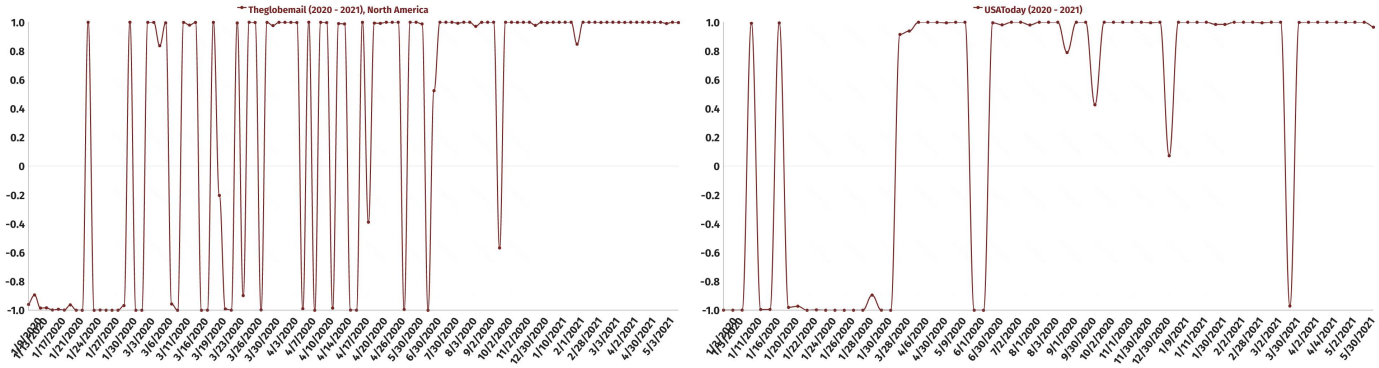


FIGURE 9. Comparison of sentiment score on scale -1 to 1 over months in news reported by top newspapers in North America related to COVID-19. Each dot represents the sentiment of news articles published on specific date. The shown sentiment score varies: first quarter (January-April 2020) negative, second quarter (May-August 2020) mostly positive, third quarter (Sep-Dec 2020) positive and, fourth and last quarter (Jan-May 2021) positive.

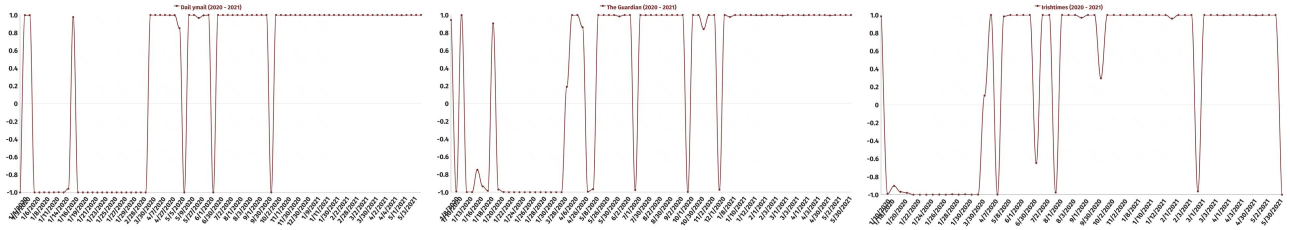


FIGURE 10. Comparison of sentiment score on scale -1 to 1 over months in news reported by top newspapers in Europe related to COVID-19. Each dot represents the sentiment of news articles published on specific date. The shown sentiment score varies: first quarter (January-April 2020) negative, second quarter (May-August 2020) mostly positive, third quarter (Sep-Dec 2020) positive and, fourth and last quarter (Jan-May 2021) positive.

TABLE 2. Political alignments (PA) of top newspapers along with definition of each PA. As RQ1 (see SectionI-C) finds political issues across different political contexts therefore this table reminds of political alignment of each newspaper and its definition.

Political Alignment	Newspaper	Description
Liberal	Dawn	• More governmental involvement to promote socio-economic equality. • Gradual speed of change in government. • The government subsidizes public education at the college level.
Socially Liberal	The Irish Times, Live Mint	• Expected to address economic and social issues such as poverty, welfare, infrastructure, health care. • Emphasize the rights and autonomy of the individual.
Centre Left	Mainichi Shimbun, The Guardian, The Jakarta Post	• It promotes a degree of social equality. • It opposes a wide gap between the rich and the poor. • Support moderate measures to reduce the economic gap such as progressive income tax
Islamic Conservative	Yeni Şafak	• It applies the teachings of particular religions to politics. • It oppose abortion, drug use, and sexual outside of marriage.
Centre Right	Daily Mail, The Globe and Mail	• It supports free markets, limited government spending.
Centrist	Dawn, USA Today	• It leans to accept or support a balance of social equality and a degree of social hierarchy. • It opposes the political changes that can shift society either to the left or right.
Left Wing	Mainichi Shimbun, The Guardian, The Jakarta Post	• It supports social equality (in a society, equal rights, liberties and status, freedom of speech, etc.)
Right Wing	Daily Mail, The Globe and Mail	• It supports the view that certain social orders and hierarchies are inevitable, natural, normal, or desirable.

On the other hand, the sentiment score of news reported by European newspapers (Dailymail, The Guardian, Irishtimes) was negative in the first quarter of 2020 (January 2020 - April 2020), whereas in all other 3 quarters (May 2020 - August 2020, September 2020 - December 2020, January 2021 - May 2021), the score mostly stayed positively.

V. DISCUSSION

Overall, the results of our spatio-temporal analysis confirm with our research hypothesis. The general findings on how political and economic issues propagated over time during the pandemic and across different political and economic barriers provided relevant insights. Left wing newspapers promote a degree of social equality (see Table 2). What we found as common and most frequent topics were emergence of virus, its causes, questions about efficacy of vaccines, effects of lockdown, financial aspects of the market, and reusable masks. Right wing newspapers support free markets and reduction of government spending (see Table 2) and the list of topics that appeared most frequently were conspiracy and misinformation about Coronavirus, protests, policies on face masks. The newspapers with moderate political alignment do not stand on either side of the spectrum (left or right wing) according to the Table 2. Since the topic COVID-19 has different associated issues it's very hard to describe what is a moderate topic during the pandemic. But overall we see topics such as conspiracy and misinformation about virus, symptoms of the virus such as breathing, heart problem, pneumonia, respiratory system diseases, allergies and need for oxygen. Generally, liberal newspapers support more governmental involvement, and subsidized public facilities such as public education (see Table 2). What we saw as most prominent was forecasting about vaccine roll-out, different vaccines with different doses and for different age groups, and the importance of ventilators. The overall findings on economic issues which appear as a consequence of COVID-19 are quite consistent with the economic situation of the country of the newspaper. Additionally those newspapers

TABLE 3. Most frequent words in different time periods representing sentiment. The sentiment is based on body text of a news article and we show the most frequent words during different time.

Newspapers	1/2020 - 4/2020	5/2020 - 8/2020	9/2020 - 12/2020	1/2021 - 5/2021
Dawn.com	virus, cases, outbreak, lockdown, infected, disease,	patients, death	prime, economic, national time, pandemic, positive, issue,pilots, vaccine, support, debate, political, gas, help	spread, authorities, outbreak, health, city, measures, travel, situation, case, global, virus, market, passengers, warn~
Thejakartapost.com	outbreak, spread, infected, killed, virus, disease, virus, tourists, coronavirus, medical, flights, respiratory	city, infections, home, authorities, ~restrictions, measures, president	pandemic, economic, year, social, vaccine, national, data, local, help, positive, tested	Wuhan, outbreak, spread, infected, citizens, virus, disease, deadly, ~suspected, world, virus, tourists, market
Livemint.com	china, outbreak, spread, infected, confirmed, markets, flights, disease, prices, death, situation, workers, investors, deadly	demand, rate, total, growth, take, country, social, testing, patients, positive, virus, vaccine, business, recovery, tested, ~increase, high,~	vaccine, growth, pandemic, data, demand, sales, positive, recovery, ~higher, increase, rate, economy, business,~	china, virus, outbreak, spread, travel, ~impact, flights, measures, novel, medical, patients, situation, workers, investors, international
Mainichi.jp	outbreak, authorities, travel, medical, measures, virus, patients, respiratory, symptoms, severe, emergency, companies, deadly	social, police, novel, masks, Chinese, week, health, testing, police, business, national	Health, spread, outbreak, travel, medical, measures, flights, infections, ~president, voters, economy, positive, vote	China, Wuhan, spread, outbreak, china, infected, travel, medical, measures, global, respiratory, symptoms, emergency, cause, deadly
Yenisafak.com	outbreak, India, flights, internet, infected, travel, virus, Pakistan, virus, news, Respiratory, ~Korea, media, Organization	spread, least, recoveries, pandemic, nuclear, killed, coronavirus, hit	total, pandemic, past, positive, ~infections, vaccine, economic, tested, deaths, recovery, restrictions	outbreak, spread, flights, coronavirus, respiratory, international, medical, lockdown, ~statement, world, capital
Theglobeandmail.com	china, global, spread, outbreak, ~health, markets, prices, stock, price, ~reported, risk,	police, workers, virus, long, week, home, long	global, market, election, vaccine, across, investors, pandemic, financial,~	virus, global, market, spread, outbreak, health, financial, prices, impact, stock, price, risk
USAToday.com	china, travel, spread, outbreak, ~health, passengers, news, flights, respiratory, symptoms, severe, killed, coronavirus	social, police, players, virus, pandemic, right, care, school, work, week	positive, election, white, tested, pandemic, early, made, season, test, mask, election, mask, way, country	Virus, spread, travel, medical, flights, impeachment, world, city, Wuhan, Chinese, passengers
Dailymail.co.uk	Wuhan, outbreak, patients, infected, ~symptoms, medical, tested, deadly, disease, virus, coronavirus,~	Chinese, outbreak, city, health, symptoms, medical, travel, masks, tested, man, virus, world, postponed, positive	lockdown, make, set, work, pandemic, restrictions, see, positive, social, support, government	spread, patients, health, travel, tested, disease, world, authorities, postponed, positive
Theguardian.com	virus, outbreak, spread, medical, ~travel, disease, symptoms, infected, risk, death, infected, risk, China	social, pandemic, think, local, support, really, police, working, restrictions	support, national, care, election, ~restrictions, social, economic, England	Medical, Health, Travel, British, City, Global, Staff, country, death, family, case, risk, citizens
Irishtimes.com	China, spread, outbreak, city, patients, global, infected, death, emergency, market, fell, medical, measures, hospital, WHO	Work, Back, Home, Family, Go, Social, Travel, Children, Country, restrictions, pandemic, life, lockdown	Work, Group, Good, Well, Positive, ~Months, Chief, Government, Business, Tested, European	Wuhan, Spread, travel, outbreak, health, city, patients, global, infected, staff, death, emergency, market, fell, measures, citizens, deaths

that focus on local political content rather global do the same when it comes to the economy. In newspapers from lower level economies the most frequent topics were funding, pays, pricing, farmers, banking, vaccine cost, marketing, lockdown effects on labour, remittance, oil prices, and starvation. In case of middle level economies the most frequent topics were funds, trade, investment, job crisis, no trade, lower business, unemployment and fragile economy. For high income economies the most frequent topics were scheme, vaccines, number of doses, stocking the vaccine, task forces, quarantine issue, protests, global pandemic, and hot-spot lockdown issues.

Regarding the second research question - How different discussions evolved during different stages of the COVID-19 epidemic? -, we see that the trends of different correlated terms were same although the frequency of occurrence of the topics were different. For example, origin, emerge, lab, misinformation, and so one has had same trend over time but their frequency of occurrence were different (see Figure 7). Regarding the third research question - What are the patterns

of emotional states during different stages of the pandemic across different countries? -, we see that the news published by European and North American newspapers depict more or less the same sentiments, whereas the Asian newspapers show some differences.

VI. CONCLUSION AND FUTURE WORK

In this paper we focused on the analysis of news articles related to COVID-19 (published between January 2020 and May 2021) by observing the political alignment of different newspapers, different economies, and sentiments across different continents. Our prime motivation was to understand news reporting by newspapers with different political alignments and news from different economies. We also include the analysis of topics which have had similar trends over time and comparison of sentiments across different continents. We firstly constructed five queries related to COVID-19 and filtered news articles for these queries for each month (total 17 months) and for each newspaper. We performed TM and generated topics. Afterward, we filtered political and economic issues and analyzed the topics on top of political alignments and different economies (see Figure 6 and 5). Next, we identified the trends of similar terms (see Figure 7). Lastly, we calculated sentiment scores for each newspaper using each news article (see Figures 8, 10, 9). Our findings suggest that 1) Left wing newspapers report on raising questions and are futured-oriented, 2) The newspapers with moderate political alignment report on conspiracies and misinformation, 3) Right wing newspapers report on conspiracies and misinformation as well as protests and rallies, 4) Liberal newspapers report on worldwide challenges such as lockdown and face masks, 5) Regarding the economy, we see that newspapers report on national economic situations more than on global economic situations, 6) With regards to sentiment, we see that the news published by European and North American newspapers depict more or less the same sentiments, whereas the Asian newspapers show some differences.

Overall, our study suggests that the political alignment of a newspaper and the economic condition of a country influence news spreading. For example, topics such as lockdown effects on labour, vaccine cost, prices appear at the same time in newspapers from lower level economies, whereas topics such as number of doses, task forces, quarantine protests appear at the same time in newspapers from high income economies. If we look at the evolution of topics we see that topics which are semantically related to each other have similar trends over time. For instance, topics such as origin of virus, lab theory, misinformation appear at the same time in newspapers. Similarly other topics such as unemployment, labor, protests appear at the same time in newspapers. Our analysis of sentiment score suggests that the news articles published by European and North American newspapers depict more or less the same sentiments whereas Asian newspapers show greater differences. We verified that sentiment score in European newspapers was negative in the beginning of the pandemic, but became positive in the rest of the pandemic. In the case of the Asian newspapers the differences between newspapers are more significant, and it is more difficult to devise a line of concordance in sentiment amongst newspapers, particularly over time.

There are a few limitations that we compromised on while analyzing the results and answering our research questions. Data size is not big enough. The preconditions to select the newspaper (see Section III-A) are very strict because the platform that we use to collect news articles (Event Registry) does not provide enough news articles for all the newspapers or it might be possible that some newspapers do not make their article available after some time period or do not publish enough articles related to COVID-19. Another reason is that we estimate the political alignment using Wikipedia-infobox which does not provide political alignment for all the newspapers. Also, there are multiple determinants to know the economic conditions of a country, [5], [6] and we only use income level to estimate the economic context of a country.

In conclusion, news offers detailed information about the current pandemic situation in different countries. Insights gained from this analysis can support government decision making and communication strategies. It can also encourage further discussion about the management of COVID-19 and other global health events in accordance to each country's political and economical context. In the future, we plan to analyze news reporting differences on top of different geographical places, different cultural values, and linguistic influences.

REFERENCES

[1] Q. Chen, R. Leaman, A. Allot, L. Luo, C.-H. Wei, S. Yan, and Z. Lu, "Artificial intelligence (AI) in action: Addressing the COVID-19 pandemic with natural language processing (NLP)," 2020, *arXiv:2010.16413*.

[2] A. Vizeu, "O telejornalismo como lugar de referência e a função pedagógica," *Revista Famecos*, vol. 16, no. 40, pp. 77–83, 2009.

[3] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving LDA topic models for microblogs via tweet pooling and automatic labeling," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2013, pp. 889–892.

[4] D. Alvarez-Melis and M. Saveski, "Topic modeling in Twitter: Aggregating tweets by conversations," in *Proc. 10th Int. AAAI Conf. Web Social Media*, 2016.

[5] I. AlAgha, "Topic modeling and sentiment analysis of Twitter discussions on COVID-19 from spatial and temporal perspectives," *J. Inf. Sci. Theory Pract.*, vol. 9, no. 1, pp. 35–53, 2021.

[6] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. MicheleValensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. A. Scala, "The COVID-19 social media infodemic," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, 2020.

[7] R. J. Medford, S. N. Saleh, A. Sumarsono, T. M. Perl, and A. U. Lehmann, "An 'infodemic': Leveraging high-volume Twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak," in *Open Forum Infectious Diseases*, vol. 7. Oxford, U.K.: Oxford Univ. Press, 2020.

[8] E. A. Iboi, O. O. Sharomi, C. N. Ngonghala, and A. B. Gumel, "Mathematical modeling and analysis of COVID-19 pandemic in Nigeria," *MedRxiv*, 2020.

[9] A. R. Ahmad and H. R. Murad, "The impact of social media on panic during the COVID-19 pandemic in Iraqi kurdistan: Online questionnaire study," *J. Med. Internet Res.*, vol. 22, no. 5, May 2020, Art. no. e19556.

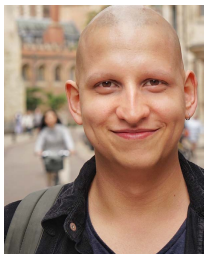
[5] https://www.prosperity.com/rankings

[6] https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

[10] A. H. Abbas, "Politicizing the pandemic: A schemata analysis of COVID-19 news in two selected newspapers," *Int. J. Semiotics Law Revue internationale de Sémiotique juridique*, pp. 1–20, Jul. 2020.
[11] A. Bridgman, E. Merkley, P. J. Loewen, T. Owen, D. Ruths, L. Teichmann, and O. Zhilin, "The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media," *Harvard Kennedy School Misinformation Rev.*, Jun. 2020.
[12] A. Alabrah, H. M. Alawadh, O. D. Okon, T. Meraj, and H. T. Rauf, "Gulf countries' citizens' acceptance of COVID-19 vaccines—A machine learning approach," *Mathematics*, vol. 10, no. 3, p. 467, Jan. 2022.
[13] H. Chen, X. Huang, and Z. Li, "A content analysis of Chinese news coverage on COVID-19 and tourism," *Current Issues Tourism*, pp. 1–8, 2020.
[14] P. Ghasiya and K. Okamura, "Investigating COVID-19 news across four nations: A topic modeling and sentiment analysis approach," *IEEE Access*, vol. 9, pp. 36645–36656, 2021.
[15] A. Sittar, D. Mladenić, and M. Grobelnik, "Analysis of information cascading and propagation barriers across distinctive news events," *J. Intell. Inf. Syst.*, vol. 58, pp. 1–34, Feb. 2021.
[16] P. Ghosh and A. Cartone, "A spatio-temporal analysis of COVID-19 outbreak in Italy," *Regional Sci. Policy Pract.*, vol. 12, no. 6, pp. 1047–1062, 2020.
[17] A. Paez, F. A. Lopez, T. Menezes, R. Cavalcanti, and M. G. D. R. Pitta, "A spatio-temporal analysis of the environmental correlates of COVID-19 incidence in Spain," *Geographical Anal.*, vol. 53, no. 3, pp. 397–421, Jul. 2021.
[18] C. Comito, "How COVID-19 information spread in U.S. the role of Twitter as early indicator of epidemics," *IEEE Trans. Services Comput.*, early access, Jun. 22, 2021, doi: 10.1109/TSC.2021.3091281.
[19] B. Gross, Z. Zheng, S. Liu, X. Chen, A. Sela, J. Li, D. Li, and S. Havlin, "Spatio-temporal propagation of COVID-19 pandemics," *Europhys. Lett.*, vol. 131, no. 5, p. 58003, Sep. 2020.
[20] A. Ceron, "Internet, news, and political trust: The difference between social media and online media outlets," *J. Comput.-Mediated Commun.*, vol. 20, no. 5, pp. 487–503, Sep. 2015.
[21] A. Sittar, D. Mladenić, and T. Erjavec, "A dataset for information spreading over the news," in *Proc. 23th Int. Multiconf. Inf. Soc. (SiKDD)*, vol. 100, 2020, pp. 5–8.
[22] A. Alambo, M. Gaur, and K. Thirunarayan, "Depressive, drug abusive, or informative: Knowledge-aware study of news exposure during COVID-19 outbreak," 2020, *arXiv:2007.15209*.
[23] J. Wilke, C. Heimprecht, and A. Cohen, "The geography of foreign news on television: A comparative study of 17 countries," *Int. Commun. Gazette*, vol. 74, no. 4, pp. 301–322, Jun. 2012.
[24] B. Martens, L. Aguiar, E. Gomez-Herrera, and F. Mueller-Langer, "The digital transformation of news media and the rise of disinformation and fake news," Tech. Rep., 2018.
[25] T.-K. Chang and J.-W. Lee, "Factors affecting Gatekeepers' selection of foreign news: A national survey of newspaper editors," *J. Quart.*, vol. 69, no. 3, pp. 554–561, Sep. 1992.
[26] H. Jo and C. Park, "Analysis of reporting characteristics of newspapers in the 19th presidential election based on random forest," *J. Korean Data Inf. Sci. Sociaty*, vol. 29, no. 2, pp. 367–375, Mar. 2018.
[27] J. Filla and M. Johnson, "Local news outlets and political participation," *Urban Affairs Rev.*, vol. 45, no. 5, pp. 679–692, May 2010.
[28] H. Wei, J. Sankaranarayanan, and H. Samet, "Enhancing local live tweet stream to detect news," in *Proc. 2nd ACM SIGSPATIAL Workshop Analytics Local Events News*, Nov. 2018, pp. 1–10.
[29] E. Bakshy, S. Messing, and L. A. Adamic, "Exposure to ideologically diverse news and opinion on Facebook," *Science*, vol. 348, no. 6239, pp. 1130–1132, 2015.
[30] A. Sittar and D. Mladenic, "How are the economic conditions and political alignment of a newspaper reflected in the events they report on?" in *Proc. Central Eur. Conf. Inf. Intell. Syst.*, 2021, pp. 201–208.
[31] L. Nemes and A. Kiss, "Social media sentiment analysis based on COVID-19," *J. Inf. Telecommun.*, vol. 5, no. 1, pp. 1–15, Jan. 2021.
[32] G. Barkur, Vibha, and G. B. Kamath, "Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India," *Asian J. Psychiatry*, vol. 51, Jun. 2020, Art. no. 102089.
[33] A. D. Dubey, "Twitter sentiment analysis during COVID-19 outbreak," Tech. Rep., 2020.
[34] D. M. El-Din Mohamed Hussein, "A survey on sentiment analysis challenges," *J. King Saud Univ.-Eng. Sci.*, vol. 30, no. 4, pp. 330–338, 2018.
[35] K. H. Manguri, R. N. Ramadhan, and P. R. M. Amin, "Twitter sentiment analysis on worldwide COVID-19 outbreaks," *Kurdistan J. Appl. Res.*, pp. 54–65, May 2020.
[36] A. Älgå, O. Eriksson, and M. Nordberg, "Analysis of scientific publications during the early phase of the COVID-19 pandemic: Topic modeling study," *J. Med. Internet Res.*, vol. 22, no. 11, Nov. 2020, Art. no. e21559.
[37] S. Boon-Itt and Y. Skunkan, "Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study," *JMIR Public Health Surveill.*, vol. 6, no. 4, Nov. 2020, Art. no. e21978.
[38] A. Amara, M. A. H. Taieb, and M. B. Aouicha, "Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis," *Int. J. Speech Technol.*, vol. 51, no. 5, pp. 3052–3073, May 2021.
[39] F. Jafarinejad, M. Rahimi, and H. Mashayekhi, "Tracking and analysis of discourse dynamics and polarity during the early corona pandemic in Iran," *J. Biomed. Informat.*, vol. 121, Sep. 2021, Art. no. 103862.
[40] C. Comito, C. Pizzuti, and N. Procopio, "Online clustering for topic detection in social data streams," in *Proc. IEEE 28th Int. Conf. Tools with Artif. Intell. (ICTAI)*, Nov. 2016, pp. 362–369.
[41] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to Information Retrieval*, vol. 39. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[42] A. A. Adebiyi, O. Ogunleye, O. M. Adebiyi, and J. O. Okesola, "A comparative analysis of TF-IDF, LSI and LDA in semantic information retrieval approach for paper-reviewer assignment," *J. Eng. Appl. Sci.*, vol. 14, no. 10, pp. 3378–3382, Nov. 2019.
[43] M. Hasan, M. A. Orgun, and R. Schwitter, "A survey on real-time event detection from the Twitter data stream," *J. Inf. Sci.*, vol. 44, no. 4, pp. 443–463, 2018.
[44] H.-Y. Kuo, S.-Y. Chen, and Y.-T. Lai, "Investigating COVID-19 news before and after the soft lockdown: An example from Taiwan," *Sustainability*, vol. 13, no. 20, p. 11474, Oct. 2021.
[45] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik, "Event registry: Learning about world events from news," in *Proc. 23rd Int. Conf. World Wide Web (WWW Companion)*, 2014, pp. 107–110.
[46] T. Kumar, "An evidence review of face masks against COVID-19," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 9, no. 12, pp. 919–923, Dec. 2021.
[47] C. A. Latkin, L. Dayton, M. Moran, J. C. Strickland, and K. Collins, "Behavioral and psychosocial factors associated with COVID-19 skepticism in the united states," *Current Psychol.*, vol. 2021, pp. 1–9, Jan. 2021.
[48] H. Sipra, F. Aslam, J. H. Syed, and T. M. Awan, "Investigating the implications of COVID-19 on PM2. 5 in Pakistan," *Aerosol Air Quality Res.*, vol. 21, no. 2, Feb. 2021, Art. no. 200459.
[49] P. Wintachai and K. Prathom, "Stability analysis of SEIR model related to efficiency of vaccines for COVID-19 situation," *Heliyon*, vol. 7, no. 4, Apr. 2021, Art. no. e06812.
[50] O. Dyer, "COVID-19: China pressured who team to dismiss lab leak theory, claims chief investigator," Tech. Rep., 2021.
[51] D. Bhowmik and S. Poddar, "Cyclical and seasonal patterns of India's GDP growth rate through the eyes of Hamilton and hodrick prescott filter models," *Asia–Pacific J. Manage. Technol.*, vol. 1, no. 3, pp. 7–17, 2021.
[52] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 8, 2014, pp. 1–10.

ABDUL SITTAR received the B.S. degree in computer science from the University of Gujrat, Gujrat, Pakistan, and the M.S. degree from the COMSATS Institute of Information Technology, Lahore, Pakistan. He is currently pursuing the Ph.D. degree with Jozef Stefan Institute, Slovenia. From 2013 to 2017, he was working as a Software Engineer at Nextbridge (Pvt.) Ltd. His development career includes iOS and android mobile application development. From 2018 to 2019, he was working as a Lecturer with PMAS-Arid Agriculture University Rawalpindi. He is currently a MSCA Fellow, working on CLEOPATRA ITN. His research interests include natural language processing (NLP) and machine learning (ML).

DANIELA MAJOR received the B.A. degree in history and the M.Litt. degree in intellectual history from the School of Advanced Study, University of London, where she is currently pursuing the Ph.D. degree in digital humanities. From 2018 to 2019, she was a Research Scholar with Arquivo.pt, Portuguese Web Archive, where she used multilingual sources kept in web archives to study the ''Commemorations of the First World War.'' Currently, she is working on a thesis on ''Ideas of Europe and the modern media coverage of the European Union.'' Her research interests include contemporary European history, conceptualizing and working with born-digital sources, and the preservation of online content.



CAIO MELLO received the B.A. degree in journalism and the M.A. degree in communication from the Universidade Federal de Pernambuco, Brazil. He is currently pursuing the Ph.D. degree in digital humanities with the School of Advanced Study, University of London. He works as an Early-Stage Researcher with the University of London, for the EU-funded Horizon 2020 Project CLEOPATRA, under the Marie Skłodowska-Curie Innovative Training Network. Before starting his Ph.D. degree, he was a Fellow Researcher with the CAIS—Center for Advanced Internet Studies, Bochum, Germany.

DUNJA MLADENIĆ is leading the Artificial Intelligence Laboratory, Jozef Stefan Institute, serving on the Institute's Scientific Council (2013–2017) as the Vice President (2015–2017) and working on a number of research projects mainly related to machine learning, data and text mining, big data analytic, semantic technologies, and their application on real-world problems. She worked at the Jozef Stefan International Postgraduate School, the University of Ljubljana, the University of Primorska, the University of Zagreb (FERI and FOI), teaching classes related to data analytics and artificial intelligence.

MARKO GROBELNIK is currently an Expert Researcher in the field of artificial intelligence (AI). He co-leads the Department for Artificial Intelligence, Jozef Stefan Institute; co-founded the UNESCO International Research Center on AI (IRCAI); and the CEO of Quintelligence.com specialized in solving complex AI tasks for the commercial world. He collaborates with major European academic institutions and major industries, such as Bloomberg, British Telecom, European Commission, Microsoft Research, and New York Times. He is the coauthor of several books and the co-founder of several start-ups and is/was involved into over 50 EU funded research projects in various fields of artificial intelligence. His research interests include machine learning, data/text/web mining, network analysis, semantic technologies, deep text understanding, and data visualization.

3.2.2 Case Study Russia and Ukraine crisis

Another crisis took over right after the situation with the COVID-19 pandemic settled down [63]. Regardless of the position in the tension between Russia and the West over Ukraine, many countries find themselves facing a serious challenge, for example the effect of the world's supply of raw materials and global supply chain [64]. We applied the previous methodology (see Section 3.2) for the Russia-Ukraine crisis to confirm that the proposed methodology is suitable only for the COVID-19 pandemic or it can be applied for different controversial events. Similar to COVID-19, we understand the effects of different political and economical context's on news reporting.

The collected dataset of news articles is based on the same newspapers (dailmail.co.uk, dawn.com, irishtimes.com, livemint.com, mainichi,jp, theglobeandmail.com, thejakarta-post.com, usatoday.com, yenisafak.com) that were used for COVID-19 because they belong to countries from different economic backgrounds except the Guardian newspaper. The Guardian newspaper does not have enough articles to be considered for the comparison across different economic and political contexts. Therefore we ignored it. Another difference in this study is timeline. Since our main purpose was to look into the results at a small scale, we set a timeline of two months only (January 2023, and February 2023). This dataset consists of 1512 news articles. Figure 3.1 shows the number of articles published by each newspaper in each month.

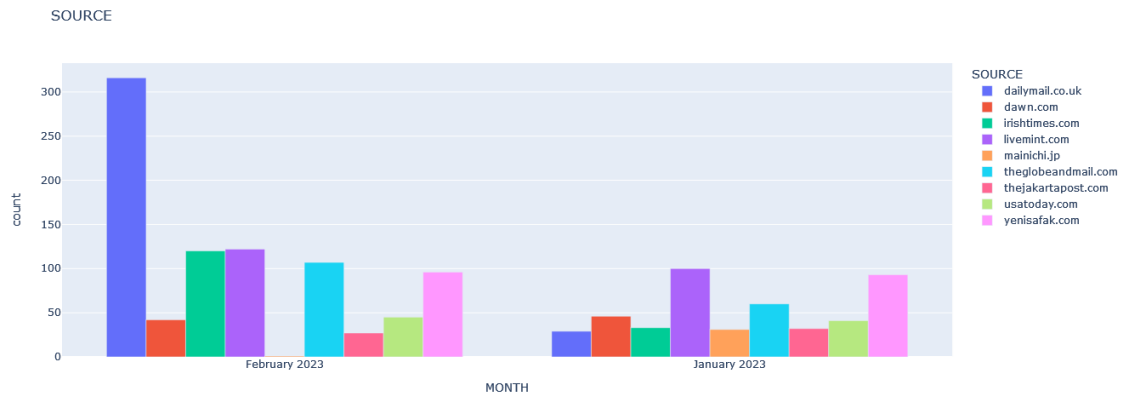


Figure 3.1: **Total number of news articles published in each month (01/2023-02/2023) by nine newspapers. Main purpose of the line graph is to show the variations in collected number of news articles reported by top nine newspapers during the set timeline.**

The adopted approach integrates computational as well as qualitative techniques to explain perceptions and attitudes towards the Russia-Ukraine crisis [65]. Several steps were performed sequentially in the workflow, including data collection, data preparation, filtering news articles, topic modeling, topic analysis, and analysis of political and economic issues. We identify two main queries to pool news articles for the two months - January 2023 and, February 2023. These two queries are: 1) What is the refugee crisis in Ukraine? and, 2) What are the implications of the Russian-Ukraine war?.
We manually filter frequent topics across different economic and political. For each type of political alignment we put together all the filtered topics and show them in word clouds. Figure 3.2 shows the word clouds for different political alignments along with all the queries. Figure 3.3 shows the word clouds for different economic levels along with all the queries.

3.2.2.1 Political Issues:

Figure 3.2 shows the word clouds of main and common topics discussed in newspapers related to different political alignments. The findings have been summarized below:

1) What is the refugee crisis in Ukraine?

Liberal newspapers have been found discussing global crisis, economic and political uncertainty, energy crisis, stock market rates, inflation, wheat prices, food shortages. Right-wing newspapers appeared to focus more on sanctions, peace-talk proposals, oil prices, fuel market whereas left-wing newspapers talk about global prices. No views appeared relating to the refugee crisis in Ukraine by newspapers with central political alignment. Lastly, quite similar discussion appeared in Islamic conservative newspapers.

2) What are the implications of the Russian-Ukraine war?

The newspapers with liberal political alignment mention the topics referendum, cut-oil-production, military force, nuclear agreement, Russian aircraft, global inflation, prices growth, global crisis, inflation rise, energy war, Russian missile. The centre-right newspapers talk about military invasion, increment in prices, allies support, military conflict, energy prices, oil debt, whereas the centre-left newspapers indicate rise in fuel prices, trade deficit, peace strategy, world tension, and global environment. The newspapers with moderate political alignment talks about inflation rise, cruise, and weapons. Islamic conservative newspapers indicates topics such as protection of families.

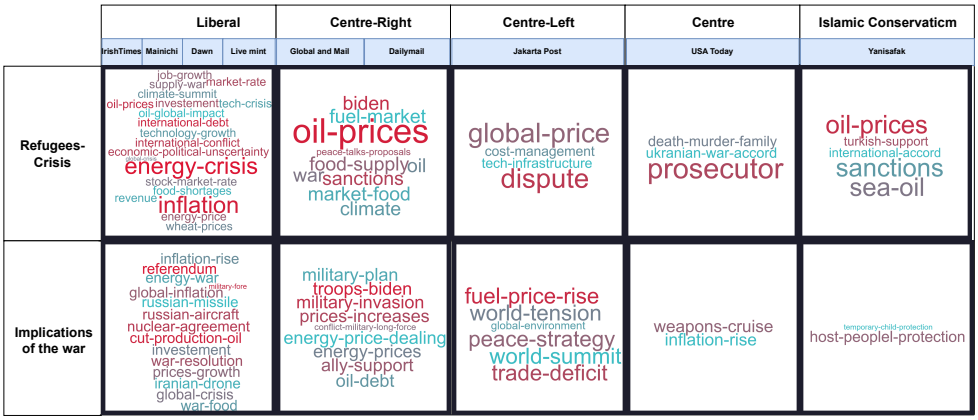


Figure 3.2: **Word clouds showing the keywords appearing most frequently related to two queries in relation to different political alignments.**

3.2.2.2 Economic Issues:

Figure 3.3 shows the word clouds of main and common topics discussed in newspapers belonging to different economies. Each topic has been discussed in different economies with different results. These findings have been summarized below:

1) What is the refugee crisis in Ukraine?

Low-income countries mention the topics international debt, oil prices, economic uncertainty, market rates, investment, global crisis, inflation, wheat prices, revenue, global prices, cost management. Topics such as oil-prices, sea oil, sanctions, and international accord appeared with high frequency in middle-income countries. Terms that appeared in newspapers from high-income countries are sanctions, peace talks, energy crisis, stock market

rate, inflation, food shortages, global impact of oil, food supply and fuel market.

2) What are the implications of the Russian-Ukraine war?

Lower-income countries discuss cutting the oil production, war resolution, global inflation, prices growth, global crisis, rise in fuel prices, trade deficit, peace strategy, world tension, and global environment. Topics such as temporary child and people protection appear as more frequent. However, in case of high-income countries, plenty of topics appeared as most frequent such as military invasion, ally support, nuclear agreement, investment, Iranian drone, energy prices, oil debt, military plan, energy price dealing, and weapon cruise.

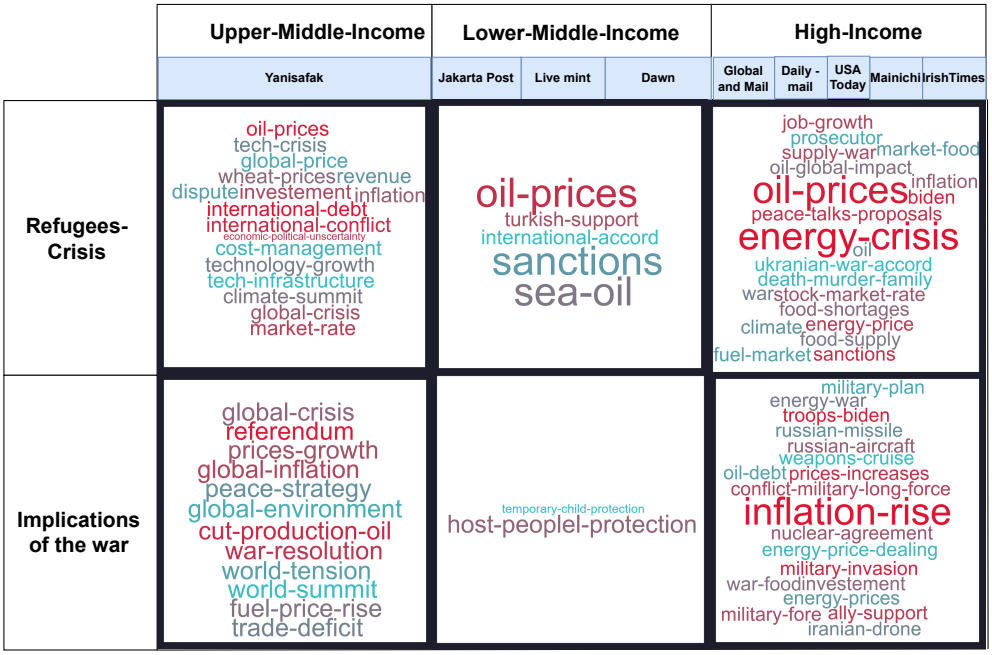


Figure 3.3: **Word clouds showing the keywords appearing most frequently related to two queries in relation to different economies.**

3.2.2.3 Analysis and conclusions

The general findings on how political and economic issues propagated across different political and economic situations do not provide relevant insights. The majority of the topics are overlapping across different political alignments, such as oil prices, sanctions, economic and political uncertainty, international accord, rise in inflation, prices growth and energy war. Similarly, the majority of the topics are overlapping across different economic contexts, such as referendum, cut oil production, military invasion, peace strategy, and global environment. Because of these overlapping topics, the differences across different conditions are not possible. There could be many reasons for this type of results. Firstly, type of event can matters a lot in news reporting differences. For instance, the COVID-19 is a natural disaster whereas the Russia-Ukraine war is a political and war event. Reporting during the COVID-19 can be more smooth, and transparent, whereas for the war event, the reporting can be political. Secondly, the timeline is short of only two months (January 2023- and February 2023), whereas in case of COVID-19 (see Section 3.2.2), the time line consists of 14 months (January 2020 - May 2021).

3.3 Clustering News Reporting

This section presents a paper titled *Stylistic features in clustering news reporting: News articles on BREXIT* by Abdul Sittar, Jason Webber and Dunja Mladenić. This paper was published in Proceedings of the 23rd International Multiconference Information Society SiKDD in Slovenia, Ljubljana, in 2022 [66]. Classification of different news reporting styles requires a collection of valid and prominent textual features. A detailed analysis of textual features is performed by [67] where they derived multiple features for creating clusters of news articles along with their comments. These features include terms in the title, terms in the first sentence, terms, in the entire article, etc. There are mainly three types of textual features such as linguistic, stylistic and content-based.

The role of content in news reporting refers to the type of language that is used in the news. It is used to convey meaning and it can impact social and psychological constructs such as social relationships, emotions, and social hierarchy. Features that could classify news reporting across different regions can be adapted to classify the news. We perform clustering of news articles using bag-of-words features and 25 stylistic features. We compare the clustering results generated by these features. Our results show that stylistic features can be better at clustering textual data than bag-of-words.

Stylistic features in clustering news reporting: News articles on BREXIT

Abdul Sittar
abdul.sittar@ijs.si
Jožef Stefan Institute and Jožef Stefan Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

Jason Webber
jason.webber@bl.uk
British Library
London, United Kingdom

Dunja Mladenić
dunja.mladenic@ijs.si
Jožef Stefan Institute and Jožef Stefan Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

ABSTRACT

We present a comparison of typical bag-of-words features with stylistic features. We group the news articles published from three different regions of the UK such as London, Wales, and Scotland. Hierarchical clustering is performed using typical bag-of-words and stylistic features. We present the performance of 25 stylistic features and compare with the bag-of-words (BOW). Our results show that stylistic features can be used for clustering the news article and their performance is much better than bag-of-words.

KEYWORDS

news reporting, topic modeling, stylistic features, clustering

1 INTRODUCTION

The role of content is an essential research topic in news spreading. Media economics scholars especially showed their interest in a variety of content forms since content analysis plays a vital role in individual consumer decisions and political and economic interactions [6]. The content basically refers to the type of language that is used in the news. It is used to convey meaning and it can impact social and psychological constructs such as social relationships, emotions, and social hierarchy [8]. The everyday act of reading the news is such a big area in which small differences in reporting may shape how events are perceived, and ultimately judged and remembered [5].

News reporting across different regions requires methods to find reporting differences. [7] characterize the relationship between the volume of online opioid news reporting and measures differences across different geographic and socio-economic levels. Scholars across disciplines have explored the institutional, organizational, and individual influences that study the quality and quantity of coverage [3].

Features that could classify news reporting across different regions can be adapted to classify the news. A detailed analysis of textual features is performed by [1] where they derived multiple features for creating clusters of news articles along with their comments. These features include terms in the title, terms in the first sentence, terms in the entire article, etc. Multi-view clustering on multi-model data can provide common semantics to improve learning effectiveness. It exploits different levels of features from the raw features, including low-level features, high-level features, and semantic features [16].

Information Society 2022, 10 October 2022, Ljubljana, Slovenia

Table 1: List of all the stylistic features that are used for clustering.

No.	Feature	No.	Feature
1.	Percentage of Question Sentences	2.	Average Sentence Length
3.	Percentage of Short Sentences	4.	Average Word Length
5.	Percentage of Long Sentences	6.	Percentage of Semicolons
7.	Percentage of Words with Six and More Letters	8.	**Percentage of Punctuations**
9.	Percentage of Words with Two and Three Letters	10.	Percentage of Pronouns
11.	Percentage of Coordinating Conjunctions	12.	**Percentage of Prepositions**
13.	**Percentage of Comma**	14.	Percentage of Adverbs
15.	Percentage of Articles	16.	Percentage of Capitals
17.	Percentage of Words with One Syllable	18.	**Percentage of Colons**
19.	Percentage of Nouns	20.	Percentage of Determiners
21.	Percentage of Verbs	22.	**Percentage of Digits**
23.	Percentage of Adjectives	24.	**Percentage of Full stop**
25.	Percentage of Interjections		

The news coverage registers the occurrence of specific events promptly and reflects the different opinions of stakeholders [4]. We take Brexit as an event to be researched on the topic of news reporting differences across the different regions of the UK. On 23 June 2016, the British electorate voted to leave the EU. This event has already been studied following different aspects such as fundamental characteristics of the voting population, driver of the vote, political and social patterns, and possible failures in communication [9, 2, 9]. In this paper, we explore how different stylistic features help in clustering the news article related to Brexit than bag-of-words.

Following are the main scientific contributions of this paper:

(1) We present a comparison of clustering (using two different textual features: bag-of-words and stylistic features) for news reporting about Brexit in three different regions (London, Scotland, and Wales) of the UK.
(2) We show in our experiments that stylistic features can be better at clustering textual data than bag-of-words.

2 RELATED WORK

In this section, we review the related literature about topic modelling, and different type of textual features.

2.1 Topic Modelling

Topic modelling is used to infer topics from the collection of text-document. Some techniques used only frequent words whereas some use pooling to generate relevant topics and maintain coherence between topics [14]. Topics are typically represented by a set of keywords. Examples of such algorithms are the Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (LSA). Clustering-based topic modelling is another solution.

Table 2: Total number of news articles about Brexit published in three different regions (London, Scotland, and Wales).

Regions	Newspapers	News articles	Total
London	bankofengland.co.uk	8	4248
	bbc.com	2209	
	dailymail.co.uk	768	
	Independent.co.uk	191	
	inews.co.uk	52	
	metro.co.uk	1	
	neweconomics.org	1	
	rspb.org.uk	8	
	theguardian.com	1167	
	theneweuropean.co.uk	1	
	thesun.co.uk	235	
	cityam.com	3	
	conservativewomen.uk	1	
	dailypost.co.uk	1	
	ft.com	2	
	mirror.co.uk	9	
	raeng.org.uk	1	
	standard.co.uk	20	
Scotland	news.stv.tv	533	533
Wales	gov.wales	3	280
	nation.wales	122	
	Walesonline.co.uk	156	

2.2 Stylistic Features

News reporting differences can be reflected through one's speech, writing, and images etc [10, 12]. A language independent features have been used for different tasks of NLP such as plagiarism detection, author diarization. These features considers the text of documents as a sequence of tokens (i.e. sentences, paragraphs, documents). On the basis of these tokens, various types of statistics could be drawn from any language [13]. Stylistic features represent the writing style of a document and have been used for understanding the author writing styles in the past [10]. We use it to explore the clustering of the news articles based on their reporting differences across different regions. Table 1 shows the list of 25 stylistic features used for the development of our proposed clustering of news articles.

2.3 Bag-of-words

A bag-of-words model is a way of extracting features from text. It is basically a representation of text that describes the occurrence of words within a document. It firstly identify a vocabulary of known words and then measure the presence of known words. Topic modelling is typically based on the bag-of-words (BOW). The essential idea of topic model is that a document can be represented by a mixture of latent topics and each topic is a distribution over words [11].

3 DATA COLLECTION

We collected news articles reporting on Brexit in the English language from the UK Web Archive (UKWA). The dataset consists on 5061 news articles after pre-processing. Due to the unavailability of news articles from other regions of the UK, we selected only the regions (London, Scotland, and Wales) which have a sufficient amount of news articles. Table 2 presents the number of news articles published from different regions and by different news publishers.

4 METHODOLOGY

The presented research focuses on clustering news articles. To this end, we experiment clustering with the combination of different features observing their performance. Our methodology consists on four steps and compares the performance of stylistic features and bag-of-words in clustering the news articles, as shown in Figure 1.

In the first step, we select Brexit under topic and themes on UK web archive [1]. After crawling the list of news articles, we extracted the meta data of news publishers from Wikipedia-infobox. The meta-data extraction process is explained in our previous work [15]. In this process, we extracted the headquarter of news publishers. Due to the unavailability of news articles from other regions of the UK, we selected only the regions (London, Scotland, and Wales) which have a sufficient amount of news articles. In the second step, we perform parsing of the html web pages and extract the body text. Since third step required pre-processing for bag-of-words, we convert the text to lower case, remove the stop words and punctuation marks. In the third step for the stylistic features, we extract the stylistics features (see Table 1) for all the three regions and perform LSA (Latent Semantic Analysis). Similarly, for the bag-of-words, we use the pre-processed text and perform LSA.

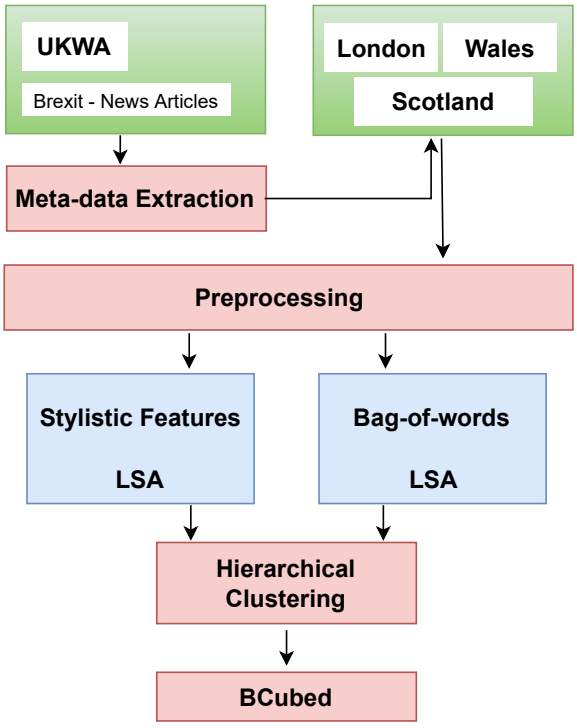


Figure 1: Methodology to clustering regional news using bag-of-words and stylistic features.

[1] https://www.webarchive.org.uk/en/ukwa/collection/910

In the last step, we compute cosine similarity among news articles published from a specific region and perform hierarchical clustering. For the evaluation of both features while clustering the news articles we utilize two different types of evaluation measures such as BCubed F1 and Silhouette Scores.

5 EXPERIMENTAL EVALUATION

We have performed experimental evaluations using intrinsic (Silhouette) and extrinsic (BCubed-F) evaluation measures. The intrinsic evaluation metrics are used to calculate the goodness of a clustering technique whereas extrinsic evaluation metrics are used to evaluate clustering performance. For extrinsic evaluation, we consider clusters generated by k-means clustering using typical bag-of-words as ground truth clusters. The value of k in k-means clustering ranges from 2 to 20. K-means identifies k number of centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. We cannot set the value of k as 1 which means there are no other clusters to allocate the nearest data point.
Silhouette is used to find cohesion. It ranges from -1 to 1. 1 means clusters are well apart from each other and clearly distinguished. 0 means clusters are indifferent, or we can say that the distance between clusters is not significant. -1 means clusters are assigned in the wrong way.
BCubed F-measure defines precision as point precision, namely how many points in the same cluster belong to its class. Similarly, point recall represents how many points from its class appear in its cluster.

- **Silhouette Score:** $S(i) = \frac{b(i)-a(i)}{max(a(i),b(i))}$

where S(i) is the silhouette coefficient of the data point i, a(i) is the average distance between i and all the other data points in the cluster to which i belongs, and b(i) is the average distance from i to all clusters to which i does not belong.

- **BCubed Precision and Recall:**

$$Correctness(i,j) = \begin{cases} 1, if\ L(i) = L(j)\ and\ C(j) = C(j) \\ 0, if otherwise \end{cases}$$

$$BCubed\ Precision = \frac{1}{N}\sum_{i=1}^{N}\sum_{j \epsilon C(i)} \frac{Correctness(i,j)}{|C(i)|}$$

$$BCubed\ Recall = \frac{1}{N}\sum_{i=1}^{N}\sum_{j \epsilon L(i)} \frac{Correctness(i,j)}{|L(i)|}$$

where |C(i)| and |L(i)| denote the sizes of the sets C(i) and L(i), respectively. L(i) and C(i) denote the class and clusters of a point i.

- **BCubed-F Score:** $F = \frac{2 \times Precision \times Recall}{Precision + Recall}$

6 RESULTS AND ANALYSIS

Figure 2 shows the three line graphs. Each graph shows Silhouette scores across a different number of clusters (from 2 to 20) representing different regions of the UK such as Scotland, Wales, and London respectively. Blue and red lines represent bag-of-words (BOW) and stylistic features.

We can see that for all three graphs, the silhouette score of stylistic features is significantly high for all three regions except at one point for Scotland. It means that the distance between the clusters is more significant using stylistic features than BOW which is mostly too close to 0. It suggests that these features are best for clustering than BOW. We explored the different combinations of all these features. But the best results are generated using the combination of 5 features. These five features are the percentage of comma, percentage of punctuation marks, percentage of propositions, percentage of colons, percentage of digits, and percentage of full stop.

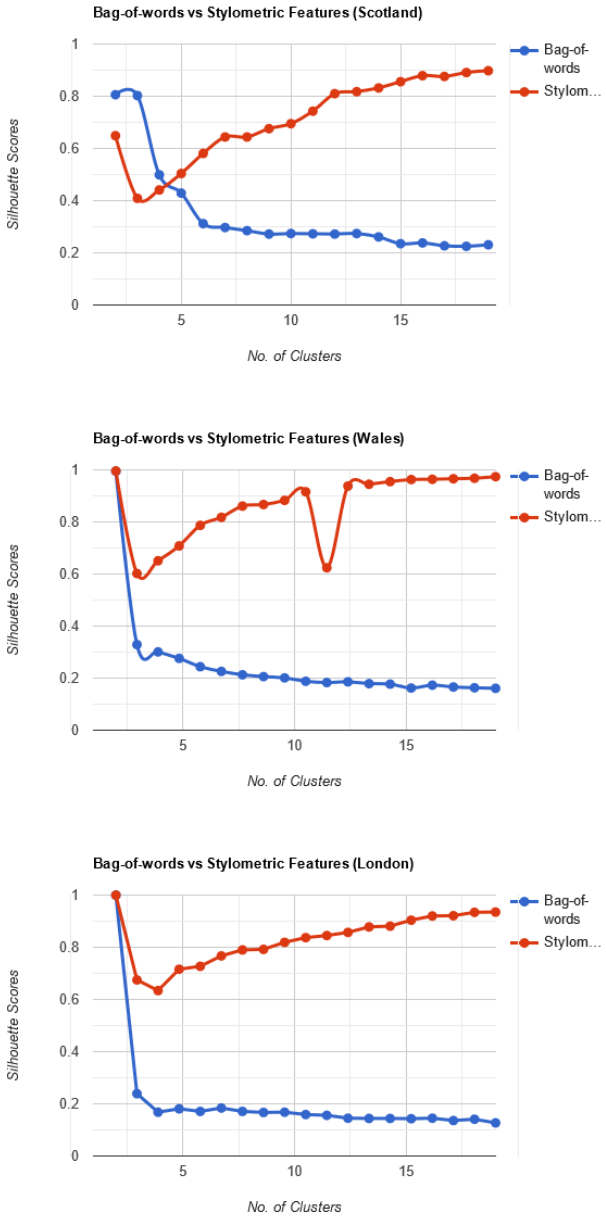


Figure 2: The line graphs represent average silhouette scores across a different number of clusters. The blue line represents the score generated using bag-of-words and the red line represents the score generated using stylistic features. The three-line graphs are generated for three different regions Scotland, Wales, and London respectively.

We consider the results of clustering with bag-of-words as the ground truth clusters. Table 3 shows Bcubed-F scores when the ground truth clusters were matched with the one that was created using stylistic features. The scores for the news articles of all three regions are well over 0.60. One thing is noticeable here. Why are the results not approximately near 1.0 when the

Table 3: Results obtained using two type of clusters: 1) The ground truth results of clustering based bag-of-words, 2) The clusters generated using stylistic features. The results are based on 20 number of clusters.

No.	Region of News Articles	Bcubed-F Score
1.	Scotland	0.66
2.	Wales	0.63
3.	London	0.60

approach of using stylistic features is better for clustering. Since we use typical bag-of-words to generate clusters as ground truth clusters which we already looked into using silhouette score (see Figure 2), considering those clusters as ground truth clusters is doubtful.

7 CONCLUSIONS

In this paper, we have presented the comparison of different features observing their performance over clustering news articles. The goal of this work was to investigate the performance of stylistic features and typical bag-of-words. The data consists on news articles about a popular event Brexit that are collected from UKWA. These news articles belong to three different regions of the UK including Scotland, London, Wales. We performed hierarchical clustering using these features. We use intrinsic and extrinsic clustering evaluation metrics such as Silhouette scores and Bcubed-F1 scores. Our experimental results suggest that the stylistic features can be better at clustering news articles related to a specific event.

ACKNOWLEDGMENTS

The research described in this paper was supported by the Slovenian research agency under the project J2-1736 Causalify and by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 812997.

REFERENCES

[1] Ahmet Aker, Monica Paramita, Emina Kurtic, Adam Funk, Emma Barker, Mark Hepple, and Rob Gaizauskas. 2016. Automatic label generation for news comment clusters. In *Proceedings of the 9th International Natural Language Generation Conference*. Association for Computational Linguistics, 61–69.

[2] Sascha O Becker, Thiemo Fetzer, and Dennis Novy. 2017. Who voted for brexit? a comprehensive district-level analysis. *Economic Policy*, 32, 92, 601–650.

[3] Danielle K Brown and Summer Harlow. 2019. Protests, media coverage, and a hierarchy of social struggle. *The International Journal of Press/Politics*, 24, 4, 508–530.

[4] Honglin Chen, Xia Huang, and Zhiyong Li. 2022. A content analysis of chinese news coverage on covid-19 and tourism. *Current Issues in Tourism*, 25, 2, 198–205.

[5] Elizabeth W Dunn, Moriah Moore, and Brian A Nosek. 2005. The war of the words: how linguistic differences in reporting shape perceptions of terrorism. *Analyses of social issues and public policy*, 5, 1, 67–86.

[6] Frederick G Fico, Stephen Lacy, and Daniel Riffe. 2008. A content analysis guide for media economics scholars. *Journal of Media Economics*, 21, 2, 114–130.

[7] Yulin Hswen, Amanda Zhang, Clark Freifeld, John S Brownstein, et al. 2020. Evaluation of volume of news reporting and opioid-related deaths in the united states: comparative analysis study of geographic and socioeconomic differences. *Journal of Medical Internet Research*, 22, 7, e17693.

[8] Qihao Ji, Arthur A Raney, Sophie H Janicke-Bowles, Katherine R Dale, Mary Beth Oliver, Abigail Reed, Jonmichael Seibert, and Arthur A Raney. 2019. Spreading the good news: analyzing socially shared inspirational news content. *Journalism & Mass Communication Quarterly*, 96, 3, 872–893.

[9] Moya Jones. 2017. Wales and the brexit vote. *Revue Française de Civilisation Britannique. French Journal of British Studies*, 22, XXII-2.

[10] Ifrah Pervaz, Iqra Ameer, Abdul Sittar, and Rao Muhammad Adeel Nawab. 2015. Identification of author personality traits using stylistic features: notebook for pan at clef 2015. In *CLEF (Working Notes)*. Citeseer, 1–7.

[11] Zengchang Qin, Yonghui Cong, and Tao Wan. 2016. Topic modeling of chinese language beyond a bag-of-words. *Computer Speech & Language*, 40, 60–78.

[12] Abdul Sittar and Iqra Ameer. 2018. Multi-lingual author profiling using stylistic features. In *FIRE (Working Notes)*, 240–246.

[13] Abdul Sittar, Hafiz Rizwan Iqbal, and Rao Muhammad Adeel Nawab. 2016. Author diarization using cluster-distance approach. In *CLEF (Working Notes)*. Citeseer, 1000–1007.

[14] Abdul Sittar and Dunja Mladenic. 2021. How are the economic conditions and political alignment of a newspaper reflected in the events they report on? In *Central European Conference on Information and Intelligent Systems*. Faculty of Organization and Informatics Varazdin, 201–208.

[15] Abdul Sittar, Dunja Mladenić, and Marko Grobelnik. 2022. Analysis of information cascading and propagation barriers across distinctive news events. *Journal of Intelligent Information Systems*, 58, 1, 119–152.

[16] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. 2022. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16051–16060.

Chapter 4

Profiling the News Spreading Barriers

For effective dissemination of events, there is a need to remove the barriers. Effective dissemination is the key to bridging the gap in information spreading. For the scientists and the practitioners, it is necessary to participate in explicit, accurate, and unbiased dissemination of their respective areas of expertise to the public [68]. To develop a better news reporting strategy, there is a need to develop an approach to automatic barrier profiling. The barrier profiling intends to assist newspapers in general, but can also be useful for the public. Researchers who want to know the reasons for cultural differences in different communities may learn by comparing the written news articles.

The rest of the chapter presents a novel approach to automatic barrier profiling based on news meta-data which shows experimental evaluation comparing machine learning classical classification methods and deep learning methods, highlighting the complex barriers as well as the role of textual features (see Section 4.1).

4.1 Profiling the News Spreading Barriers

Profiling the news spreading barriers can help to improve media gate-keeping, event-centric news analysis, suspicious news detection, and content recommendations to readers and subscribers. In the world of foreign affairs, editors as gatekeepers in the long chain of news flow undoubtedly hold a central and crucial position in providing news and information to the audiences. One of the most enduring areas of research in media sociology is media gate-keeping, the process by which countless occurrences and ideas are reduced to the few messages we are offered in our news media. News work the process of news-gathering, news writing, and dissemination has come under scrutiny in no small part because people's sense of reality is influenced by what gets into the news and what gets left out [11].

There is a need to profile each barrier separately on top of different events because the flow of news spreading is different across different events and domains. Also, since journalists have to report on uncertain and unexpected events, various factors influence it at different levels. The influence of different factors varies depending upon the issues that journalists deal with (see Section 1.2). The news relating to local events involves domestic factors, whereas the news relating to global events involves the national and international factors that affect their news flow. These factors include economic, political, cultural, linguistic, and geographical influences. Because the news selection process can take place within a complex framework shaped by socio-cultural. Economic stability is one of the factors that influence media coverage [26]. Moreover, the influence of economic power varies across different events and issues (e.g. protests, online privacy, disasters).

Classification based on news headlines using common sense knowledge and sentiments We set another hypothesis to perform automatic barrier profiling. Our hypothesis states that common sense-based semantic knowledge and sentiments of news headlines can help in classifying the news spreading barriers. To verify this hypothesis, we collect news articles that belong to ten different categories (business, computers, games, health, home, recreation, science, shopping, society, and sports). The main focus of this study was to answer the following three questions on the given hypothesis:

- Q1: Do the sentiments of the news headlines of different topics vary across the different barriers?
- Q2: What are the properties (statistics and ratio) of the common sense knowledge relations in news headlines to different topics?
- Q3: Which classification methods (classical or deep learning methods or transformers-based methods) yield the best performance to barrier classification task?

The next section presents the results of the paper *Profiling the barriers to the spreading of news using news headlines* that was published in Frontiers in Artificial Intelligence-Natural Language Processing in 2023 [69]. The authors of this paper are Abdul Sittar, Dunja Mladenić, and Marko Grobelnik.

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 29 August 2023
DOI 10.3389/frai.2023.1225213

OPEN ACCESS

EDITED BY
Alejandro Rodríguez González,
Polytechnic University of Madrid, Spain

REVIEWED BY
Adrian M. P. Brasoveanu,
Modul University Vienna, Austria
Craig S. Webster,
The University of Auckland, New Zealand
Weiqiang Jin,
Xi'an Jiaotong University, China

*CORRESPONDENCE
Abdul Sittar
✉ abdul.sittar@ijs.si

RECEIVED 18 May 2023
ACCEPTED 03 August 2023
PUBLISHED 29 August 2023

CITATION
Sittar A, Mladenić D and Grobelnik M (2023) Profiling the barriers to the spreading of news using news headlines.
Front. Artif. Intell. 6:1225213.
doi: 10.3389/frai.2023.1225213

Profiling the barriers to the spreading of news using news headlines

Abdul Sittar[1,2]*, Dunja Mladenić[1,2] and Marko Grobelnik[2]

[1]Information and Communication Technologies, Jožef Stefan International Postgraduate School (IPS), Ljubljana, Slovenia, [2]Department for Artificial Intelligence - E3, Jozef Stefan Institute, Ljubljana, Slovenia

News headlines can be a good data source for detecting the barriers to the spreading of news in news media, which can be useful in many real-world applications. In this study, we utilize semantic knowledge through the inference-based model COMET and the sentiments of news headlines for barrier classification. We consider five barriers, including cultural, economic, political, linguistic, and geographical and different types of news headlines, including health, sports, science, recreation, games, homes, society, shopping, computers, and business. To that end, we collect and label the news headlines automatically for the barriers using the metadata of news publishers. Then, we utilize the extracted common-sense inferences and sentiments as features to detect the barriers to the spreading of news. We compare our approach to the classical text classification methods, deep learning, and transformer-based methods. The results show that (1) the inference-based semantic knowledge provides distinguishable inferences across the 10 categories that can increase the effectiveness and enhance the speed of the classification model; (2) the news of positive sentiments cross the political barrier, whereas the news of negative sentiments cross the cultural, economic, linguistic, and geographical barriers; (3) the proposed approach using inferences-based semantic knowledge and sentiment improves performance compared with using only headlines in barrier classification. The average F1-score for 4 out of 5 barriers has significantly improved as follows: for cultural barriers from 0.41 to 0.47, for economic barriers from 0.39 to 0.55, for political barriers from 0.59 to 0.70 and for geographical barriers from 0.59 to 0.76.

KEYWORDS

news spreading barriers, profiling news spreading barriers, common-sense inferences, sentiment analysis, economic barrier, political barrier, cultural barrier, linguistic barrier

1. Introduction

News spreading comes across many barriers due to different reasons including cultural, economic, political, linguistic, or geographical. The term barrier refers to the abstract fences that are in place between different societies, nations, and countries while transferring information. We know that the storylines of the news are anchored to the time, places, or entities; therefore, the coverage of news is hampered by news publisher's preferences (Rospocher et al., 2016; Sittar et al., 2022b). The roots of the existence of the mentioned barriers relate to their influences. The classification of such barriers can be useful in the context of numerous real-world applications, such as trend detection and content recommendations for readers and subscribers (Heydari et al., 2015; Gulla et al., 2017). Thus, it is highly important to classify the barriers to massive news spreading related to different events.

Culture is multifaceted, subsuming behaviors, values, and attitudes that are dominant and unique to a particular group of people. The news media has a strong relationship with many macro-level factors in society, ranging from the economy, governments, the public, and other organizational structures (Ng and Tan, 2021). Within a cultural barrier, media diversity provides different opinions and perspectives across different cultures (d'Haenens et al., 2009). The publishing language of a news media also influences the diffusion of news about global and local events (Wright, 2022), so we can say that there is a language barrier. Similarly, the political alignment of news publishers has a direct relationship with the published content, and it is called as a political barrier by Sittar et al. (2022b). Another contextual variable that has a direct relationship with different types of news is the economic situation (we call it the economic barrier), surrounding the news publisher. Since the economic differences in living styles affect the need, the news is likely to propagate according to the needs of locals.

Another important variable in this context is news sentiments about different events across different locations. Several studies use sentiment from textual data, including social media, and news articles, to forecast financial variables (Barbaglia et al., 2022; Consoli et al., 2022; Kumbure et al., 2022). The sentiment of the news plays an important role in news spreading, as Bustos et al. (2011) found that the price movement on the stock exchange has a direct relationship with the news spreading patterns. Similarly, news about global events has different sentiment polarities across the geographical barrier. Moreo et al. (2012) calls it the popularity measurement of news in a global context. Market behavior is also predictable through sentiments (Godbole et al., 2007; Shah et al., 2018), and sentiments can vary by demographic group, news source, and geographic location (Mehler et al., 2006).

When it comes to news headlines, they reflect the vital information of news articles. It reduces the interpretation time and effort of reading the whole article (Shrawankar and Wankhede, 2016). The first thing in the news article is its headline, which makes the first and foremost impression on the news readers. Plenty of news articles are published every day and spread *via* news and social media (Nassirtoussi et al., 2015; Gabielkov et al., 2016; Gravanis et al., 2019). These headlines have different emotional scores with a negative, positive, or neutral polarity, which directly impacts the readers' actions (Aslam et al., 2020).

Barrier classification with news headlines is a challenging task due to incorporating insufficient information as well as misinformation in the headlines. News coverage in different fields, including sports, health, and computers, has different impact levels. We focus on five different types of barriers, including cultural, political, linguistic, economic, and geographic, as these are important barriers that can influence news spreading (Sittar et al., 2022b). In this study, we assume that common sense-based semantic knowledge and sentiments of news headlines will help to classify barriers to the spreading of news. We are interested in exploring the variations in sentiments across different barriers where news headlines belong to different events. We explore a range of different common-sense descriptions generated by the Natural Language Processing Knowledge Inference Tool (Ismayilzada and Bosselut, 2022). In addition, we present an approach to barrier classification that aims to classify barriers across the news. This approach combines information based on news headlines, their inferences, and their sentiment.

The contributions of this research can be summarized as follows:

1. A novel approach to information barrier annotation based on news meta-data.
2. A dataset for the barrier classification in the news that has been labeled automatically using the metadata and the semantic similarity.
3. An approach to the classification of barriers to the spreading of news based on semantic knowledge, including a wide range of common sense inferences and sentiments of news headlines.

The rest of the study is organized as follows: Section 2 reviews the related work on barrier classification; Section 3 presents the approach; Section 4 describes the benchmark dataset construction; Section 5 discusses the experimental results; Section 8 concludes the study and highlights the theoretical and practical implications of our study.

2. Related work

In this section, we present the related work on barriers to the spreading of news and the role of semantic knowledge and sentiments of the news headlines for different tasks.

2.1. Barriers to the spreading of news

Effective dissemination is the key to bridging the gap in information spreading. For the scientists and practitioners, it is necessary to participate in explicit, accurate, and unbiased dissemination of their respective areas of expertise to the public (Kelly et al., 2019). The result of communication is not only situation-specific but also inherently culturally bound because it is entrenched in human acts with intentions, interests, and wants as well as larger institutional, social, and cultural systems (Jiang and Tang, 2020). Culture-specific ideology is defined as the values, beliefs, attitudes, or interests expressed in a source text that is associated with a particular culture or source and that may be viewed as undesirable or incompatible with the dominant values, beliefs, attitudes, or interests of another culture or subculture. It defines the strategies adopted by text producers to bridge the divides in global news transmission. According to MCNelly's theory, the more distance an intermediary communicator has to travel before learning about a news occurrence, the less personally invested he is in it and the more he considers its "marketability" to editors or readers (Vuorinen, 1994). It has been said that countries with close distances share culture, and the news reporting on the same events will not differ due to ideology, culture, and geopolitics (Segev, 2015; Ma et al., 2017). Countries that share a common culture are expected to have heavier news flow between them when reporting on similar events (Wu, 2007). There are many quantitative studies that found demographic, psychological, sociocultural, source, system, and content-related aspects (Al-Samarraie et al., 2017).

The role of content is an essential research topic in news spreading. Media economics scholars especially showed their interest in a variety of content forms since content analysis plays a vital role in individual consumer decisions and political and economic interactions (Fico et al., 2008). In content, a frame is a means to highlight certain elements of a seen reality in a communication text, so as to support a specific problem definition, causal interpretation, moral assessment, and/or therapy proposal for the thing being described. There are four places where frames can be found during communication, such as text, recipient, communicator, and culture (Reese, 2007). The inverted pyramid reporting method, where the most significant facts are presented in order of importance, is a key component of news framing. Bias in the news can manifest in a variety of ways, such as "source bias," "unbalanced presentation of contested themes," and "frequent usage of packaged formula" (Walter and Ophir, 2019). Scheufele identifies five factors that influence how journalists frame news. These include societal expectations and ideals, organizational demands and restrictions, pressure from interest groups, journalistic practices, and journalists' ideological or political leanings (Obijiofor, 2010). A vast body of literature exists on how the news media frame news events and consequently influence public perception of those events (Lamidi and Olisa, 2016). Existing literature posits that framing is often used intentionally for the purpose of changing the perception of content, and to cater to this, different computational methods have been applied (King et al., 2017; Sheshadri et al., 2021).

2.2. Inference-based semantic knowledge

Common-sense transformer (COMET) is an automatic construction of common-sense knowledge bases. It is a framework for adapting the weights of language models to learn to produce novel and diverse common-sense knowledge tuples (Bosselut et al., 2019). Abductive natural language inference can be used to interpret between the lines in natural language (Bhagavatula et al., 2019). Inferences allow us to connect pieces of knowledge to reach a new conclusion. Humans perform natural language inference based on a vast amount of external knowledge about language and the world. To comprehend human language, machines first need linguistic knowledge, i.e., knowledge about the language. This includes the understanding of word meanings, grammar, syntax, semantics, and discourse structure. Having linguistic knowledge gives a human or machine the basic capabilities of understanding language and virtually is a required property of any NLP system, even those not created for NLI tasks. Common knowledge refers to well-known facts about the world that are often explicitly stated (Cambria et al., 2011). This type of knowledge is often referred to human communication (Cambria et al., 2014). Some types of common knowledge may be domain-specific. While domain-specific knowledge is obviously useful for domain-specific applications, much of this knowledge may not be needed for general-purpose communication with humans. Common-sense knowledge, on the other hand, is typically unstated, as it is considered obvious to most humans and consists of universally accepted beliefs about the world. common-sense knowledge provides a deeper understanding of language. While it is rarely referred to language, humans rely on it in communication (Gao et al., 2016), as it is required to reach a common ground. It consists of everyday assumptions about the world and is generally learned through one's own experience with the world but can also be inferred by generalizing over common knowledge. While common knowledge can vary by region, culture, and other factors, we expect that common-sense knowledge should be roughly typical of all humans (Davis et al., 2017).

To tackle the challenging benchmark tasks, many computational models have been developed. These range from earlier symbolic and statistical approaches to recent approaches based on deep neural networks. Explicit textual content is used for different tasks, such as hate speech detection systems, and the primary challenge for statistical and neural classifiers is to infer the implicit messages in text. Recent studies have highlighted the need to use implicit messages to detect textual content (ElSherief et al., 2021). Knowledge graphs have been constructed to answer user questions by identifying the reasoning relations (Jin et al., 2023b). Similarly, an external knowledge base was used with the transformer to perform emotion recognition and bias prediction (Ghosal et al., 2020; Swati and Grobelnik, 2022). Semantic knowledge also proved to enhance the existing model to learn a general representation (Razniewski et al., 2021). There are many examples of recommendation systems that utilize semantic knowledge consisting of several attributes and multi-model knowledge (Zhou et al., 2020; Lei et al., 2021). Taking the semantic information through knowledge graphs is also one of the best ways to associate semantic information with the data for different tasks (Colon-Hernandez et al., 2021). Common-sense knowledge consists of many spatiotemporal features, including spatial, physical, temporal, and psychological aspects of everyday life. It has proven to be crucial for many NLP tasks, including dialogue understanding and generation, event prediction, and question answering (Fang et al., 2021). For the development of new approaches to address different tasks, one of the critical tasks is creating benchmark datasets to evaluate the approaches (Storks et al., 2019).

2.3. Sentiments as semantic knowledge

Sentiment classification of news deals with the identification of positive and negative news that can be used to predict trends related to different tasks (Yazdani et al., 2017). Sentiment of the news has already been used for news classification and other features, including entities and special phrases (Demirsoz and Ozcan, 2017; Hui et al., 2017). In the task of sentiment classification approaches, DistilBERT can transfer basic semantic understanding to further domains, and lead to greater accuracy than the baseline TFIDF (Dogra et al., 2021). For the task of fake news detection, the textual content of the news along with the headline has been used to extract the features (Cui et al., 2019). Taj et al. (2019) used dictionary-based and corpus-based methods for sentiment analysis of news related to business, entertainment, politics, sport, and technology. Li et al. (2017) have used sentiments along with a bag of features to predict the stock market prediction. Aspect-based sentiment analysis has

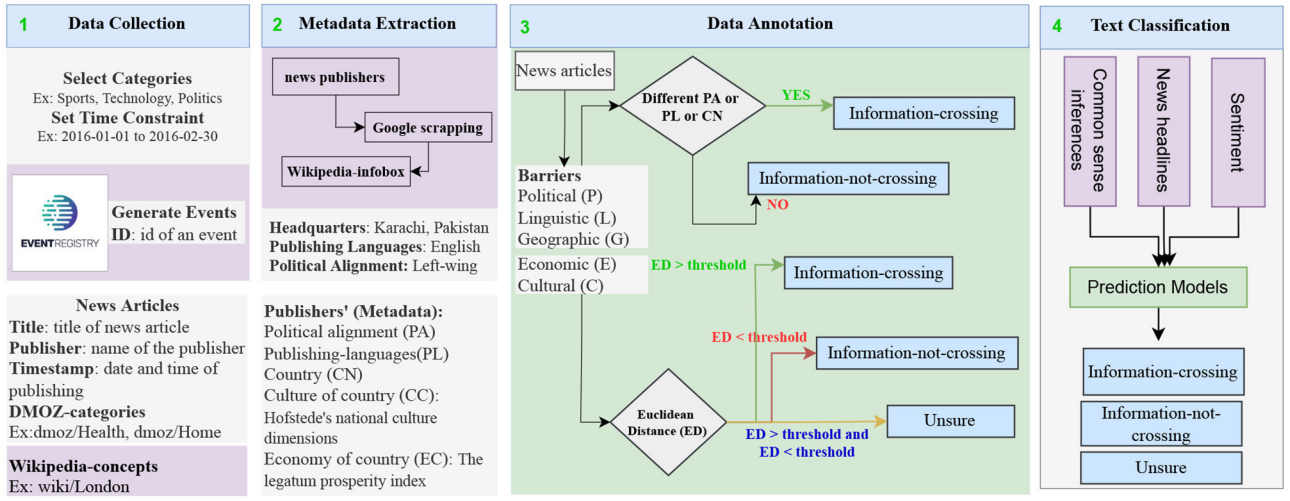


FIGURE 1
An approach to automatic barrier profiling based on the news meta-data. Data extraction from the Event Registry is the first step. Meta-data extraction through Google and Wikipedia scraping is the second step. The third step is to label the news articles after calculating the Euclidean distances. The classification with classical machine learning models, deep learning, and transformer-based methods is performed in the last step.

been performed by infusing external background knowledge in the form of triples (Jin et al., 2023c). Bhutani et al. (2019) prove that sentiments of fake news increase the accuracy of fake news detection, and there exists a strong relationship between news and its sentiments, such as negative emotions tend to spread fast (Ajao et al., 2019).

3. Approach description

To perform the classification of news published across barriers (geographical, cultural, economic, etc.) and, in that attempt, to recommend and identify trends of news spreading belonging to different categories, some methodological considerations are necessary.

This research article presents a novel approach to barrier annotation utilizing news meta-data and an approach to news classification utilizing inference-based semantic knowledge, as shown in Figure 1. In the first step, we execute a query that extracts the news articles from the Event Registry belonging to different categories (business, computers, games, health, home, recreation, science, shopping, society, and sports) and publishes them within a certain time span – in our case, between 2016 and 2021 (see Section 4). Then, we parse and save these news articles along with the source information, such as publishers' names and publishing dates.

In the second step, we extract meta-data related to news publishers *via* searching the news publishers' on Google and extracting their Wikipedia links. Using these links, we obtain the necessary information from Wikipedia-Infobox (Sittar et al., 2022b). We use the Bright Data service[1] to crawl and parse Wikipedia-Infobox.

In the third step, we perform the annotation of news articles. To label the news articles, we set the annotation guidelines (see Section 4). For cultural and economic barriers, we assign ternary labels to news articles, whereas for linguistic, geographical, and political barriers, we assign binary labels to news articles.

In the fourth step, we conduct a detailed analysis of the sentiments of the news headlines for each category (see Figures 2, 3) and provide a list of comprehensive trends of sentiments across different categories and barriers (see Figures 4, 5). Next, we extract semantic knowledge through the inference-based model COMET (Bosselut et al., 2019) (see Figure 6). We analyze the properties of the relations to the news headlines of different topics (see Figures 7, 8). Afterward, we conduct experiments comparing machine learning state-of-the-art (LR, NB, SVC, kNN, and DT), deep learning (LSTM), and transformer-based methods (BERT) using a combination of news headlines with inference-based semantic knowledge and its sentiments. The results are presented in Sections 6.1 and 6.2 showing the performance of different features and methods. The source code for this approach is available in the GitHub repository.[2]

More specifically, we focus on the following research questions:

1. RQ1: Do the sentiments of the news headlines on different topics vary across the different barriers?
2. RQ2: What are the properties (statistics and ratio) of the common-sense knowledge relations in news headlines to different topics?
3. RQ3: Which classification methods (classical or deep learning methods or transformer-based methods) yield the best performance to barrier classification task?

1 https://brightdata.com/

2 https://github.com/abdulsittar/BC-Inferences-Sentiments.git

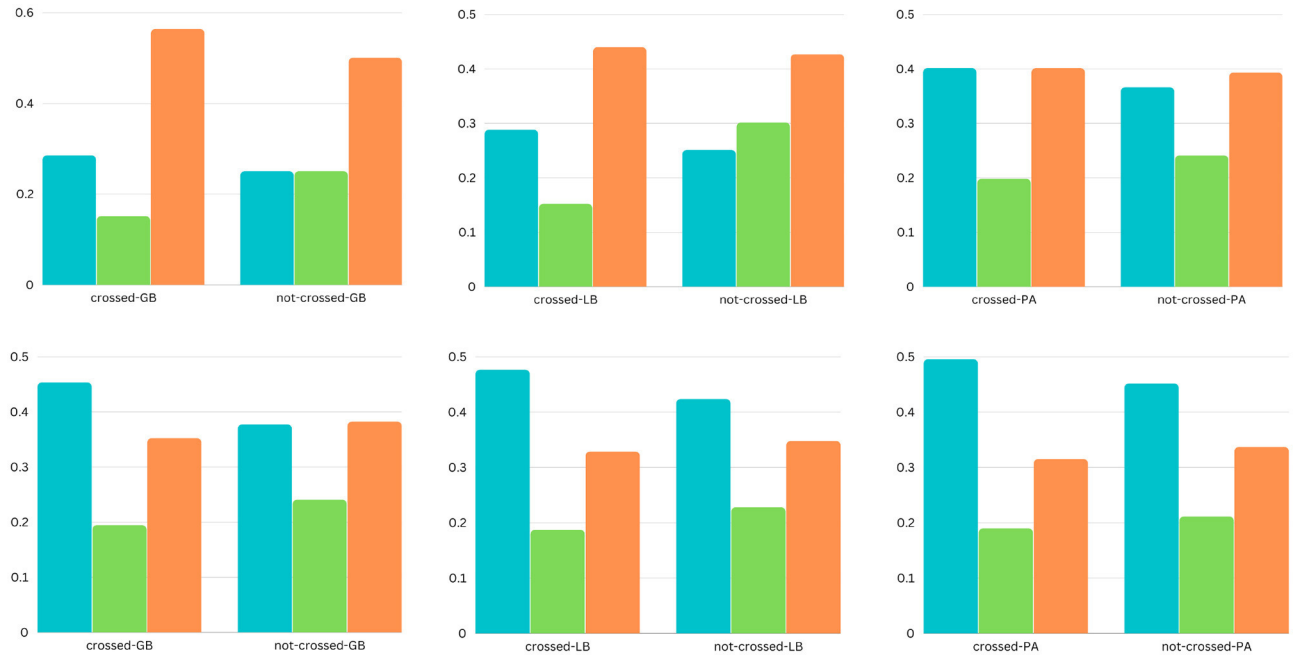


FIGURE 2
The bar charts present the comparison of sentiments (positive, neutral, and negative) in the headlines for two categories (shopping and society). Each column represents (from top to bottom) the two categories, and each row represents (from left to right) the two barriers (cultural and economic).

4. Benchmark dataset construction

We collected the news articles reporting on different events published between 2016 and 2021 in the English language using Event Registry (Leban et al., 2014) APIs.[3] The dataset consists of approximately 1.7 million news articles. Each news article belongs to a different category (see Table 1). Each news article consists of a few attributes, such as title, body text, name of the news publisher, date and time of publishing, event-ID, DMOZ-categories, and Wikipedia concepts.

A few attributes are self-explanatory, such as title, body text, name of the news publisher, and date and time of publication. An event-id represents a unique number that is associated with all the news articles that belong to the same event. The DMOZ-categories represent the topics of the content or news article. It is a project that has a hierarchical collection of web page links organized by subject matter.[4] Approximately 50,000 categories are used by the Event Registry (top 3 layers of the DMoz taxonomy).[5] The statistics of all the categories for all five barriers are presented in Table 1. The Wikipedia concepts are used as a semantic annotation for news articles and can represent entities (locations, people, or organizations) or non-entities (things such as personal computers and toys). In the Event Registry, Wikipedia's URLs are used as concept URIs.

To fetch the metadata for each barrier, the essential thing is the news publisher's headquarter name (see Figure 9). For each news publisher, we get this information from Wikipedia-Infobox. We used the Bright Data service (see text footnote 1) to crawl and parse Wikipedia-Infobox for almost more than 10,000 news websites. We retrieved the country name of the news publisher's headquarters. For the economical barrier, we fetched the economical profile for each country using "The Legatum Prosperity Index"[6] as done by Sittar et al. (2022b). It has 12 dimensions that represent different economic aspects. For the cultural barrier, we calculated differences among different regions using six Hofstede's national culture dimensions (HNCD). For the economic and cultural barriers, we calculated the Euclidean distance among all the countries (for the economic barrier using the economical profile and for the cultural barrier using the HNCD). Two countries were labeled as "information-not-crossing" if the distance score was $\leq$0.1, "unsure" if the distance score was $>$0.1 and $\leq$0.4, and "information-crossing" if the distance score was $>$0.4. For the geographical barrier, we stored general latitude and longitude. For the political barrier, we utilized the political ideology or alignment of the newspaper or magazine that we determined based on Wikipedia-Infobox at their Wikipedia page (Sittar et al., 2022a). The barriers, including cultural, economic, and geographic, do not have a standard representation. They have been estimated by utilizing a relevant set of features. In case of the political and linguistic barriers, we utilized the available political alignments and publishing languages from the Wikipedia-Infobox of a specific publisher. The statistics about the labeled dataset are presented in Figures 10–12 and

3 https://github.com/EventRegistry/event-registry-python/blob/master/eventregistry/examples/QueryArticlesExamples.py

4 https://dmoz-odp.org/

5 https://eventregistry.org/documentation?tab=terminology

6 https://www.prosperity.com/

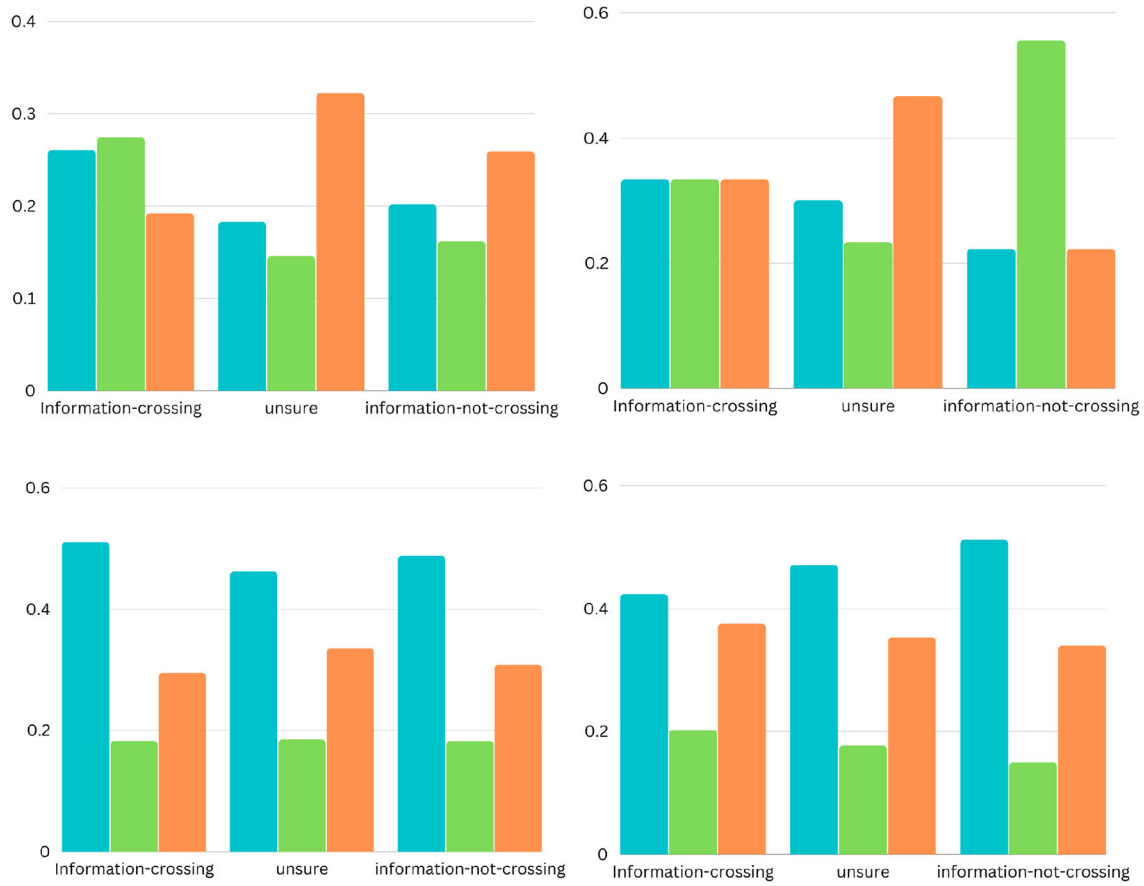


FIGURE 3
The bar charts present the comparison of sentiments (positive, neutral, and negative) in the headlines for two categories (shopping and society). Each column represents (from top to bottom) the two categories, and each row represents (from left to right) the three barriers (geographic, linguistic, and political).

Table 1. Data can be found in the GitHub repository (see text footnote 2).

We set the following annotation questions based on the definitions mentioned above in order to classify the barriers to news spreading.

Q1: *Do all the news articles reporting on an event publish from a particular or the same geographical location?*
Q2: *Do all the news articles reporting on an event publish from the locations having equal economic prosperity?*
Q3: *Do all the news articles reporting on an event publish from a particular or thesame locations having equal cultures?*
Q4: *Do all the news articles reporting on an event publish from sources with a particular or similar political class?*
Q5: *Do all the news articles reporting on an event publish by the newspapers where the publishing language was same?*

Question 1 (Q1) intends to identify whether the news was published across different geographical places or not. The question is answered "yes" for all the news articles reported on an event if they are published in one country otherwise "no." Question 2 (Q2) intends to identify whether the news was published across different economies or not. The economic similarity has been calculated using Euclidean distance. The question is answered with "information-crossing" for all the news articles reported on an event if they are published from countries with similar economic situations. The question is answered with "unsure" for all the news articles reported on an event if at least one of the news articles is published from a country that is labeled with "unsure" otherwise "information-not-crossing." Question 3 (Q3) intends to identify whether the news was published across different cultures or not. The question is answered with "information-crossing" for all the news articles reported on an event if they are published from countries with a similar culture. The question is answered with "unsure" for all the news articles reported on an event if at least one of the news articles is published from a country that is labeled with "unsure" otherwise "information-not-crossing." The cultural similarity has been calculated using the Euclidean distance (see Section 4). Question 4 (Q4) intends to identify whether the news was published in newspapers with the same political alignments or not. The question is answered "yes" for all the news articles reporting on an event if they are published in the newspapers following similar political alignments, otherwise

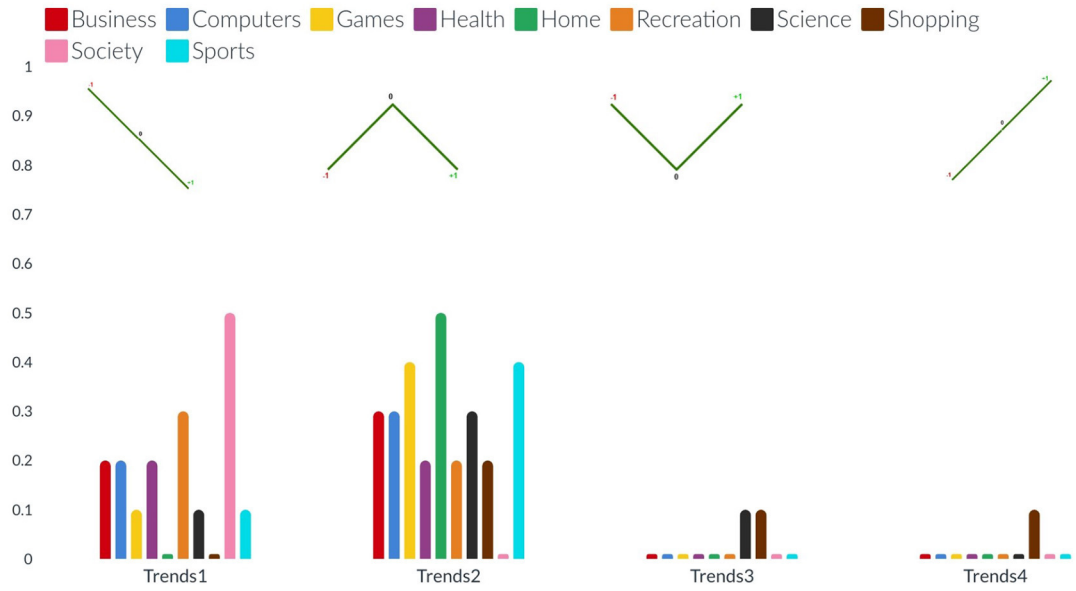


FIGURE 4
The bar charts present the distribution of different possible trends of sentiments across the ten categories (from left to right). The sentimental trends vary in four different types (see on the x-axis): trend1, and trend4 represent decrement and increment respectively in the percentage of news articles (see on the y-axis) with negative sentiment to neutral and then to positive: trend2, and trend3 represent decrement and increment respectively in the percentage of news articles with neutral sentiments than positive and negative sentiments.

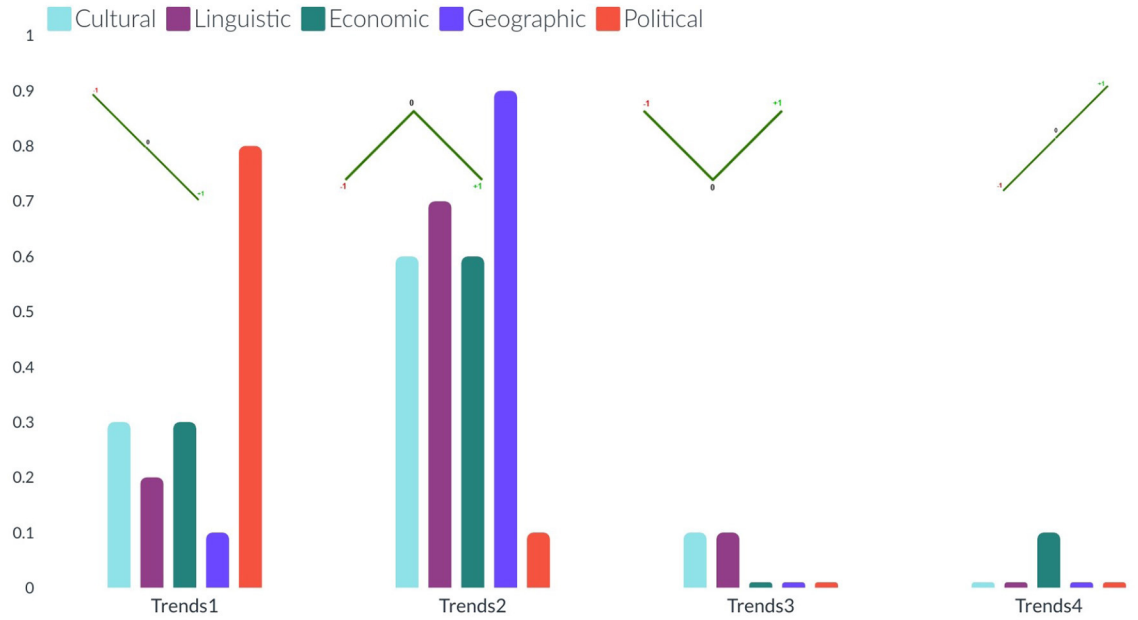


FIGURE 5
The bar charts present the distribution of different possible trends of sentiments across the five barriers (from left to right). The sentimental trends vary in four different types (see on the x-axis): trend1, and trend4 represent decrement and increment respectively in the percentage of news articles (see on the y-axis) with negative sentiment to neutral and then to positive: trend2, and trend3 represent decrement and increment respectively in the percentage of news articles with neutral sentiments than positive and negative sentiments.

"no." Question 5 (Q5) intends to identify whether the news was published in the newspapers where the publishing language was the same or not. The question is answered "yes" for all the news articles reporting on an event if they are published from different newspapers, where the publishing language is the same otherwise "no."

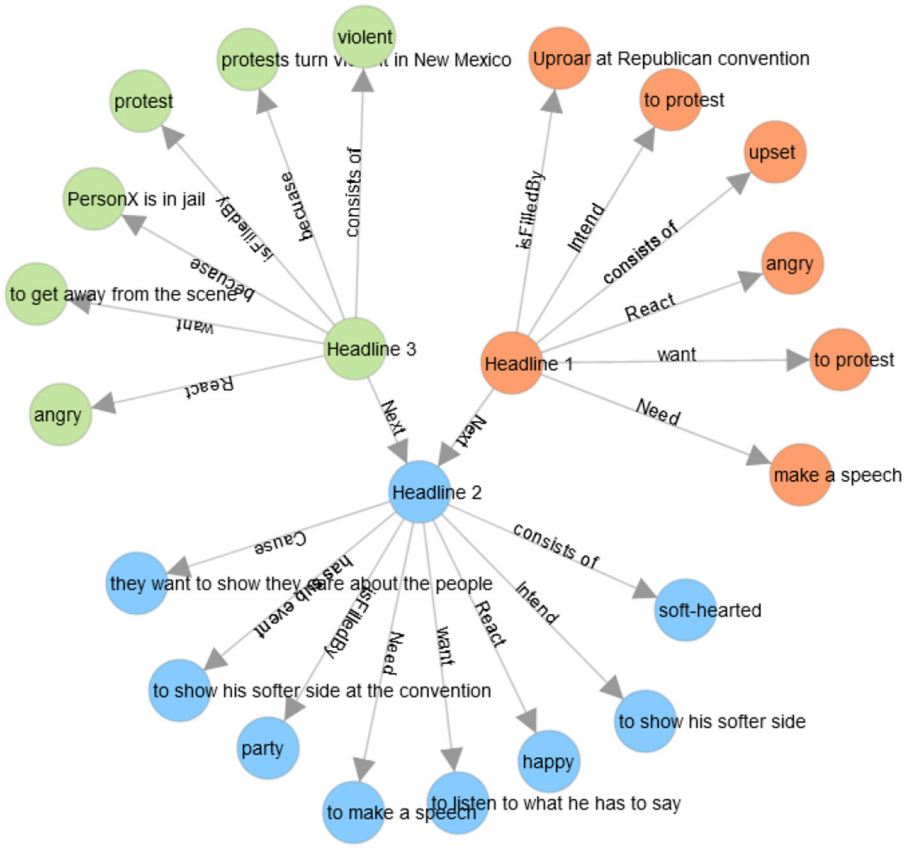


FIGURE 6
The network diagram presents an example of headlines with common-sense knowledge. Headline 1 is "Uproar at Republican convention as anti-Trump delegates revolt," Headline 2 is "Trump aims to show his softer side at Cleveland convention," and Headline 3 is "Protests turn violent outside Trump rally in New Mexico" [these headlines belong to the cultural barrier (see Table 2)].

4.1. Annotated dataset

Initially, we collected approximately 1.7 million news articles. After filtering the news based on the unavailability of the metadata information, the news articles were limited to a few thousand articles. Similarly, based on not having any common sense inferences, the news articles were reduced to a few thousand articles. The number of news articles was reduced from 75 to 96%. The statistics of the news belonging to the 10 categories across the five barriers are presented in Table 1. The dataset is available in the GitHub repository (see text footnote 2). Labels for annotations of the five types of barriers are derived as follows:

- Economic barrier classes: *information-not-crossing*, *unsure*, and *information-crossing*.
- Cultural barrier classes: *information-not-crossing*, *unsure*, and *information-crossing*.
- Geographical barrier classes: *Not-crossed-GB* and *Crossed-GB*.
- Political barrier classes: *Not-crossed-PB* and *Crossed-PB*.
- Linguistic barrier classes: *Not-crossed-LB* and *Crossed-LB*.

5. Materials and methods

In this section, we present an analysis of sentiments across different barriers, followed by the properties of common-sense inference knowledge, classification baselines, and evaluation metrics.

5.1. Analysis of sentiments

We use the Vader rule-based model to obtain the emotional and sentiment polarity of the news headlines to analyze the variation of sentiments across the different categories of the different barriers. Vader provides a polarity range for the news headlines in the interval from −1 to +1. The −1 value represents a negative polarity, and +1 indicates a positive polarity (Martín et al., 2021). The bar charts illustrate the differences in sentiments across binary and ternary classes in two categories of the three barriers (see Figures 2, 3). For each instance, we have one of the three sentiments, such as positive, neutral, or negative.

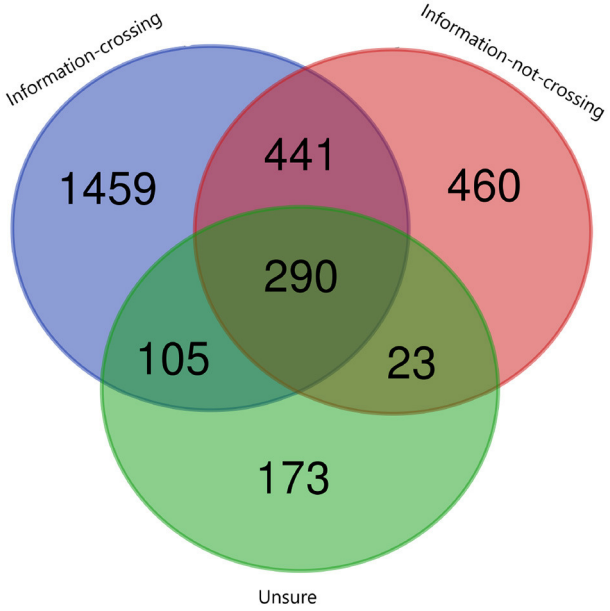


FIGURE 7
The Venn diagram shows the intersection between inferences across ternary classes of the cultural barrier.

For the binary class classification of the political, linguistic, and geographical barrier, the headlines that have been labeled as crossing the barrier have the following sentimental differences: The categories business, home, health, recreation, science, shopping, and society have more instances of negative sentiments than positive and neutral, with considerable differences of (8, 1, 10, 5, 8, 5, and 5%), (6, 5, 2, 5, 8, 5, and 5%), and (7, 12, 4, 12, 9, 3, and 5%), respectively. The game category of news headlines with annotations of crossing the political barrier has 12, 6, and 1% more instances of positive sentiments for the political, linguistic, and geographical barriers; the news headlines that have been labeled as not crossing the political barrier have the following differences: the categories computers, health, recreation, science, society, and sports have more instances of positive sentiments than neutral and negative sentiments, with considerable differences of (5, 8, 5, 5, 2, and 5%), (7, 2, 5, 5, 2, and 3%), and (4, 2, 1, 8, 3, and 5%), respectively; the game category has 10, 1, and 3% more instances, respectively, with negative instances than other classes; With regard to the ternary classification of the cultural barrier, the news headlines that have been labeled as crossing the barrier have the following sentimental differences: the categories games, health, shopping, and society have more instances of negative sentiments than other classes, with considerable differences of 12, 5, 6, and 2% respectively. The news headlines that have been labeled as not crossing the cultural barrier have the following differences: the categories business, computers, health, recreation, science, shopping, society, and sports have more instances of positive sentiments than other classes with considerable differences of 5, 6, 5, 5, 10, 3, and 8%, respectively. The news headlines that have been labeled as unsure have the following differences: the categories computers and shopping have more instances of positive sentiments than other classes, with considerable differences of 6 and 5%, respectively, whereas the category game has 20% more instances of negative sentiments; with regard to the ternary classification of the economic barrier, the headlines that have been labeled as crossing the barrier have the following sentimental differences: the categories business, home, recreation, science, shopping, and sports have more instances of negative sentiments than other classes, with the considerable differences of 5, 7, 8, 5, 2, and 14%, respectively. The news headlines that have been labeled as not crossing the economic barrier have the following differences: games, home, recreation, science, and shopping have more instances of positive sentiments than other classes, with considerable differences of 18, 16, 7, 5, and 8% respectively. The headlines that have been labeled as unsure have the following differences: The categories business and games have 5 and 22% respectively, more instances of negative sentiments, computers, and sports have 8 and 10% respectively more instances of positive sentiments; and recreation and shopping have 12 and 13% more instances of neutral sentiment, respectively.

Overall, with regard to the binary class classification for the political, linguistic, and geographical barriers, we see that the news headlines that are labeled as crossing the barrier, have more instances of negative sentiments, whereas the news headlines that are labeled as not crossing the barrier, have more instances of positive sentiments. With regard to the ternary class classification for the economic and cultural barrier, we see that the news headlines that are labeled as crossing the barrier have more instances of negative sentiments, whereas the news headlines that are labeled as not crossing the barrier have more instances of positive sentiments. However, in the case of news headlines that are labeled as unsure, there are more instances of negative sentiments for the economic barrier and positive sentiments for the cultural barrier.

The bar charts present the distribution of different possible trends of sentiments across the ten categories and five barriers (see Figures 4, 5). The purpose of this figure is to show the investigation that we did while finding the variations of sentiments across different categories and barriers. It tells the readers what type of news has more positive, neutral, or negative polarity. It especially helps us across the barriers to know what type of barriers are crossing positive or negative news. We analyzed the sentimental trends and found that, among other possible trends, these four trends cover more than 95% of the data. The first trend shows that the number of positive instances is higher than the number of neutral instances, and then both are higher than the number of negative instances. The fourth trend is the reverse of it. The second trend shows that the number of neutral instances is higher than that of positive and negative instances, and negative and positive instances are approximately equal to each other. The first bar chart shows that more than 30% of news of society and recreation categories consist of news headlines with positive sentiment, whereas more than 30% of news of games, home, and sports categories consist of news headlines with neutral sentiments. The second line graph shows that 80% of news headlines belonging to the political barrier have negative sentiment, whereas ninety 90% of news headlines belonging to the geographical barrier have neutral sentiments.

The results suggest the following conclusions for the **Q1**: (1) The political barrier has been crossed by the news with positive sentiments and reversed for the other four barriers (linguistic,

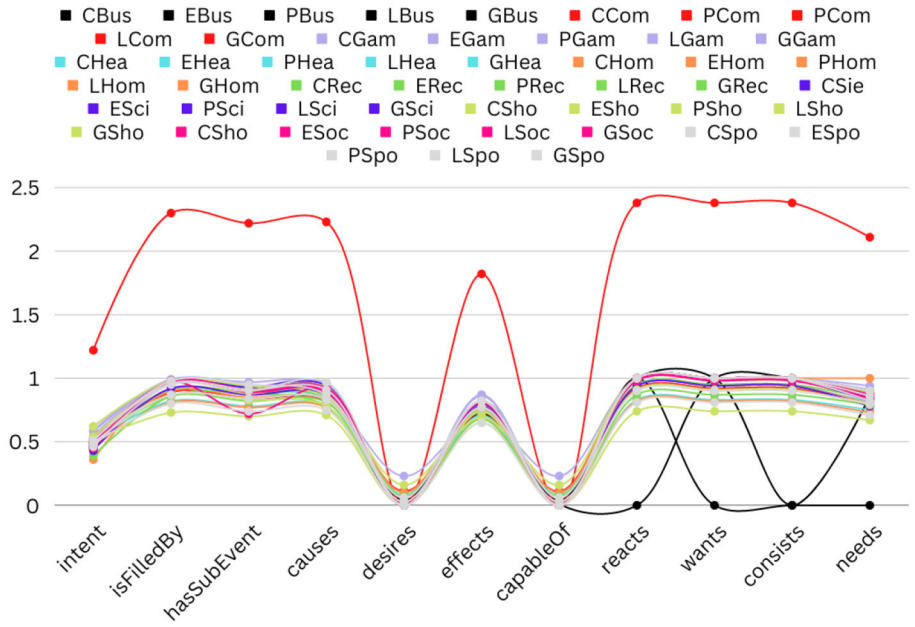


FIGURE 8
The line graph shows the frequency of all the inferences to all categories of all the barriers. The list of inferences has been shown on the x-axis, whereas the average number of inferences per news headline has been shown on the y-axis.

TABLE 1 Statistics of the news articles based on common-sense knowledge extraction and data annotation for the 10 categories (business, computers, games, health, home, recreation, science, shopping, society, and sports) of the five barriers.

Categories	Cultural	Economic	Geographic	Linguistic	Political
Business	3,455	1,015	3,550	7,974	7037
Computers	913	310	1,181	2,194	1895
Games	186	68	549	504	316
Health	1,159	90	1,533	3,295	3368
Home	1,065	86	1,321	2,796	3258
Recreation	1,695	161	1,783	3697	3236
Science	4,377	378	7,877	14,665	14925
Shopping	513	42	796	1,685	1287
Society	14,238	1,003	13,472	28,431	28447
Sports	1,533	39	1,021	2,054	1289

geographic, cultural, and economic). The news with negative sentiments has not been crossing the political barrier but has been crossing the linguistic, geographic, cultural, and economic barriers; (2) the variations in the sentiments across binary and ternary class classifications of the different categories of news and the barriers suggest that we should take sentiment score as a feature in barrier classification. Alonso et al. (2021) have considered sentiments of news for fake news detection based on the fact that sentiment is a complementary element to fake news.

5.2. Common-sense knowledge

We use the common sense knowledge resource COMET atomic through an inference toolkit called *kogito* to generate common-sense inferences in a given situation by assessing their intentions and behaviors. This toolkit provides the interface to interact with natural language generation models that can be used to infer common-sense from a textual input (Hwang et al., 2021; Ismayilzada and Bosselut, 2022). These models consist of triplets of head entity, relation, and tail entity. We present an example illustrating the results of the common-sense of different relations about three news headlines (taken from the Table 2). The first headline ("Uproar at Republican convention as anti-Trump delegates revolt") has six relations such as *react, need, intend, want, isFilledBy, and react*. To convert common-sense knowledge into a meaningful text, we consider each tuple consisting of the relation and tail as a sentence and then concatenate them. To make the tuples as sentences, we change the relation to the past form, such as reacted angry, needed to make a speech, intended to protest,

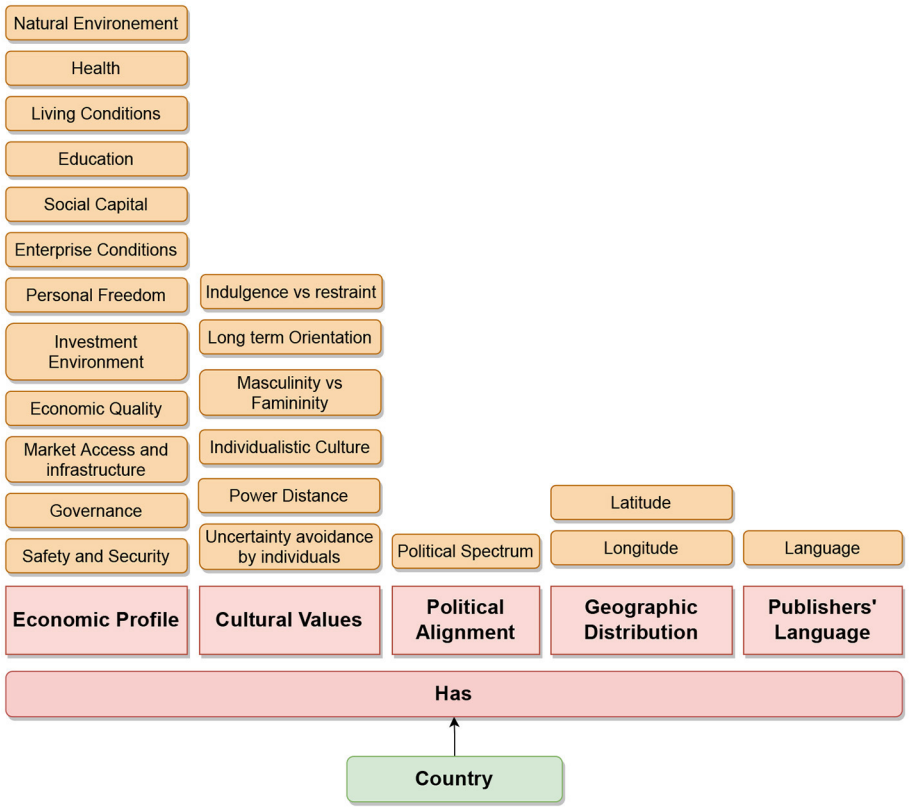


FIGURE 9
Metadata for the five barriers (cultural, economic, geographical, linguistic, and political).

wanted to protest, isFilledBy uproar at the Republican Convention, and reacted upset.

The purpose of using semantic knowledge in the form of common-sense knowledge was to improve text classification. We analyzed the associated inferences to all the barriers. We present an example to illustrate the comparison. We choose the cultural barrier, and to perform a comparison between the categories, we select the category of society. The results of the intersection between the inferences belonging to three different classes (information crossing, information not crossing, and unsure) have been shown in Figure 7. There are 290 inferences that are common among all three classes, and there are 441, 105, and 23 inferences that are common between classes one and two, class two and three, and class one and three, respectively. The most important fact is that there are 1,459, 173 and 460 unique inferences for classes one, two, and three, respectively, that can be useful for the classification in this ternary class classification.

To answer the **Q2** the line graphs in Figure 8 present the statistics of inferences across the ten different categories of the five barriers. Since the main purpose of this figure is to analyze the statistics of the inferences across the categories, we keep the same color for a category across the five barriers. The x axis shows the names of all the inferences, and the y axis shows the average number of inferences per news headline in each category. The categories that have significant differences are computer and business. The average inferences are significantly higher in the computer category than in all other categories. Each news headline contains 1.5% inference type "intent" and consists of "isFilledBy," "hasSubEvent," "Causes," "reacts," "wants," "consists, and "needs" inference types with an average of approximately 2.5. The existence of the inference type is almost 0% per news headline for "desires" and "capableOf". For the category business, the existence of a few types of inferences is equal to zero, such as "reacts," "wants," "consists," and "needs." Otherwise, the average of the existence of all the inferences per news headline is approximately equal for all the categories of all the barriers.

This analysis helps us to understand the distribution of different inferences across different categories, as well as the associated semantic knowledge per news headline. Since different features, such as sentiments and semantic knowledge, possess different discriminative capabilities in classification (Zhai et al., 2011; Nassirtoussi et al., 2015), we use them as classification of barriers to the spreading of news.

5.3. Evaluation methodology and baselines

We used the Scikit-learn implementation of classical and deep learning models, considering the following parameters, which are usually the default: hidden layers = 3, hidden units = 64, no. of epochs = 10, batch size = 64, and dropout = 0.001. We provide the

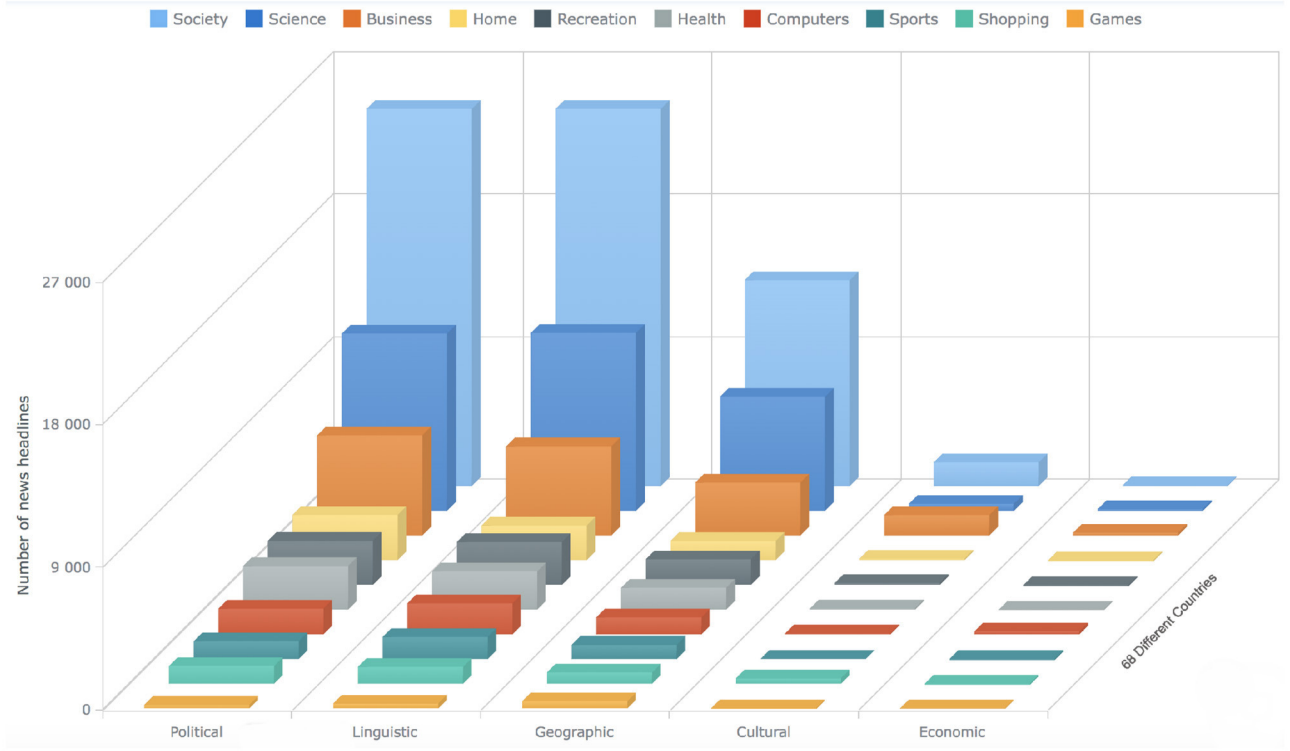


FIGURE 10
The bar chart shows the statistics about the news articles that has the label "Information-crossing" for all the 10 different categories. The x-axis shows the different barriers, the y-axis shows the count of the news articles, whereas the bars on the z-axis represent 10 different categories (see Section 4).

pseudocode of the classification models along with the features in Algorithm 1 and present a detailed description of each component. For the training process of the political, geographical, and linguistic barrier, we used Adam as the optimizer, binary cross-entropy as the loss function, and sigmoid as the activation function. For economic and cultural barriers, we used Adam as the optimizer, categorical cross-entropy as the loss function, and SoftMax as the activation function. The data about each barrier is split into train-sets and test-sets with a ratio of 80–20%. To maintain the class proportion in the train and test sets, we use stratified sampling. It means that the training and testing sets must have an equal proportion of all the classes. The Figures 13, 14 show the class distribution for each category of a barrier. For instance, in the case of the business category of the culture barrier, there are a total of 100 news headlines where 40 instances have the label "information crossing," 40 instances have label "information not crossing," and 20 instances have the label "Unsure." And, we suppose, we split our train and test sets in the ratio of 80:20. Then, the train set and test set must have 20, 20, and 10 instances of each label ("information crossing," "information crossing," and "Unsure") respectively.

For comparison with the proposed common sense inferences and semantic knowledge, we evaluated the barrier classification task using the news headline text only. After performing the preprocessing steps such as lower case conversion and stop word removal, we adopted the term frequency (TF) and inverted document frequency (IDF) methods to represent the bag of words in each news article (Yazdani et al., 2017). For the barrier classification task, the experiments were conducted by utilizing three different types of machine learning algorithms: (1) classical machine learning algorithms, including Logistic Regression (LR), Naive Bayes (NB), Support Vector Classifier (SVC), k-nearest Neighbor (kNN), and Decision Tree (DT): The performance of LR for text classification problems is the same as that of the SVM algorithm (Shah et al., 2020). SVMs use kernel functions to find separating hyper-planes in high-dimensional spaces (Colas and Brazdil, 2006). SVM is difficult to interpret, and there have to be many parameters that need to be set for performing the classification and one parameter that performs well in one task might perform poorly in other tasks (Shah et al., 2020). Therefore, many information retrieval systems use decision trees and naive bayes. However, these models lack accuracy (Kowsari et al., 2017; Kamath et al., 2018). (2) LSTM (long-short-term memory): With the emergence of deep learning algorithms, the accuracy of text categorization has been greatly improved. Convolutional neural networks (CNN) and long short-term memory networks (LSTM) are widely used (Wang et al., 2017; Kamath et al., 2018; Luan and Lin, 2019; Yu et al., 2020). (3) The state-of-the-art pre-training language model BERT (Bidirectional Encoder Representations from Transformers): It is trained on a large network with a large amount of unlabeled data and adopts a fine-tuning approach that requires almost no specific architecture for each end task and has achieved great success in a couple of NLP tasks, such as natural language inference, and text classification (Yu et al., 2019; González-Carvajal and Garrido-Merchán, 2020; Jin et al., 2020).

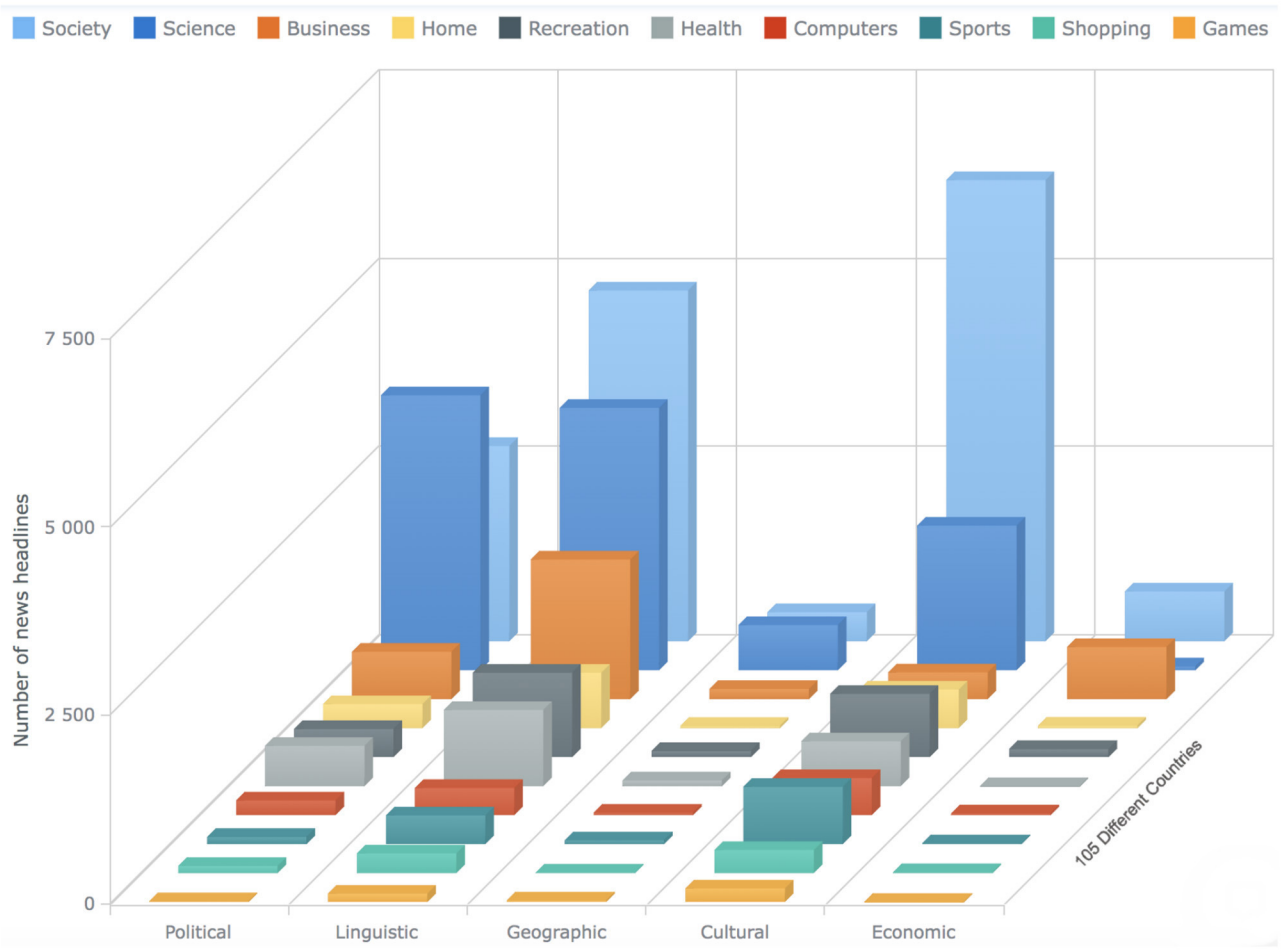


FIGURE 11
The bar chart shows the statistics about the news articles that has the label "Information-not-crossing" for all the 10 different categories. The x-axis shows the different barriers, the y-axis shows the count of the news articles, whereas the bars on the z-axis represent 10 different categories (see Section 4).

5.4. Evaluation metric

To evaluate the performance of binary and multi-class barrier classification models, the F1-score is used as an evaluation measure.

- **F1-score:** It combines the precision and recall of a classifier into a single metric by taking their harmonic mean. It is defined as:

$$F_1 = \frac{2(Precision * Recall)}{Precision + Recall}$$

6. Results

In this section, we present the experimental results comparing simple (LR, SVM, DT, RF, kNN), deep learning (LSTM), and transformers (BERT) for the barrier classification task.

6.1. Comparative analysis of the ten categories

We compare the results of all ten news categories based on the evaluation metric F1-score. Since the results of LR among the five (LR, SVC, NB, DT, and kNN) classical machine learning algorithms were higher in all categories, we exclude the others. Table 3 compares the results of LR, LSTM, and BERT with our proposed approach that is based on common-sense-based semantic knowledge and sentiment. The words PM-LSTM (proposed model LSTM) and PM-BERT (proposed model BERT) mean the usage of LSTM and BERT utilizing our approach with the inference-based semantic knowledge and sentiments. For the cultural barrier, F1-scores using BERT or LSTM with common-sense-based semantic knowledge and sentiment are higher than LR, LSTM, and BERT for business, computers, games, health, home, recreation, science, shopping, society, and sports (with an improvement of 0.02, 0.05, 0.01, 0.09, 0.11, 0.14, 0.09, 0.12, 0.06, and 0.03, respectively). For the economic barrier, F1-scores are higher than LR, LSTM, and BERT for business, computers, health, home, and sports (with an improvement of 0.06, 0.03, 0.1, 0.14, and 0.01, respectively). For the

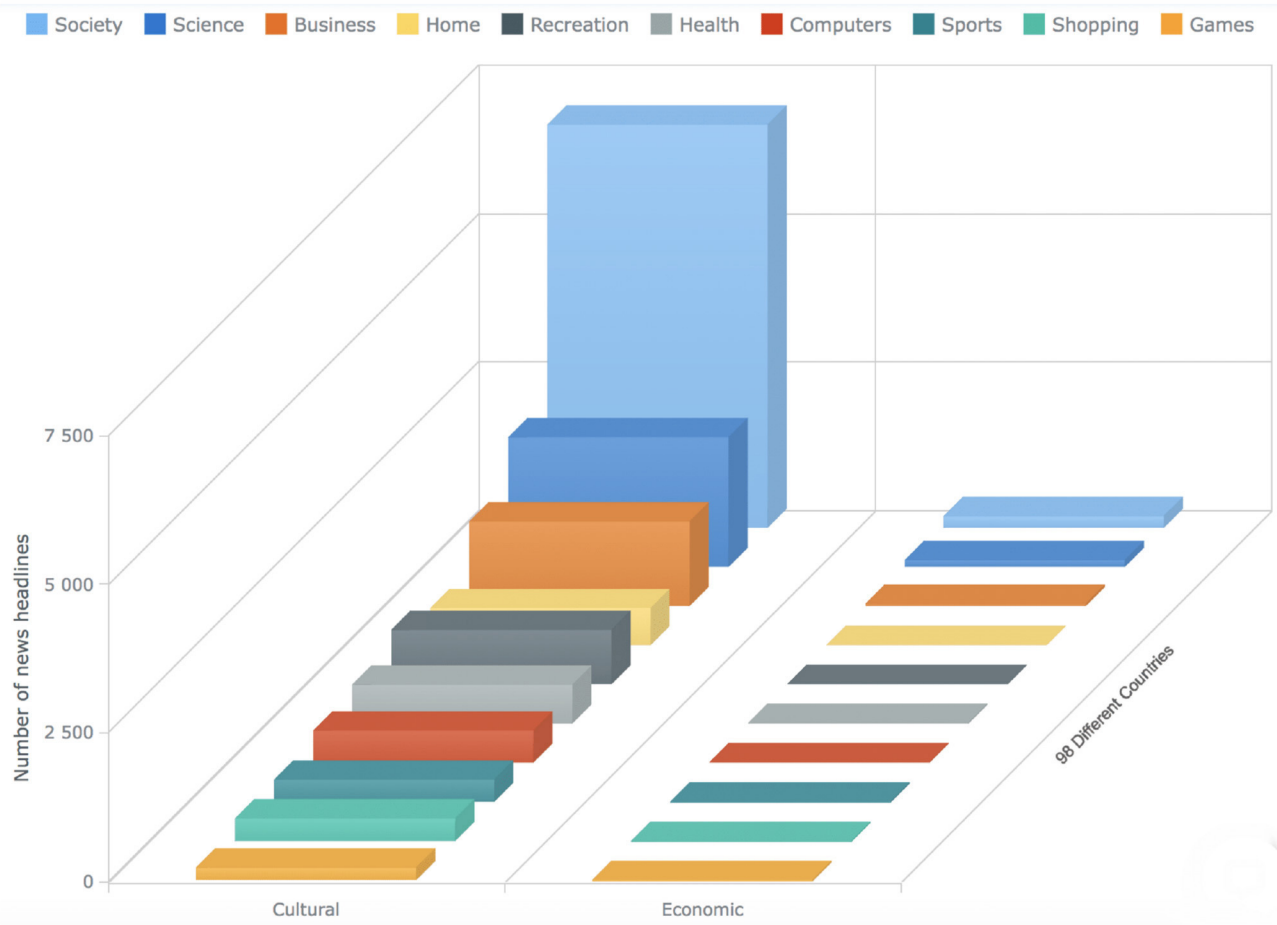


FIGURE 12
The bar chart shows the statistics about the news articles that has the label "Unsure" for all the ten different categories. The x-axis shows the different barriers, the y-axis shows the count of the news articles, whereas the bars on the z-axis represent ten different categories (see Section 4).

political barrier, F1-scores are higher than LR, LSTM, and BERT for business, computers, games, health, home, recreation, science, shopping, and society (with an improvement of 0.12, 0.28, 0.08, 0.13, 0.21, 0.07, 0.1, 0.14, and 0.24, respectively). For the linguistic barrier, F1-scores are higher than LR, LSTM, and BERT for science, and society (with an improvement of 0.26 and 0.24, respectively). For the geographical barrier, F1-scores are higher than LR, LSTM, and BERT for business, computers, health, home, recreation, and society (with an improvement of 0.03, 0.44, 0.25, 0.17, 0.26, and 0.28, respectively).

The results presented in Table 3 indicate that the best results for each barrier are obtained by PM-BERT and BERT, closely followed by PM-LSTM. The LR performs a little less compared to the other algorithms tested. Moreover, it can be seen that the obtained F1-scores vary significantly across different categories. While the obtained F1-score is very high for the two categories (health and society) of the geographical barrier and for the three categories (business, shopping, and science) of the political barrier, and a quite good score is obtained for the recreation category of the cultural barrier, the games category of the economic barrier, and the computers category of the linguistic barrier, the score is low comparatively for the other categories of the different barriers.

The best results obtained for the task of classifying the barriers for the ten different categories are a direct consequence of the class distribution and sentiments of the classes, to some extent. As far as the results are concerned for all the barriers, we see that the highest F1-score is produced for the health (0.97) and society (0.97) categories of the geographical barrier, recreation (0.66) category of the cultural barrier, games (0.72) category of the economic barrier, computers (0.97) category of the linguistic barrier, and business (0.97), shopping (0.97), and science (0.97) categories of the political barrier. The F1-score for the society category of the geographic barrier and business category of the political barrier is as high as 0.97. An obvious reason for this is the fact that the data is heavily imbalanced, with 95 and 91% instances of majority classes. However, both are showing improvements. This can be due to a slight variation in sentiments across its binary classes. For the health category of the geographical barrier, and the shopping and science category of the political barrier, the class distribution is not very imbalanced (78, 85, and 75% instances of the majority class), but the F1-score is really high, which means PM-BERT is best suited for these categories. Regarding its best results, it might be possible that sentiments across these binary classes have variations, such as the label "Crossed-GB"

TABLE 2 Examples of annotation for all five types of barriers.

Barrier (Category)	Time	Title	Location/Publisher/ Language	Meta-data	Class
Cultural (Games)	2016-07-18T19:48:00Z 2016-07-18T22:04:00Z	Trump aims to show his softer side at Cleveland convention Uproar at Republican convention as anti-Trump delegates revolt	Ireland (irishtimes.com) Thailand (bangkokpost.com)	Same Culture	Information -not-crossing
	2016-04-16T22:23:00Z 2016-04-18T16:02:00Z	Another small victory for Cruz Romney: 3-man race throws Trump the nomination	New Zealand (odt.co.nz) United States (wcyb.com)	Different Culture	Unsure
	2016-05-25T05:19:00Z 2016-05-25T11:40:00Z	Protests turn violent outside Trump rally in New Mexico Protests turn violent outside Trump rally in New Mexico	japan (japantoday.com) United States (newschannel5.com)	Different Culture	Information -crossing
Economic (Recreation)	2016-07-13T21:36:00Z 2016-07-16T19:59:00Z	File to seek Gulen's US extradition ready Erdogan calls on Barack Obama to extradite Fethullah	Azerbaijan (en.trend.az) Armenia (news.am)	Similar economic Situations (ES)	Information -not-crossing
	2018-03-17T20:46:00Z 2018-03-19T17:14:00Z 2018-03-22T01:50:00Z	Trump consultants harvested data from 50 million Facebook users: reports Officials question Facebook's protection of personal data Ex-Facebook manager says company was sluggish in stopping data harvesting	Pakistan (geo.tv) United States (union-bulletin.com) Kenya (businessdailyafrica.com)	Different ES	unsure
	2018-04-03T17:19:00Z 2018-04-04T18:13:00Z	Trump seeks Syria pullout as advisers warn on Islamic State White House appears to delay Trump's order for Syrian withdrawal	Egypt (english.ahram.org.eg) Iraq (kurdistan24.net)	Different ES	Information -crossing
Political (Society)	2021-04-07T21:55:00Z 2021-04-08T06:52:00Z 2021-04-09T03:22:00Z	Thugs petrol bomb bus as violent riots continue in Belfast Thugs petrol bomb bus as violent riots continue in Belfast Police use water cannon during continued unrest in belfast	conservatism (thesun.ie) conservatism (thesun.ie) centre-right (theaustralian.com.au)	Similar political alignment (PA)	information -not-crossing
	2016-01-06T14:51:00Z 2016-01-08T01:49:00Z 2016-01-09T01:08:00Z	Iraq offers to mediate in Saudi-Iran crisis stemming from cleric's execution Iran not seeking tension with Saudi Arabia: Zarif Hammond fails to condemn Saudia political executions	Neutral (brandonsun.com) Conservatism (tehrantimes.com) Left-wing (morningstaronline.co.uk)	Different PA	information -crossing
Linguistic (Society)	2016-01-15T14:50:00Z 2016-01-15T17:51:00Z 2016-01-16T02:20:00Z	Why Amal Clooney doesn't think she's a celebrity Amal Clooney talks about her new celebrity status for the first time Amal Clooney sits down for First U.S. TV interview Watch!	English (pagesix.com) English (vanityfair.com) English (usmagazine.com)	Similar publishing Language (PL)	information -not-crossing
	2016-01-15T17:38:00Z 2016-01-15T14:00:00Z 2016-01-15T16:27:00Z	Friendly no more: Trump, Cruz erupt in bitter fight at Republican debate The fight everyone wanted to see finally happened The 6th republican debate in 100 words (and 4 videos)	Spanish (ecodiario.eleconomista.es) English (charismanews.com) English (scpr.org)	Different PL	information -crossing
Geographic (Society)	2021-04-28T11:48:00Z 2021-04-28T22:52:00Z	Lawmaker says schools must teach the "Good of slavery" (Video) Backlash on Louisiana lawmaker grows following his comments about slavery	United States (patheos.com) United States (theadvertiser.com)	Publishers' headquarters in same country	information -not-crossing
	2016-06-10T11:21:00Z 2016-06-10T14:29:00Z 2016-06-10T15:18:00Z	Queen's dedication praised at 90th Queen's dedication praised at 90th Queen's dedication praised at 90th	England (haverhillecho.co.uk) United Kingdom (newsletter.co.uk) Australia (adelaidenow.com.au)	Publishers' headquarters in different country	information -crossing

The annotation is performed using the meta-data.

having more positive and fewer negative instances than "Not-crossed-GB" and vice versa. Similarly, the shopping and science categories of the political barrier consist of more news headlines with negative sentiments for the label "Crossed-PB" and vice versa for the label "Not-crossed-PB." PM-BERT has proved to be best suited for the classification of computers and games categories of the linguistic and the economic barriers. We see that the data is quite balanced for computers category of the linguistic (75 and 25% instances of "Crossed-LB," "Not-crossed-LB" respectively) and games category of the economic barrier (29, 49, and 22% instances of "information-crossing," "information-not-crossing," and "unsure" classes, respectively). Looking into the sentiments of each class of computers category (see Figures 2, 3), we observe that one has more news headlines with positive headlines and

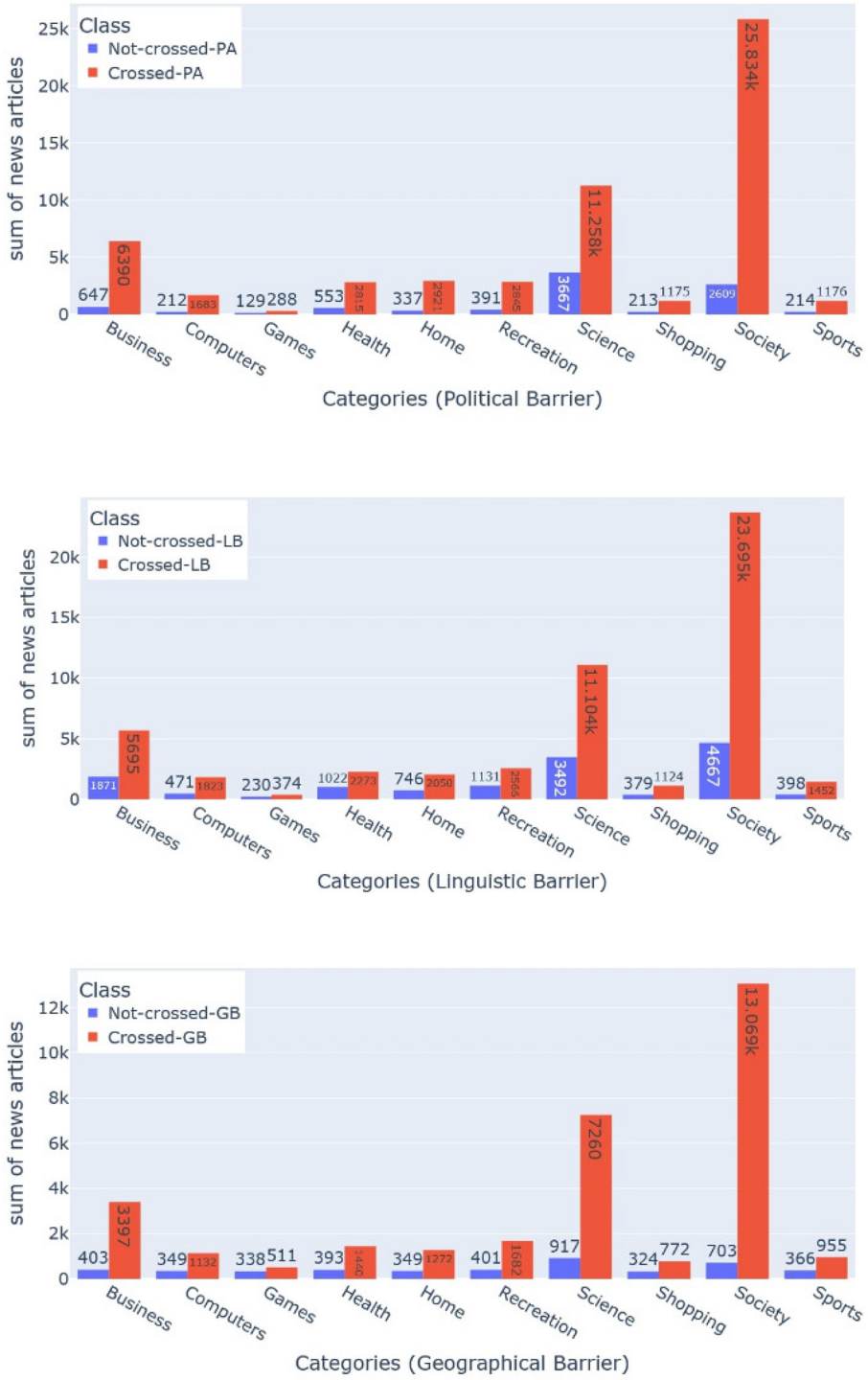


FIGURE 13
This bar charts show the class distribution for the political, linguistic, and geographical barriers (from left to right). The bar with blue color shows the distribution for the class "Information-crossing" a barrier, whereas the bar with orange color shows the distribution for the class "Information-not-crossing" a barrier. Each of the three-bar charts presents the class distribution for all ten categories.

less negative news headlines and vice versa. However, for the games category of the economic barrier, sentiments are varying across all three classes: 12, 25, and 50% news headlines with positive, neutral, and negative sentiments, respectively, for the label "Information-crossing." The label "Unsure" does not have news headlines with neutral sentiments, whereas all the news headlines labeled "Information-not-crossing" have only positive sentiments. For the recreation category of the cultural barrier, although the distribution of positive, neutral, and negative sentiments across all three barriers is almost equal and the class distribution is

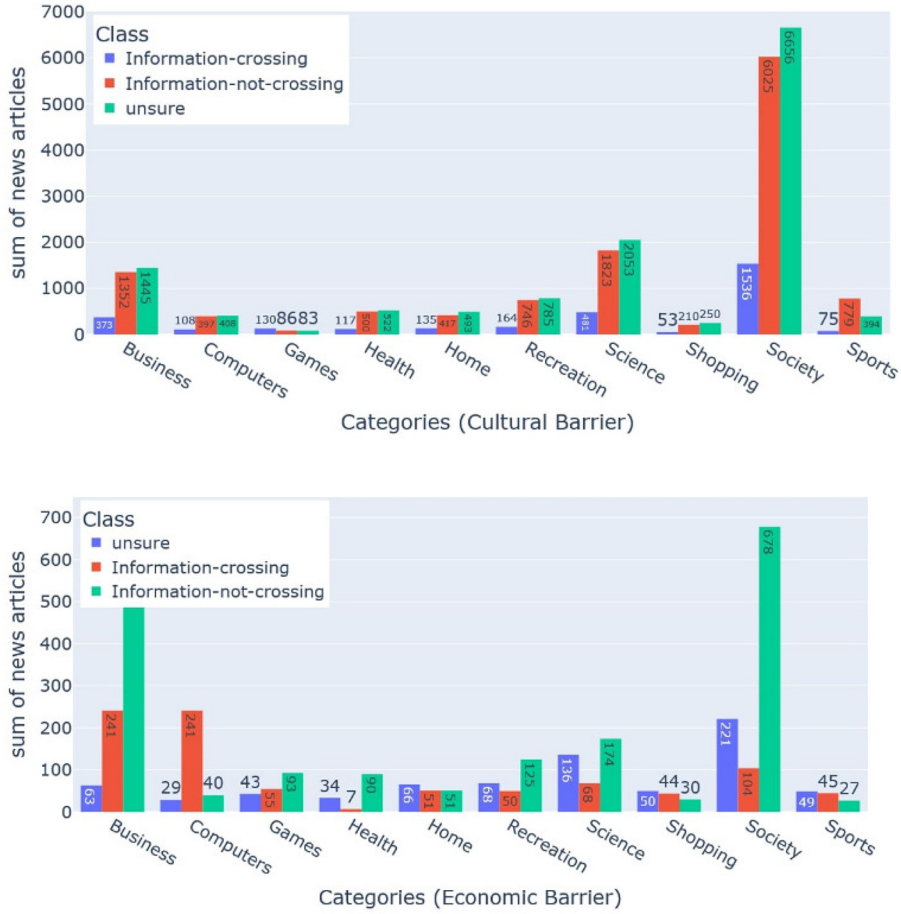


FIGURE 14
This bar chart shows the class distribution for economic, and cultural barriers (from left to right). The bar with orange color shows the distribution for the class "Information-not-crossing" whereas the bar with green color shows the distribution for the class "Unsure" a barrier. The bar with the blue color shows the distribution for the class "Information-crossing." Each of the two bar charts presents the class distribution for all ten categories.

balanced (10, 44, and 46% instances for "Information-crossing," "Information-not-crossing," and "Unsure," respectively, the PM-BERT performs really well (0.66 F1-score).

6.2. Comparative analysis of three types of algorithms

After discussing the results of all ten news categories, we compare all five different types of barriers using the average F1-scores of all ten categories. Figure 15 presents the comparison of the average F1-score of all the categories. The highest average F1-score for the cultural barrier is obtained using PM-BERT (0.47) and BERT (0.38), whereas this score was low using PM-LSTM (0.22). For the economic, political, and geographical barriers, the average F1-score obtained using PM-BERT (0.55, 0.70, and 0.76 respectively) was followed by a slight lower average F1-score using BERT (0.48, 0.59, and 0.60, respectively) and then PM-LSTM (0.28, 0.49, and 0.57, respectively).

With regards to the linguistic barrier, the highest average F1-score was achieved using BERT instead of PM-BERT or PM-LSTM, which is quite interesting and questionable. Instead of average, we look for the F1-score of all the ten categories of this barrier. The obtained F1-score for the eight categories (business, computers, health, home, science, shopping, sports, and society) using PM-BERT is higher then that using BERT (using PM-BERT : 0.80, 0.97, 0.82, 0.83, 0.78, 0.67, 0.73, and 0.79, respectively; using BERT : 0.74, 0.75, 0.79, 0.81, 0.78, 0.63, 0.73, and 0.76, respectively). However, the F1-score for the game and recreation categories of the linguistic barrier using PM-BERT (0.47 and 0.48, respectively) was considerably lower than BERT (0.85 and 0.81, respectively). On the other hand, the obtained average F1-score is better for four barriers, including economic, political, linguistic, and geographical, than the cultural barrier, which is just under 0.5. However, for the other barriers, it is well over 0.50 and extends to near 0.80. To answer the **Q3**, we can say that our proposed methods (LSTM and BERT with semantic knowledge) outperform for the four cultural, economic, political, and geographical barriers.

7. Analysis and discussion

Experiments with the novel approach on the ten different kinds of news and the five different barriers have brought some

TABLE 3 F1-score of the five different machine learning algorithms (LR, LSTM, BERT, PM-LSTM, and PM-BERT) for the ten different categories (business, computers, games, health, home, recreation, science, shopping, society, and sports).

Model	Category	Cul	Eco	Pol	Ling	Geo	Category	Cul	Eco	Pol	Ling	Geo
LR	Business	0.40	0.48	0.71	0.73	0.61	Recreation	0.37	0.30	0.59	0.73	0.57
	Computers	0.42	0.35	0.58	0.63	0.56	Science	0.42	0.42	0.62	0.71	0.65
	Games	0.52	0.35	0.59	0.59	0.60	Shopping	0.36	0.27	0.49	0.61	0.52
	Health	0.36	0.40	0.58	0.67	0.64	Society	0.40	0.45	0.62	0.68	0.62
	Home	0.39	0.57	0.49	0.68	0.59	Sports	0.44	0.28	0.59	0.62	0.57
LSTM	Business	0.19	0.28	0.49	0.55	0.47	Recreation	0.20	0.39	0.48	0.41	0.47
	Computers	0.20	0.08	0.49	0.52	0.46	Science	0.20	0.20	0.48	0.43	0.43
	Games	0.14	0.29	0.74	0.70	0.48	Shopping	0.23	0.44	0.49	0.44	0.49
	Health	0.18	0.17	0.49	0.59	0.53	Society	0.20	0.27	0.49	0.46	0.48
	Home	0.19	0.21	0.49	0.59	0.63	Sports	0.26	0.43	0.48	0.43	0.49
BERT	Business	0.42	0.54	0.62	0.74	0.50	Recreation	0.32	0.29	0.74	0.81	0.59
	Computers	0.38	0.30	0.47	0.75	0.66	Science	0.39	0.47	0.68	0.78	0.60
	Games	0.40	0.65	0.77	0.85	0.68	Shopping	0.34	0.44	0.49	0.63	0.50
	Health	0.38	0.54	0.66	0.79	0.67	Society	0.39	0.51	0.49	0.73	0.48
	Home	0.32	0.60	0.54	0.81	0.67	Sports	0.48	0.50	0.48	0.76	0.66
PM-LSTM	Business	0.19	0.28	0.49	0.43	0.48	Recreation	0.21	0.47	0.48	0.44	0.46
	Computers	0.21	0.06	0.49	0.46	0.47	Science	0.19	0.21	0.48	0.74	0.43
	Games	0.18	0.22	0.48	0.40	0.48	Shopping	0.18	0.50	0.49	0.50	0.48
	Health	0.20	0.27	0.48	0.43	0.46	Society	0.19	0.25	0.49	0.64	0.48
	Home	0.20	0.27	0.49	0.45	0.48	Sports	0.27	0.22	0.48	0.44	0.48
PM-BERT	Business	0.48	0.66	**0.97**	0.80	0.96	Recreation	**0.66**	0.35	0.74	0.48	0.65
	Computers	0.46	0.28	0.54	**0.97**	0.70	Science	0.47	0.53	**0.97**	0.78	0.66
	Games	0.33	**0.72**	0.48	0.47	0.72	Shopping	0.45	0.50	**0.97**	0.67	0.56
	Health	0.46	0.60	0.72	0.82	**0.97**	Society	0.52	0.69	0.55	0.73	**0.97**
	Home	0.41	0.66	0.54	0.83	0.73	Sports	0.49	0.50	0.54	0.79	0.63

The bold values indicate the highest F1-scores that have been achieved by a method for a category.

insights regarding information spreading. In order to support the hypothesis, we have set three research questions.

To answer the first research question (Do the sentiments of the headlines of different topics vary across the different barriers?), We compare the sentiments of the headlines for all the categories across the five barriers, performing sentiment analysis at the granularity of ten negative and ten positive points as well as at overall negative, positive, and neutral sentiment (see Figures 2–5). The comparative analysis indicates that the political barrier has been crossed by the news with positive sentiments whereas for the other four barriers (linguistic, geographical, cultural, and economic), news with positive sentiments is not crossing the barriers. With regard to the binary class classification for the political, linguistic, and geographical barrier, we see that the news headlines that are labeled as crossing the barrier have more instances of negative sentiments, whereas the news headlines that are labeled as not crossing the barrier have more instances of positive sentiments. With regard to the ternary class classification for the economic and cultural barriers, the sentiments were the same as in the binary class classification, where the news headlines are labeled as crossing or not crossing the barriers. However, for the news headlines that are labeled as unsure, there were negative and positive sentiments about the economic and cultural barriers, respectively. The implication for the general audience is that the news with different categories cross the barriers. But the news that cross the political barrier are mostly of positive sentiments. Moreover, another interesting method that can be helpful to support this classification is to have domain-specific phrases. For instance, in case of economic barrier and science category, a pre-defined phrases along with their semantic description can be detected in the headlines. These phrases can further support this classification of either what type of science-related news are crossing the economic barrier or vice verse. One of the available datasets includes *Fintech key-phrase dataset* (Jin et al., 2023a).

To answer the second research question (What are the properties (statistics and ratio) of the common-sense knowledge relations in news headlines to different topics?), we find the intersection between the inferences belonging to different barriers and categories (see Figures 6–8). The results suggest that although inferences are being shared among the classes, there are some

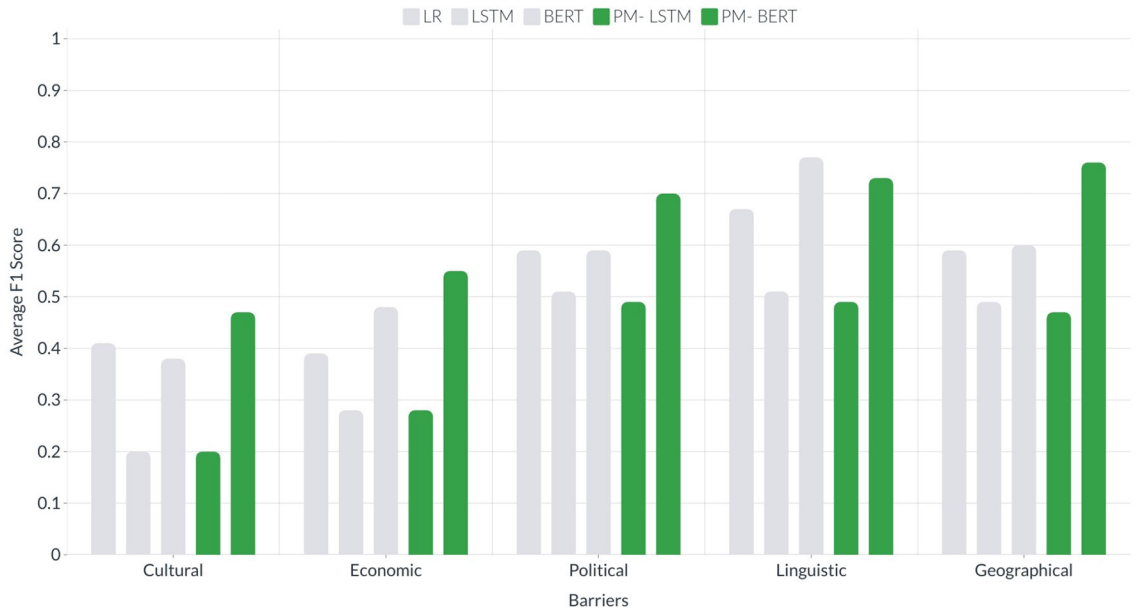


FIGURE 15
It presents bars of two colors for each barrier. The green bars show the average F1-scores of all the ten categories for LSTM and BERT using common-sense-based semantic knowledge and sentiments. The gray bars show the average F1-scores of all the ten categories for LR, LSTM, and BERT using only headline text. The x-axis shows the groups of bars for all five barriers, whereas the y-axis shows the average F1-score.

unique inferences for each class. Similarly, the same fact exists between the different categories. Therefore, it might be possible that it will help to improve the classification results. The results of the annotation show that there are variations in class distributions across different categories. Therefore, we use stratified sampling to maintain the class proportion in the train and test sets (see Figures 13, 14).

To answer our third research question (Which classification methods (classical or deep learning methods) yield the best performance to the barrier classification task?), We perform classification with classical machine learning methods, including Logistic Regression (LR), Naive Bayes (NB), Support Vector Classifier (SVC), k-nearest Neighbor (kNN), and Decision Tree (DT). Afterward, we perform classification with and without inferences using LSTM and BERT. We evaluate the models using the F1-score (see Section 5.4). We analyzed the classification results by comparing the ten categories (see Section 6.1) and three types of classification methods (see Section 6.2).

The results suggest that our proposed methods (LSTM and BERT with inferences based semantic knowledge and sentiments) offer better performance for the four barriers (cultural, economic, political and geographical).

8. Conclusion and perspective

In this paper, we focused on the classification of barriers to the spreading of news by utilizing semantic knowledge in the form of common-sense knowledge and sentiments. We consider the news related to the ten different categories (business, computers, games, health, home, recreation, science, shopping, society, and sports). After completing the automatic annotation of news data for the five barriers, including cultural, economic, political, linguistic, and geographical (binary class classification of the linguistic, political, and geographical barrier and ternary class classification of the cultural and political barrier), we perform classification with classical machine learning methods (LR, NB, SVC, kNN, and DT), deep learning (LSTM) and transformer-based methods (BERT). Our findings suggest that common-sense based semantic knowledge and sentiments help in achieving a higher F1-score. The classification of news across the barrier can help to recommend the news belonging to different categories and to identify the trends of different kinds of news across different barriers. The main theoretical contributions of this work are an approach to information barrier annotation based on news meta-data and labeling and classifying the news, including semantic knowledge across different barriers (cultural, economic, political, linguistic, and geographical). The annotation process includes meta-data extraction that requires too many requests to find the corresponding Wikipedia URLs for news publishers. Although many news publishers, including some local news publishers, do not have an entry in the Wikipedia database, popular and a significantly large number of news publishers do have their profiles available at the Wikipedia-Infobox. The annotation process and data statistics demonstrate that this approach to extracting profiles of news publishers is feasible to perform barrier classification to news spreading as well as for other important tasks such as understanding fake news propagation. The labeling process involves the demographic values and profile of news publishers, such as cultural and economic differences, political alignment, and publishing language. To the best of our knowledge, our proposed approach is the first of its kind to the classification of

Input a set of news headline H
Output predicted label: information-crossing or information-not-crossing or unsure

```
Find the sentiment S of a news headline (see Section 5.1)
for A headline in H do
    S_A = VADER(A)
end for
Extract inferences-based semantic knowledge K in form of tuples (see Section 5.2)
for A headline in H do
    K_A = COMET(A)
    for each relation r in K_A do
        if r is not in (react, need, intend, want, isFilledBy, and react) then
            ignore r
        end if
    end for
end for
Convert the K tuples into sentences(see Section 5.2).
Calculate the feature vectors of H, K, and S and merge them together (see Section 5.3)
Hyper-parameter settings and training of the models
if classes == binary then
    activation = sigmoid
    optimizar = adam
    loss = binary cross entropy
else if classes == ternary then
    activation = softmax
    optimizar = adam
    loss = categorical cross entropy
end if
Evaluation on test data using F1-score (see Section 5.4)
```

Algorithm 1. Barrier classification algorithms—PM-LSTM and PM-BERT.

barriers to the spreading of news. There are basically two practical contributions: (1) an annotated data set, and (2) an approach to the classification of barriers to the spreading of news based on semantic knowledge, including a wide range of common sense knowledge and sentiments of news headlines. Since the existing work lacks an annotated dataset for this task, it presents an annotated data set for the classification of barriers to the spreading of news. It presents the sentiment analysis of annotated news headlines as well as the properties of common-sense knowledge relations in news headlines. Our experimental evaluation shows that deep learning (LSTM) and transformer-based methods (BERT) can be useful for classifying barriers using common-sense-based knowledge and sentiments.

In the future, we plan to analyze the performance of prompt learning and GPT-based generative classification models for barrier classification to the spreading of news. Moreover, currently, geographical barrier is calculated in a binary way. In the future, we would like to extend the classes based on the distance between countries and time zone. Similarly, for the political and linguistic barriers, we will incorporate more information while annotating the news.

Data availability statement

The datasets generated for this study can be found in the GitHub repository via the following link: https://github.com/abdulsittar/BC-Inferences-Sentiments.git.

Author contributions

AS: methodology, data curation, writing—original draft preparation, software, and writing—reviewing and editing. DM: supervision, validation, and writing—reviewing and editing. MG: conceptualization and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

The research described in this paper was supported by the Slovenian Research Agency under the project J2-1736 Causality, by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 812997 (Cleopatra), and by the EU's Horizon Europe Framework under grant agreement number 101095095.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Ajao, O., Bhowmik, D., and Zargari, S. (2019). Sentiment aware fake news detection on online social networks," in *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE), 2507–2511. doi: 10.1109/ICASSP.2019.8683170

Alonso, M. A., Vilares, D., Gómez-Rodríguez, C., and Vilares, J. (2021). Sentiment analysis for fake news detection. *Electronics* 10, 1348. doi: 10.3390/electronics10111348

Al-Samarraie, H., Eldenfria, A., and Dawoud, H. (2017). The impact of personality traits on users' information-seeking behavior. *Inf. Proc. Manage*. 53, 237–247. doi: 10.1016/j.ipm.2016.08.004

Aslam, F., Awan, T. M., Syed, J. H., Kashif, A., and Parveen, M. (2020). Sentiments and emotions evoked by news headlines of coronavirus disease (covid-19) outbreak. *Human. Soc. Sci. Commun*. 7, 1–9. doi: 10.1057/s41599-020-0523-3

Barbaglia, L., Consoli, S., and Manzan, S. (2022). Forecasting with economic news. *J. Bus. Econ. Stat*. 41, 708–719. doi: 10.1080/07350015.2022.2060988

Bhagavatula, C., Bras, R. L., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., et al. (2019). Abductive commonsense reasoning. *arXiv preprint arXiv*:1908.05739.

Bhutani, B., Rastogi, N., Sehgal, P., and Purwar, A. (2019). "Fake news detection using sentiment analysis," in *2019 Twelfth International Conference on Contemporary Computing (IC3)* (IEEE), 1–5. doi: 10.1109/IC3.2019.8844880

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y. (2019). Comet: Commonsense transformers for automatic knowledge graph construction. arXiv preprint arXiv:1906.05317.

Bustos, S. M., Andersen, J. V., Miniconi, M., Nowak, A., Roszczynska-Kurasinska, M., and Brée, D. (2011). Pricing stocks with yardsticks and sentiments. *Algor. Finance* 1, 183–190. doi: 10.3233/AF-2011-013

Cambria, E., Olsher, D., and Rajagopal, D. (2014). "Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*. doi: 10.1609/aaai.v28i1.8928

Cambria, E., Song, Y., Wang, H., and Hussain, A. (2011). "Isanette: A common and common sense knowledge base for opinion mining," in *2011 IEEE 11th International Conference on Data Mining Workshops (IEEE)*, 315–322. doi: 10.1109/ICDMW.2011.106

Colas, F., and Brazdil, P. (2006). "Comparison of svm and some older classification algorithms in text classification tasks," in *IFIP International Conference on Artificial Intelligence in Theory and Practice* (Springer), 169–178. doi: 10.1007/978-0-387-34747-9_18

Colon-Hernandez, P., Havasi, C., Alonso, J., Huggins, M., and Breazeal, C. (2021). Combining pre-trained language models and structured knowledge. arXiv preprint arXiv:2101.12294.

Consoli, S., Barbaglia, L., and Manzan, S. (2022). Fine-grained, aspect-based sentiment analysis on economic and financial lexicon. *Knowl. Based Syst*. 247, 108781. doi: 10.1016/j.knosys.2022.108781

Cui, L., Wang, S., and Lee, D. (2019). "Same: sentiment-aware multi-modal embedding for detecting fake news," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 41–48. doi: 10.1145/3341161.3342894

Davis, E., Morgenstern, L., and Ortiz, C. L. (2017). The first winograd schema challenge at ijcai-16. *AI Magaz*. 38, 97–98. doi: 10.1609/aimag.v38i4.2734

Demirsoz, O., and Ozcan, R. (2017). Classification of news-related tweets. *J. Inf. Sci*. 43, 509–524. doi: 10.1177/0165551516653082

d'Haenens, L., Antoine, F., and Saeys, F. (2009). Belgium: Two communities with diverging views on how to manage media diversity. *Int. Commun. Gazette* 71, 51–66. doi: 10.1177/1748048508097930

Dogra, V., Singh, A., Verma, S., Jhanjhi, N., Talib, M., et al. (2021). Analyzing distilbert for sentiment classification of banking financial news," in *Intelligent Computing and Innovation on Data Science* (Springer), 501–510. doi: 10.1007/978-981-16-3153-5_53

ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., et al. (2021). Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv*:2109.05322. doi: 10.18653/v1/2021.emnlp-main.29

Fang, T., Zhang, H., Wang, W., Song, Y., and He, B. (2021). "Discos: Bridging the gap between discourse knowledge and commonsense knowledge," in *Proceedings of the Web Conference 2021* 2648–2659. doi: 10.1145/3442381.3450117

Fico, F. G., Lacy, S., and Riffe, D. (2008). A content analysis guide for media economics scholars. *J. Media Econ*. 21, 114–130. doi: 10.1080/08997760802069994

Gabielkov, M., Ramachandran, A., Chaintreau, A., and Legout, A. (2016). "Social clicks: What and who gets read on twitter?" in *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science* 179–192. doi: 10.1145/2896377.2901462

Gao, Q., Doering, M., Yang, S., and Chai, J. (2016). "Physical causality of action verbs in grounded language understanding," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* 1814–1824. doi: 10.18653/v1/P16-1171

Ghosal, D., Majumder, N., Gelbukh, A., Mihalcea, R., and Poria, S. (2020). Cosmic: Commonsense knowledge for emotion identification in conversations. arXiv preprint arXiv:2010.02795. doi: 10.18653/v1/2020.findings-emnlp.224

Godbole, N., Srinivasaiah, M., and Skiena, S. (2007). Large-scale sentiment analysis for news and blogs. *ICWSM* 7, 219–222.

González-Carvajal, S., and Garrido-Merchán, E. C. (2020). Comparing bert against traditional machine learning text classification. arXiv preprint arXiv:2005.13012.

Gravanis, G., Vakali, A., Diamantaras, K., and Karadais, P. (2019). Behind the cues: A benchmarking study for fake news detection. *Expert Syst. Applic*. 128, 201–213. doi: 10.1016/j.eswa.2019.03.036

Gulla, J. A., Zhang, L., Liu, P., Özgöbek, Ö., and Su, X. (2017). "The adressa dataset for news recommendation," in *Proceedings of the International Conference on Web Intelligence* 1042–1048. doi: 10.1145/3106426.3109436

Heydari, A., ali Tavakoli, M., Salim, N., and Heydari, Z. (2015). Detection of review spam: A survey. *Exp. Syst. Applic*. 42, 3634–3642. doi: 10.1016/j.eswa.2014.12.029

Hui, J. L. O., Hoon, G. K., and Zainon, W. M. N. W. (2017). Effects of word class and text position in sentiment-based news classification. *Procedia Comput. Sci*. 124, 77–85. doi: 10.1016/j.procs.2017.12.132

Hwang, J. D., Bhagavatula, C., Le Bras, R., Da, J., Sakaguchi, K., Bosselut, A., et al. (2021). "(comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs," in *Proceedings of the AAAI Conference on Artificial Intelligence* 6384–6392. doi: 10.1609/aaai.v35i7.16792

Ismayilzada, M., and Bosselut, A. (2022). kogito: A commonsense knowledge inference toolkit. arXiv preprint arXiv:2211.08451.

Jiang, S., and Tang, B. (2020). "Relying on multi-modal contextual cross-cultural communication ability training big data analysis," in *2020 13th International Conference on Intelligent Computation Technology and Automation (ICICTA)* 602–605. (IEEE), doi: 10.1109/ICICTA51737.2020.00133

Jin, D., Jin, Z., Zhou, J. T., and Szolovits, P. (2020). "Is bert really robust? A strong baseline for natural language attack on text classification and entailment," in *Proceedings of the AAAI Conference on Artificial Intelligence* 8018–8025. doi: 10.1609/aaai.v34i05.6311

Jin, W., Zhao, B., and Liu, C. (2023a). "Fintech key-phrase: A new chinese financial high-tech dataset accelerating expression-level information retrieval," in *International Conference on Database Systems for Advanced Applications* (Springer), 425–440. doi: 10.1007/978-3-031-30675-4_31

Jin, W., Zhao, B., Yu, H., Tao, X., Yin, R., and Liu, G. (2023b). Improving embedded knowledge graph multi-hop question answering by introducing relational chain reasoning. *Data Mining Knowl. Disc*. 37, 255–288. doi: 10.1007/s10618-022-00891-8

Jin, W., Zhao, B., Zhang, L., Liu, C., and Yu, H. (2023c). Back to common sense: Oxford dictionary descriptive knowledge augmentation for aspect-based sentiment analysis. *Inf. Proc. Manage*. 60, 103260. doi: 10.1016/j.ipm.2022.103260

Kamath, C. N., Bukhari, S. S., and Dengel, A. (2018). Comparative study between traditional machine learning and deep learning approaches for text classification," in *Proceedings of the ACM Symposium on Document Engineering* 2018 1–11. doi: 10.1145/3209280.3209526

Kelly, M. P., Martin, N., Dillenburger, K., Kelly, A. N., and Miller, M. M. (2019). Spreading the news: History, successes, challenges and the ethics of effective dissemination. *Behav. Analy. Pract*. 12, 440–451. doi: 10.1007/s40617-018-0238-8

King, G., Schneer, B., and White, A. (2017). How the news media activate public expression and influence national agendas. *Science* 358, 776–780. doi: 10.1126/science.aao1100

Kowsari, K., Brown, D. E., Heidarysafa, M., Meimandi, K. J., Gerber, M. S., and Barnes, L. E. (2017). Hdltex: Hierarchical deep learning for text classification," in *2017 16th IEEE international conference on machine learning and applications (ICMLA)* (IEEE), 364–371. doi: 10.1109/ICMLA.2017.0-134

Kumbure, M. M., Lohrmann, C., Luukka, P., and Porras, J. (2022). Machine learning techniques and data for stock market forecasting: a literature review. *Expert Syst. Applic*. 197, 116659. doi: 10.1016/j.eswa.2022.116659

Lamidi, I. K., and Olisa, D. (2016). Newspaper framing of the apc change mantra in the 2015 nigerian presidential election: A study of the punch and guardian newspapers. *J. Commun. Media Res*. 8, 201–218.

Leban, G., Fortuna, B., Brank, J., and Grobelnik, M. (2014). "Event registry: learning about world events from news," in *Proceedings of the 23rd International Conference on World Wide Web* 107–110. doi: 10.1145/2567948.2577024

Lei, Z., Haq, A. U., Zeb, A., Suzauddola, M., and Zhang, D. (2021). Is the suggested food your desired?: Multi-modal recipe recommendation with demand-based knowledge graph. *Exp. Syst. Appl*. 186, 115708. doi: 10.1016/j.eswa.2021.115708

Li, J., Bu, H., and Wu, J. (2017). "Sentiment-aware stock market prediction: A deep learning method," in *2017 International Conference on Service Systems and Service Management* (IEEE), 1–6.

Luan, Y., and Lin, S. (2019). "Research on text classification based on cnn and lstm," in *2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)* (IEEE), 352–355. doi: 10.1109/ICAICA.2019.8873454

Ma, M., Fang, P., Gao, J., and Song, C. (2017). "Does ideology affect the tone of international news coverage?" in *2017 International Conference on Behavioral, Economic, Socio-Cultural Computing (BESC)* (IEEE), 1–5. doi: 10.1109/BESC.2017.8256368

Martín, A. G., Fernández-Isabel, A., González-Fernández, C., Lancho, C., Cuesta, M., and de Diego, I. M. (2021). Suspicious news detection through semantic and sentiment measures. *Eng. Appl. Artif. Intell.* 101, 104230. doi: 10.1016/j.engappai.2021.104230

Mehler, A., Bao, Y., Li, X., Wang, Y., and Skiena, S. (2006). Spatial analysis of news sources. *IEEE Trans. Visual. Comput. Graph*. 12, 765–772. doi: 10.1109/TVCG.2006.179

Moreo, A., Romero, M., Castro, J., and Zurita, J. M. (2012). Lexicon-based comments-oriented news sentiment analyzer system. *Exp. Syst. Appl.* 39, 9166–9180. doi: 10.1016/j.eswa.2012.02.057

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., and Ngo, D. C. L. (2015). Text mining of news-headlines for forex market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment. *Exp. Syst. Appl.* 42, 306–324. doi: 10.1016/j.eswa.2014.08.004

Ng, R., and Tan, Y. W. (2021). Diversity of covid-19 news media coverage across 17 countries: The influence of cultural values, government stringency and pandemic severity. *Int. J. Environ. Res. Public Health* 18, 11768. doi: 10.3390/ijerph182211768

Obijiofor, L. (2010). Press coverage of hiv/aids in nigeria and the socio-cultural barriers that inhibit media coverage. *China Media Report Overseas* 6, 24–32.

Razniewski, S., Tandon, N., and Varde, A. S. (2021). "Information to wisdom: Commonsense knowledge extraction and compilation," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* 1143–1146. doi: 10.1145/3437963.3441664

Reese, S. D. (2007). The framing project: A bridging model for media research revisited. *J. Commun*. 57, 148–154. doi: 10.1111/j.1460-2466.2006.00334.x

Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., et al. (2016). Building event-centric knowledge graphs from news. *J. Web Semant*. 37, 132–151. doi: 10.1016/j.websem.2015.12.004

Segev, E. (2015). Visible and invisible countries: News flow theory revised. *Journalism* 16, 412–428. doi: 10.1177/1464884914521579

Shah, D., Isah, H., and Zulkernine, F. (2018). "Predicting the effects of news sentiments on the stock market," in *2018 IEEE International Conference on Big Data (Big Data)* (IEEE), 4705–4708. doi: 10.1109/BigData.2018.s8621884

Shah, K., Patel, H., Sanghvi, D., and Shah, M. (2020). A comparative analysis of logistic regression, random forest and knn models for the text classification. *Augmented Hum. Res.* 5, 1–16. doi: 10.1007/s41133-020-00032-0

Sheshadri, K., Shivade, C., and Singh, M. P. (2021). Detecting framing changes in topical news. *IEEE Trans. Comput. Soc. Syst*. 8, 780–791. doi: 10.1109/TCSS.2021.3063108

Shrawankar, U., and Wankhede, K. (2016). "Construction of news headline from detailed news article," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (IEEE), 2321–2325.

Sittar, A., Major, D., Mello, C., Mladenić, D., and Grobelnik, M. (2022a). Political and economic patterns in covid-19 news: From lockdown to vaccination. *IEEE Access* 10, 40036–40050. doi: 10.1109/ACCESS.2022.3164692

Sittar, A., Mladenić, D., and Grobelnik, M. (2022b). Analysis of information cascading and propagation barriers across distinctive news events. *J. Intell. Inf. Syst*. 58, 119–152. doi: 10.1007/s10844-021-00654-9

Storks, S., Gao, Q., and Chai, J. Y. (2019). Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. arXiv preprint arXiv:1904.01172.

Swati, S., and Grobelnik, M. (2022). Ic-bait: An inferential commonsense-driven model for predicting political polarity in news headlines. Available at SSRN 4114271. doi: 10.2139/ssrn.4114271

Taj, S., Shaikh, B. B., and Meghji, A. F. (2019). "Sentiment analysis of news articles: a lexicon based approach," in *2019 2nd international conference on computing, mathematics and engineering technologies (iCoMET)* (IEEE), 1–5. doi: 10.1109/ICOMET.2019.8673428

Vuorinen, E. (1994). "Crossing cultural barriers in international news transmission: A translational approach," in *Translation and the (re) Location of Meaning, Selected papers of the CETRA Research Seminars in Translation Studies* 161–171.

Walter, D., and Ophir, Y. (2019). News frame analysis: An inductive mixed-method computational approach. *Commun. Methods Measur*. 13, 248–266. doi: 10.1080/19312458.2019.1639145

Wang, Y., Zhou, Z., Jin, S., Liu, D., and Lu, M. (2017). "Comparisons and selections of features and classifiers for short text classification," in *IOP Conference Series: Materials Science and Engineering* (IOP Publishing), 012018. doi: 10.1088/1757-899X/261/1/012018

Wright, K. B. (2022). "Social media misinformation about extreme weather events and climate change: Structures, communication processes, and individual factors that influence the diffusion of misinformation," in *Communication and Catastrophic Events: Strategic Risk and Crisis Management*, 137. doi: 10.1002/9781119751847.ch9

Wu, H. D. (2007). A brave new world for international news? Exploring the determinants of the coverage of foreign news on us websites. *Int. Commun. Gazette* 69, 539–551. doi: 10.1177/1748048507082841

Yazdani, S. F., Murad, M. A. A., Sharef, N. M., Singh, Y. P., and Latiff, A. R. A. (2017). Sentiment classification of financial news using statistical features. *Int. J. Pattern Recogn. Artif. Intell.* 31, 1750006. doi: 10.1142/S0218001417500069

Yu, S., Liu, D., Zhu, W., Zhang, Y., and Zhao, S. (2020). Attention-based lstm, gru and cnn for short text classification. *J. Intell. Fuzzy Syst*. 39, 333–340. doi: 10.3233/JIFS-191171

Yu, S., Su, J., and Luo, D. (2019). Improving bert-based text classification with auxiliary sentence and domain knowledge. *IEEE Access* 7, 176600–176612. doi: 10.1109/ACCESS.2019.2953990

Zhai, Z., Xu, H., Kang, B., and Jia, P. (2011). Exploiting effective features for chinese sentiment classification. *Exp. Syst. Applic.* 38, 9139–9146. doi: 10.1016/j.eswa.2011.01.047

Zhou, Y., Mishra, S., Verma, M., Bhamidipati, N., and Wang, W. (2020). "Recommending themes for ad creative design via visual-linguistic representations," in *Proceedings of The Web Conference* 2521–2527. doi: 10.1145/3366423.3380001

Classification results via fine tuning of GPT-3 models This section presents additional experiments to complement the results of the above paper *Profiling the barriers to the spreading of news using news headlines*. In the paper, we utilized classical machine learning methods (LR, SVM, DT, RF, kNN), deep learning (LSTM), and transformers (BERT) for the barrier classification task. Since the OpenAI's GPT-based generative classification models showed excellent performance for different classification methods, we carried out additional experiments for our task using fine-tuned model of GPT-3 [70]. The training of OpenAI's GPT (generative pre-trained transformer) models has been performed on a large amount of text from the open internet. When given a prompt with just a few examples, it can often grasp what task you are trying to perform and generate a plausible completion. This is often referred to as a few-shot learning. Another way is fine tuning the base models of GPT-3 (ada, babbage, curie, and davinci). We prepared the training data according to the instructions provided by OpenAI [1] and created a fine-tuned model with a base-model ada.

The table 4.1 shows the comparison of LR, LSTM, BERT, PM-LSTM, PM-BERT and fine-tuned model across all the ten news categories based on the evaluation metric F1-score. Since the explanation of F1-score of LR, LSTM, BERT, PM-LSTM, and PM-BERT is available in the paper *Profiling the barriers to the spreading of news using news headlines*, we compare the F1-score of fine-tuned model with highest F1-score produced with those models. According to the table, we see that the results of the fine-tuned model for the cultural and economic barrier are quite similar to the other standard models. However, in case of binary classification, for the linguistic, geographical, and political barriers, the F1-score of fine-tuned model is higher for 8, 9, and 8 categories out of 10, respectively. The fine-tuned model achieved lower F1-score on recreation, and computers categories of the linguistic barrier, health category of the geographical barrier, and science and shopping categories of the political barrier.

[1] `https://platform.openai.com/docs/guides/fine-tuning`

Table 4.1: **F1-score of the five different machine learning algorithms (LR, LSTM, BERT, PM-LSTM, and PM-BERT) and fine-tuned model for the ten different categories (business, computers, games, health, home, recreation, science, shopping, society, and sports).**

Model	Category	Cul	Eco	Pol	Ling	Geo	Category	Cul	Eco	Pol	Ling	Geo
LR	Business	0.40	0.48	0.71	0.73	0.61	Recreation	0.37	0.30	0.59	0.73	0.57
	Computers	0.42	0.35	0.58	0.63	0.56	Science	0.42	0.42	0.62	0.71	0.65
	Games	0.52	0.35	0.59	0.59	0.60	Shopping	0.36	0.27	0.49	0.61	0.52
	Health	0.36	0.40	0.58	0.67	0.64	Society	0.40	0.45	0.62	0.68	0.62
	Home	0.39	0.57	0.49	0.68	0.59	Sports	0.44	0.28	0.59	0.62	0.57
LSTM	Business	0.19	0.28	0.49	0.55	0.47	Recreation	0.20	0.39	0.48	0.41	0.47
	Computers	0.20	0.08	0.49	0.52	0.46	Science	0.20	0.20	0.48	0.43	0.43
	Games	0.14	0.29	0.74	0.70	0.48	Shopping	0.23	0.44	0.49	0.44	0.49
	Health	0.18	0.17	0.49	0.59	0.53	Society	0.20	0.27	0.49	0.46	0.48
	Home	0.19	0.21	0.49	0.59	0.63	Sports	0.26	0.43	0.48	0.43	0.49
BERT	Business	0.42	0.54	0.62	0.74	0.50	Recreation	0.32	0.29	0.74	0.81	0.59
	Computers	0.38	0.30	0.47	0.75	0.66	Science	0.39	0.47	0.68	0.78	0.60
	Games	0.40	0.65	0.77	0.85	0.68	Shopping	0.34	0.44	0.49	0.63	0.50
	Health	0.38	0.54	0.66	0.79	0.67	Society	0.39	0.51	0.49	0.73	0.48
	Home	0.32	0.60	0.54	0.81	0.67	Sports	0.48	0.50	0.48	0.76	0.66
PM-LSTM	Business	0.19	0.28	0.49	0.43	0.48	Recreation	0.21	0.47	0.48	0.44	0.46
	Computers	0.21	0.06	0.49	0.46	0.47	Science	0.19	0.21	0.48	0.74	0.43
	Games	0.18	0.22	0.48	0.40	0.48	Shopping	0.18	0.50	0.49	0.50	0.48
	Health	0.20	0.27	0.48	0.43	0.46	Society	0.19	0.25	0.49	0.64	0.48
	Home	0.20	0.27	0.49	0.45	0.48	Sports	0.27	0.22	0.48	0.44	0.48
PM-BERT	Business	0.48	0.66	**0.97**	0.80	0.96	Recreation	**0.66**	0.35	0.74	0.48	0.65
	Computers	0.46	0.28	0.54	**0.97**	0.70	Science	0.47	0.53	**0.97**	0.78	0.66
	Games	0.33	**0.72**	0.48	0.47	0.72	Shopping	0.45	0.50	**0.97**	0.67	0.56
	Health	0.46	0.60	0.72	0.82	**0.97**	Society	0.52	0.69	0.55	0.73	**0.97**
	Home	0.41	0.66	0.54	0.83	0.73	Sports	0.49	0.50	0.54	0.79	0.63
Fine-tuned Model	Business	0.47	**0.78**	**0.97**	**0.90**	**0.98**	Recreation	0.48	**0.67**	**0.93**	0.80	**0.95**
	Computers	0.42	**0.72**	**0.93**	0.96	**0.95**	Science	**0.52**	0.51	0.85	**0.90**	**0.97**
	Games	**0.55**	0.42	**0.86**	**0.87**	**0.91**	Shopping	0.43	0.34	0.9	**0.87**	**0.91**
	Health	**0.48**	**0.69**	**0.88**	**0.86**	0.95	Society	0.49	**0.72**	**0.96**	**0.78**	**0.98**
	Home	**0.46**	0.51	**0.93**	**0.88**	**0.95**	Sports	**0.64**	0.49	**0.90**	**0.80**	**0.95**

Evaluation using statistical testing This section introduces supplementary experiments aimed at complementing the findings presented in the aforementioned paper *Profiling the barriers to the spreading of news using news headlines*. Within the paper, we conducted a comparative analysis of performance, measured by F1-score, across various methods, including classical machine learning techniques (LR, SVM, DT, RF, kNN), deep learning (LSTM), and transformers (BERT), specifically for the task of classifying barriers. We utilized the average raw F1-score across all barrier categories to determine the efficacy of each method. Now, we perform statistical test that is used in null-hypothesis testing. It assume a null hypothesis of no relationship or no difference between groups.

Statistical tests calculate a score that describes how much the relationship between variables in our test differs from the null hypothesis of no relationship [71]. It then calculates a p-value (probability value) [72]. The p-value estimates how likely it is that we would see the difference described by the test statistic if the null hypothesis of no relationship were true. If the value of the test statistic is more extreme than the statistic calculated from the null hypothesis, then we can infer a statistically significant relationship between the predictor and outcome variables [73]. If the value of the test statistic is less extreme than the one calculated from the null hypothesis [74], then we can infer no statistically significant relationship between the predictor and outcome variables. We perform statistical testing to more robustly understand the comparative effectiveness of the both models (namely kNN and Logistic Regression) used. For comparing the performance of two models, we perform Wilcoxon Rank-Sum test. It is used to solve problem of comparing two methods. After dividing the dataset into five folds, we run both models five time, using a different fold for testing each time. We repeat the procedure with kNN and LR, to form pairs of values $(F1_{kNN}, F1_{LR})$, where $F1_{kNN}$ and $F1_{LR}$ represent the classification metric F1 generated using kNN and LR respectively. We define the null and alternate hypothesis for this test below:

Null hypothesis: There is no statistical difference between the two distributions of data observed with kNN and LR.

Alternate hypothesis: There is a statistical difference between the two distributions of data observed with kNN and LR.

On conducting the null hypothesis test using a standard threshold for significance of $p < 0.05$ (significance level, $= 0.05$), it is evident that kNN and LR have statistically significant difference with respect to the F1 score. Initially, employing 5-fold cross-validation while training kNN and Logistic Regression across all barrier categories, we utilized the F1 score for each of the 5 folds. Subsequently, we employed the Wilcoxon Rank-Sum test to evaluate the results, as presented in Table 4.2. The first sub-table showcases the mean F1-score obtained through LR, while the second demonstrates the mean F1-score acquired via kNN. The bold values in the table indicate categories where statistically significant differences exist between the data distributions, indicating tangible performance differences between LR and kNN than statistical chance.

In the first step, we performed 5-fold cross validation while training kNN and Logistic Regression for all the categories of a barrier. Using the F1 score for all the 5 folds, we performed Wilcoxon Rank-Sum test and presented its results in Table 4.2. The first sub-table presents the average F1-score obtained using LR and the second table presents the average F1-score obtained using kNN. The bold values in the table represents the categories where there are statistically significant differences between the distributions of data. And

Table 4.2: **Results of statistical (Wilcoxon Rank-Sum) test with a standard threshold for significance of $p < 0.05$ (significance level, $= 0.05$) to understand the comparative effectiveness of the both models - kNN and Logistic Regression. The bold values indicate the categories where the F1-score as well as the statistical differences are high in all the barriers.**

F1-score (LR)

Category	Cultural	Economic	Political	Linguistic	Geographical
Business	**0.40**	**0.48**	**0.71**	**0.73**	**0.61**
Computers	**0.42**	**0.35**	0.58	**0.63**	**0.56**
Games	**0.52**	**0.35**	0.59	0.59	**0.60**
Health	0.36	0.40	0.58	**0.67**	**0.64**
Home	0.39	**0.57**	0.49	**0.68**	**0.59**
Recreation	**0.37**	**0.30**	**0.59**	**0.73**	**0.57**
Science	**0.42**	0.42	**0.62**	**0.71**	**0.65**
Shopping	0.36	0.27	**0.49**	**0.61**	**0.52**
Society	0.40	0.45	0.62	**0.68**	**0.62**
Sports	0.4	0.28	0.59	**0.62**	**0.57**

F1-score (kNN)

Category	Cultural	Economic	Political	Linguistic	Geographical
Business	0.30	0.27	0.49	0.52	0.49
Computers	0.35	0.29	0.47	0.46	0.49
Games	0.31	0.29	0.48	0.72	0.48
Health	0.28	0.31	**0.59**	0.51	0.48
Home	0.30	0.25	**0.51**	0.58	0.49
Recreation	0.32	0.29	0.50	0.46	0.49
Science	0.33	0.2	0.55	0.58	0.50
Shopping	0.29	0.27	0.48	0.44	0.52
Society	0.36	0.27	0.53	0.55	0.49
Sports	0.4	0.28	0.47	0.44	0.48

it means that there is real performance difference between LR and kNN than statistical chance. These outcomes from the statistical test are consistent with the observations detailed in the aforementioned paper. Specifically, the data distributions differ significantly across all geographical and linguistic barrier categories, with LR exhibiting a higher average F1-score compared to kNN. However, within the linguistic barrier's games category, kNN shows a superior average F1-score to LR. In the cultural, economic, and political barrier categories, there are five, four, and six respective categories displaying statistical differences between LR and kNN's data distributions. Nonetheless, among a total of fifty categories across all barriers, seventeen reveal no statistical differences in data distribution, with six out of fifteen instances where kNN attains an equal or higher average F1-score than LR.

Chapter 5

Conclusions

5.1 Thesis Summary

This thesis investigates the influence of different barriers to information spreading in news related to different events. After giving a general introduction of information spreading through news, Section 1.2.1 explains the influence of different factors including journalists' professional routines, organization influences, gender and political bias, and effects of travelling to different socio-cultural environments for news reporting. Several cross-cultural studies investigate the effect of culture on news such as coverage of news relevant to the local audiences between those countries which are geographically or culturally close. Similarly, common national languages promote the more coverage of news. These studies also explain that cultural diffusion is performed for people using language to communicate with each other. Sections from 1.2.1.1 to 1.2.1.5 explain the effect of economic, political, geographical, cultural, and language barriers to news spreading. An overview of news framing and agenda-setting is explained in Section 1.2.2.

To perform the experiments related to news spreading barriers, one of the main problems was limited availability of news datasets containing barrier-related information. The data collection process is explained in Section 2.1. This data is collected focusing on the idea that news about different domains spreads differently. Therefore, after filtering top contrasting events, a dataset is built consisting of FIFA World Cup, earthquakes, global warming of three contrasting domains (sports, natural disasters, and climate changes, respectively). Different research questions are constructed to understand the influence of six barriers (cultural, economic, linguistic, geographic, time zone, and political). These research questions support the hypothesis (depending on the nature of an event, there will be variations in information spreading behaviour across the observed barriers) from different perspectives. Since the news representation in a form of chains can help to understand the coverage and spreading range, cosine similarity among the news articles is computed to compare the temporal chains of news spreading across three contrasting domains. Moreover, other visualizations have been used to look inside the distributions across different economies, cultures, political alignments. The results show that 1) the scope of a specific event significantly affects the news spreading across languages, 2) geographical size of a news publisher's country is directly proportional to the number of publishers and articles reporting on the same information, 3) countries with shorter time-zone differences and similar cultures tend to propagate news between each other, 4) news related to global warming comes across economic barriers more smoothly than news related to FIFA World Cup and earthquakes, and 5) events which may in some way involve political benefits are mostly published by those publishers which are not politically neutral.

Apart from analysis of influence of different barriers on news spreading related to three

contrasting domains, the news reporting differences across different barriers is another area which can help to understand the role of content in understanding the news spreading barriers. In this context, news classification belonging to different categories including business, health, recreation, science etc. has been performed across different cultures. Prior to this, clusters of countries based on similar cultures have been created (see Section 3.1) to support this classification.

Global events become famous and catch the attentions in all corners of the world. These events have some direct relationships to characteristics of the newspapers in terms of political alignment and economic conditions. Firstly, a study has been conducted to find this relationship between world events and characteristics of the newspapers. This study performs clustering among top read newspapers in a world based on a set of concepts and content categories. For a detail overview see Section 3.1.1. The results show that 1) the representation of the news events with the Wikipedia concepts and DMOZ categories appears, an appropriate way to understand relationships between the newspapers, 2) economic conditions of the country of the newspaper publisher reflect better in Wikipedia concepts than when using representation with DMOZ categories, whereas for identifying politically aligned groups of newspapers, DMOZ categories stand out more suitable, 3) for capturing economic groups, clustering using the Dynamic Time Warping similarity between the trend lines of newspapers is better aligned with the ground truth groups than other tested similarities, whereas for capturing a political group, Jaccard distance using the frequent terms and Euclidean distance between the trend lines turns more useful. On the other hand, an approach has been developed to understand the evaluation of discussions using topic modeling. This approach of using LDA with a combination of 1-6 grams and article's pooling based on queries is evaluated based on a coherence score. Here, the focus was on the same barriers (economic and political) to understand their effects on news reporting (see Section 3.2). The results show that the news reporting by newspapers with different political alignment supports the reported content. Also, economic issues reported by newspapers depend on economy of the place where a newspaper resides. Similarly, a detailed list of features utilized to cluster the news reporting. The experiments include the comparison of typical bag-of-words features with stylistic features. News articles related to BREXIT and published across three regions of the UK (London, Wales, and Scotland) are considered for such experiments. The results show that stylistic features can be used for clustering the news article and their performance is much better than bag-of-words.

After understanding the influence of the barrier news spreading and analyzing the role of content in news reporting differences, a mechanism of automatically annotating and classifying the news articles across different barriers is developed. Approximately one million news articles related to ten different kinds of events are selected (health, sports, science, recreation, games, homes, society, shopping, computers, and business). This data is then labeled following automatic mechanism and a set of already defined guidelines. To perform the classification, semantic annotation such as Wikipedia concepts and news headlines along with common sense inferences and sentiments have been used. The classification of news articles has been performed using classical machine learning methods, deep learning and transformer-based methods (see Section 4.1).

In conclusion, this thesis investigates the influence of different barriers information spreading in news. The concept of cross-lingual event-centric open analytics recently necessitates the need of developing analytical methods where the main concerns are about analyzing the cross-lingual, multi-cultural, and event-centric data. We have proposed several novel methodologies, providing their theoretical justification, and evaluated their performance in different types of datasets including the FIFA World Cup, earthquakes, global warming, COVID-19, and Brexit.

5.2 Scientific Contributions

The main contributions of the thesis are:

1. A novel methodology to analyze the news spreading barriers on different kinds of news events.
A corpus has been collected and annotated containing news articles related to three different domains: natural disasters, sports, and climate change. Further, it has been extended with background information related to six barriers (linguistic, economic, time zone, geographic, political, and cultural). For mono-lingual news analysis, network analysis is performed, whereas for multi-lingual temporal information cascading analysis, visualization is developed to portray the temporal spread of a list of news articles. For other barriers, alluvial and chord diagrams, and Google maps have been used.

2. A novel approach to enhance the topic modeling technique and understand political and economic differences in news reporting.
An approach is proposed to enhance the topic modeling technique that uses LDA with a combination of 1-6 grams and article pooling based on queries to improve the quality of topics without modifying the structure of LDA. Using this approach, political and economic differences in news reporting across different geographical places have been identified. The proposed approach has been applied to news articles related to COVID-19. A complete phase of the COVID pandemic has been covered in collecting the data (from January 2020 to May 2021).

3. An approach to barrier profiling by automatically annotating and classifying the news articles for the different barriers.
The task of barrier profiling has been proposed along with a novel approach to automatic annotation of news articles and classification based on semantic annotation as well as news headlines using common sense inferences and sentiments. A corpus has been collected from the Event Registry that belongs to ten different categories (business, computers, games, health, home, recreation, science, shopping, society) and, published from 2016 to 2021. Metadata information is collected from Wikipedia-infox for more than ten thousand news publishers. To annotate the news articles, annotation guidelines have been set. For the binary and ternary class labels, experiments have been conducted comparing machine learning classical classification methods, deep learning methods, and transformer-based methods.

5.3 Future Work

This thesis has shown the current analytical methods to analyzing news spreading barriers. It has also shown that machine learning methods are appropriate to detect and profile these barriers. However, there are certainly many opportunities for further advancing the proposed approaches, as well as finding new practical applications.

The proposed methodology for analyzing the influence of the barriers on news spreading is utilizing simple mono- and cross-lingual similarity measures. And the methods such as graphs, charts, and animations are providing the prototypes. We intend to utilize the latest semantic similarity methods across cross-lingual news. Additionally, we intend to develop a single large framework that will be special for presenting how the cross-lingual news related to specific events spread over time.

The datasets of event-centric information propagation are about three events (FIFA World Cup, global warming, and earthquake) in different domains (natural disasters, sports, and climate changes) and consist of only a few attributes. We intend to add more events for these three domains and enhance them with more domains. Now the attributes that mention the meta-data are only about news publishers and news articles. Since the socio-economic contexts are more informative, therefore, we would like to enhance this dataset with economic and cultural differences alongside the news publishers' metadata.

The proposed methodology for profiling the news spreading barriers is based on a small set of features and classification models. The used representation was tf-idf score of word n-grams. Since there are many advanced features and representation methods available in natural language processing, therefore, we would like to extend the features with lexical and stylistic features along with already trained models on news data. Also, we intend to use transformer-based trained language models.

References

[1] A. O. SANNI and N. A. ISSA, "Challenges of newspaper circulation in nigeria,"

[2] G. Graham and A. Greenhill, "Exploring interaction: Print and online news media synergies," *Internet Research*, 2013.

[3] P. Ghasiya and K. Okamura, "Investigating covid-19 news across four nations: A topic modeling and sentiment analysis approach," *Ieee Access*, vol. 9, pp. 36 645–36 656, 2021.

[4] D. A. Atkins, "Sense of community, political participation, and civic engagement: An examination of the relationships between local daily newspapers, news websites, and their communities," Ph.D. dissertation, Virginia Tech, 2016.

[5] C. Flavián, M. Guinalıu, and R. Gurrea, "The influence of familiarity and usability on loyalty to online journalistic services: The role of user experience," *Journal of Retailing and Consumer Services*, vol. 13, no. 5, pp. 363–375, 2006.

[6] D. Peksen, T. M. Peterson, and A. C. Drury, "Media-driven humanitarianism? news media coverage of human rights abuses and the use of economic sanctions," *International Studies Quarterly*, vol. 58, no. 4, pp. 855–866, 2014.

[7] S. K. Mustafaa, O. S. Alib, M. S. Awlqadirb, and R. J. Mahmoodb, "Investigating factors affecting poor reading culture among efl university students," 2019.

[8] N. Couldry and J. Curran, *Contesting media power: Alternative media in a networked world*. Rowman & Littlefield Publishers, 2003.

[9] J. Jennings and M. Rubado, "Newspaper decline and the effect on local government coverage," *Annette Strauss Institute for Civic Life. Available online at https://moody. utexas. edu/sites/default/files/Strauss_ Research_ Newspaper_ Decline_ 2019–11-Jennings. pdf (accessed January 25, 2020)*, 2019.

[10] P. J. Shoemaker and T. P. Vos, "Media gatekeeping," in *An integrated approach to communication theory and research*, Routledge, 2014, pp. 89–103.

[11] T.-K. Chang and J.-W. Lee, "Factors affecting gatekeepers' selection of foreign news: A national survey of newspaper editors," *Journalism Quarterly*, vol. 69, no. 3, pp. 554–561, 1992.

[12] L. Camaj, "Media framing through stages of a political discourse: International news agencies' coverage of kosovo's status negotiations," *International Communication Gazette*, vol. 72, no. 7, pp. 635–653, 2010.

[13] F. Firdaniza, B. N. Ruchjana, D. Chaerani, and J. Radianti, "Information diffusion model in twitter: A systematic literature review," *Information*, vol. 13, no. 1, p. 13, 2021.

[14] A. Chakraborty, S. Ghosh, N. Ganguly, and K. P. Gummadi, "Optimizing the recency-relevancy trade-off in online news recommendations," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 837–846.

[15] G. Jiang, S. Li, and M. Li, "Dynamic rumor spreading of public opinion reversal on weibo based on a two-stage spnr model," *Physica A: Statistical Mechanics and its Applications*, vol. 558, p. 125 005, 2020.

[16] K. Oida, "Impact of network density on cascade size and community growth," *Applied Network Science*, vol. 4, no. 1, pp. 1–17, 2019.

[17] S. Kumar, M. Saini, M. Goel, and B. Panda, "Modeling information diffusion in online social networks using a modified forest-fire model," *Journal of intelligent information systems*, vol. 56, no. 2, pp. 355–377, 2021.

[18] J. Filla and M. Johnson, "Local news outlets and political participation," *Urban Affairs Review*, vol. 45, no. 5, pp. 679–692, 2010.

[19] N. Newman, D. A. Levy, and R. K. Nielsen, *Reuters Institute digital news report 2015: Tracking the future of news*. Reuters Institute for the Study of Journalism, 2015.

[20] L. Guo and C. J. Vargo, "Global intermedia agenda setting: A big data analysis of international news flow," *Journal of Communication*, vol. 67, no. 4, pp. 499–520, 2017.

[21] H. Seo, "International media coverage of north korea: Study of journalists and news reports on the six-party nuclear talks," *Asian Journal of Communication*, vol. 19, no. 1, pp. 1–17, 2009.

[22] A. Schwarz, "The theory of newsworthiness applied to mexico's press. how the news factors influence foreign news coverage in a transitional country," 2006.

[23] E. Segev, "Visible and invisible countries: News flow theory revised," *Journalism*, vol. 16, no. 3, pp. 412–428, 2015.

[24] T. M. Jones, P. Van Aelst, and R. Vliegenthart, "Foreign nation visibility in us news coverage: A longitudinal analysis (1950-2006)," *Communication Research*, vol. 40, no. 3, pp. 417–436, 2013.

[25] H. D. Wu, "Investigating the determinants of international news flow: A meta-analysis," *Gazette (Leiden, Netherlands)*, vol. 60, no. 6, pp. 493–512, 1998.

[26] C. Grasland, *International news flow theory revisited through a space–time interaction model: Application to a sample of 320,000 international news stories published through rss flows by 31 daily newspapers in 2015*, 2020.

[27] S. Liu, M. Boukes, and K. De Swert, "Strategy framing in the international arena: A cross-national comparative content analysis on the china-us trade conflict coverage," *Journalism*, p. 14 648 849 211 052 438, 2022.

[28] M. S. Al-Zaman and T. Khan, "Framing environmental news in bangladesh," *Media Asia*, vol. 49, no. 2, pp. 98–110, 2022.

[29] N. X. Liu, *News framing through English-Chinese translation: A comparative study of Chinese and English media discourse*. Routledge, 2018.

[30] M. R. Sobel, "Chronicling a crisis: Media framing of human trafficking in india, thailand, and the usa," *Asian Journal of Communication*, vol. 24, no. 4, pp. 315–332, 2014.

[31] E. Mitchelstein and P. J. Boczkowski, "Between tradition and change: A review of recent research on online news production," *Journalism*, vol. 10, no. 5, pp. 562–586, 2009.

[32] M. Malling, "The story behind the news: Informal and invisible interactions between journalists and their sources in two countries," Ph.D. dissertation, Mid Sweden University, 2022.

[33] G. Greenwald, "Israeli media coverage of international male and female politicians: Gender and ethnopolitical aspects," *Communications*, 2022.

[34] F. Hanusch, "The geography of travel journalism: Mapping the flow of travel stories about foreign countries," *International Communication Gazette*, vol. 76, no. 1, pp. 47–66, 2014.

[35] A. Rafeeq and S. Jiang, "From the big three to elite news sources: A shift in international news flow in three online newspapers thenational. ae, nst. com. my, and nzherald. co. nz," *The Journal of International Communication*, vol. 24, no. 1, pp. 96–114, 2018.

[36] D. Vogler and L. Udris, "Transregional news media coverage in multilingual countries: The impact of market size, source, and media type in switzerland," *Journalism Studies*, vol. 22, no. 13, pp. 1793–1813, 2021.

[37] J. Xue and W. Zuo, "English dominance and its influence on international communication," *Theory and Practice in Language Studies*, vol. 3, no. 12, p. 2262, 2013.

[38] M. He and J. Lee, "Social culture and innovation diffusion: A theoretically founded agent-based model," *Journal of Evolutionary Economics*, vol. 30, no. 4, pp. 1109–1149, 2020.

[39] K. Jiang, G. A. Barnett, and L. D. Taylor, "Dynamics of culture frames in international news coverage: A semantic network analysis," *International Journal of Communication*, vol. 10, p. 27, 2016.

[40] M. Khosrowjerdi, A. Sundqvist, and K. Byström, "Cultural patterns of information source use: A global study of 47 countries," *Journal of the Association for Information Science and Technology*, vol. 71, no. 6, pp. 711–724, 2020.

[41] P. Müller, "National identity building through patterns of an international third-person perception in news coverage," *International Communication Gazette*, vol. 75, no. 8, pp. 732–749, 2013.

[42] H. D. Wu, "A brave new world for international news? exploring the determinants of the coverage of foreign news on us websites," *International Communication Gazette*, vol. 69, no. 6, pp. 539–551, 2007.

[43] I. Rösel, "The impact of cross-cultural dynamics on the effectiveness of change management processes," in *Global Business Conference 2021 Proceedings*, 2021, p. 169.

[44] E. Vuorinen, "Crossing cultural barriers in international news transmission: A translational approach," in *Translation and the (re) location of meaning, selected papers of the CETRA Research Seminars in Translation Studies*, vol. 1996, 1994, pp. 161–171.

[45] I. E. Aguilar Rodrıguez, C. A. Bernal Torres, J. C. Aldana Bernal, A. Acosta Aguinaga, C. H. Artieda Cajilema, and P. Chalá, "Relationship between social culture, industry 4.0, and organizational performance in the context of emerging economies," *Journal of Industrial Engineering and Management*, vol. 14, no. 4, pp. 750–770, 2021.

[46] M. Quezada, V. Peña-Araya, and B. Poblete, "Location-aware model for news events in social media," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015, pp. 935–938.

[47] G. Sylvie and H. I. Chyi, "One product, two markets: How geography differentiates online newspaper audiences," *Journalism & mass communication quarterly*, vol. 84, no. 3, pp. 562–581, 2007.

[48] D. Laniado, Y. Volkovich, S. Scellato, C. Mascolo, and A. Kaltenbrunner, "The impact of geographic distance on online social interactions," *Information Systems Frontiers*, vol. 20, no. 6, pp. 1203–1218, 2018.

[49] J. Maldonado, V. Peña-Araya, and B. Poblete, "Spatio and temporal characterization of chilean news events in social media," in *SIGIR 2015 Workshop on Temporal, Social and Spatially-aware Information Access (TAIA'15)*, 2015.

[50] E. Humprecht and F. Esser, "Diversity in online news: On the importance of ownership types and media system types," *Journalism Studies*, vol. 19, no. 12, pp. 1825–1847, 2018.

[51] M. A. Ahmad, M. Ersoy, and T. H. Dambo, "Influence of political tweets on campaign coverage: Building the news agenda in twittersphere," *Journalism Practice*, vol. 16, no. 1, pp. 103–121, 2022.

[52] P. Maurer and M. Beiler, "Networking and political alignment as strategies to control the news: Interaction between journalists and politicians," *Journalism Studies*, vol. 19, no. 14, pp. 2024–2041, 2018.

[53] A. Ceron, "Internet, news, and political trust: The difference between social media and online media outlets," *Journal of computer-mediated communication*, vol. 20, no. 5, pp. 487–503, 2015.

[54] E. S. Van der Goot, T. G. Van der Meer, and R. Vliegenthart, "Reporting on political acquaintances: Personal interactions between political journalists and politicians as a determinant of media coverage," *International Journal of Communication*, vol. 15, p. 23, 2021.

[55] N. A. G. Fredheim, "Dancing in the dark: Source coordination and strategic media alliances in the health field," *Journalism Studies*, vol. 22, no. 1, pp. 96–113, 2021.

[56] A. Alambo, M. Gaur, and K. Thirunarayan, "Depressive, drug abusive, or informative: Knowledge-aware study of news exposure during covid-19 outbreak," *arXiv preprint arXiv:2007.15209*, 2020.

[57] H. Kwak and J. An, "A first look at global news coverage of disasters by using the gdelt dataset," in *International conference on social informatics*, Springer, 2014, pp. 300–308.

[58] A. Sittar, D. Mladenić, and T. Erjavec, "A dataset for information spreading over the news," in *Proceedings of the 23th International Multiconference Information Society SiKDD*, vol. 100, 2020, pp. 5–8.

[59] A. Sittar, D. Mladenić, and M. Grobelnik, "Analysis of information cascading and propagation barriers across distinctive news events," *Journal of Intelligent Information Systems*, vol. 58, no. 1, pp. 119–152, 2022.

[60] A. Sittar and D. Mladenic, "Classification of cross-cultural news events," in *Proceedings of the 24th International Multiconference Information Society SiKDD*, vol. 100, 2021, pp. 29–32.

[61] ——, "How are the economic conditions and political alignment of a newspaper reflected in the events they report on?" In *Central European Conference on Information and Intelligent Systems*, Faculty of Organization and Informatics Varazdin, 2021, pp. 201–208.

[62] A. Sittar, D. Major, C. Mello, D. Mladenić, and M. Grobelnik, "Political and economic patterns in covid-19 news: From lockdown to vaccination," *IEEE Access*, vol. 10, pp. 40 036–40 050, 2022.

[63] A. Cafruny, V. K. Fouskas, W. D. Mallinson, and A. Voynitsky, "Ukraine, multipolarity and the crisis of grand strategies," *Journal of Balkan and Near Eastern Studies*, vol. 25, no. 1, pp. 1–21, 2023.

[64] N. AlQershi, R. B. A. Saufi, N. A. Ismail, *et al.*, "The moderating role of market turbulence beyond the covid-19 pandemic and russia-ukraine crisis on the relationship between intellectual capital and business sustainability," *Technological Forecasting and Social Change*, vol. 186, p. 122 081, 2023.

[65] A. Sittar, D. Mladenić, and M. Grobelnik, "Political and economic patterns in covid-19 news: From lockdown to vaccination," *IEEE Access*, vol. 10, pp. 40 036–40 050, 2022.

[66] A. Sittar, J. Webber, and D. Mladenić, "Stylistic features in clustering news reporting: News articles on brexit," 2022.

[67] A. Aker, M. Paramita, E. Kurtic, *et al.*, "Automatic label generation for news comment clusters," in *Proceedings of the 9th International Natural Language Generation Conference*, Association for Computational Linguistics, 2016, pp. 61–69.

[68] M. P. Kelly, N. Martin, K. Dillenburger, A. N. Kelly, and M. M. Miller, "Spreading the news: History, successes, challenges and the ethics of effective dissemination," *Behavior Analysis in Practice*, vol. 12, no. 2, pp. 440–451, 2019.

[69] A. Sittar, D. Mladenić, and M. Grobelnik, "Profiling the barriers to the spreading of news using news headlines," *Frontiers in Artificial Intelligence*, vol. 6, p. 1 225 213,

[70] R. Zhang, Y.-S. Wang, and Y. Yang, "Generation-driven contrastive self-training for zero-shot text classification with instruction-tuned gpt," *arXiv preprint arXiv:2304.11872*, 2023.

[71] A. W. Edwards, "Ra fischer, statistical methods for research workers, (1925)," in *Landmark writings in western mathematics 1640-1940*, Elsevier, 2005, pp. 856–870.

[72] J. Neyman and E. S. Pearson, "Ix. on the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, no. 694-706, pp. 289–337, 1933.

[73] J. P. Romano and E. Lehmann, *Testing statistical hypotheses*, 2005.

[74] G. Casella and R. L. Berger, "Statistical inference duxbury press," *Pacific Grove, CA.[Google Scholar]*, 2002.

Bibliography

Publications Related to the Thesis

Journal Articles

A. Sittar, D. Mladenić, and M. Grobelnik, "Analysis of information cascading and propagation barriers across distinctive news events," *Journal of Intelligent Information Systems*, vol. 58, no. 1, pp. 119–152, 2022.

A. Sittar, D. Mladenić, and M. Grobelnik, "Political and economic patterns in covid-19 news: From lockdown to vaccination," *IEEE Access*, vol. 10, pp. 40 036–40 050, 2022.

A. Sittar, D. Mladenić, and M. Grobelnik, "Profiling the barriers to the spreading of news using news headlines," *Frontiers in Artificial Intelligence-Natural Language Processing*, pp. 000–000, 2023.

Conference Papers

A. Sittar and D. Mladenic, "How are the economic conditions and political alignment of a newspaper reflected in the events they report on?" In *Central European Conference on Information and Intelligent Systems*, Faculty of Organization and Informatics Varazdin, 2021, pp. 201–208.

A. Sittar, D. Mladenić, and T. Erjavec, "A dataset for information spreading over the news," in *Proceedings of the 23th International Multiconference Information Society SiKDD*, vol. 100, 2020, pp. 5–8.

A. Sittar and D. Mladenic, "Classification of cross-cultural news events," in *Proceedings of the 24th International Multiconference Information Society SiKDD*, vol. 100, 2021, pp. 29–32.

A. Sittar and D. Mladenic, "Using the profile of publishers to predict barriers across news articles.," in *CLEOPATRA@ WWW*, vol. 001, 2021, pp. 47–60.

S. Gottschalk, E. Kacupaj, S. Abdollahi, *et al.*, "Oekg: The open event knowledge graph.," in *CLEOPATRA@ WWW*, 2021, pp. 61–75.

G. S. Cheema, S. Hakimov, A. Sittar, E. Müller-Budack, C. Otto, and R. Ewerth, "Mm-claims: A dataset for multimodal claim detection in social media," *arXiv preprint arXiv:2205.01989*, 2022.

Biography

Abdul Sittar was born in Gujrat, Pakistan on December 14, 1991.
He studied computer science at the Department of Computer Science and Information Technology in University of Gujrat. He obtained his Bachelor of Science degree in Computer Science in August 2013. The graduation thesis titled as *iQuran with voice tags* and its novelty included the training of a machine learning model to learn voice tags and deployment of the model in iPhone application.
Abdul Sittar was a senior software engineer in the mobile application department at Nextbridge (Pvt.) limited from December 2013 to October 2018, where he worked on developing android and iOS applications.
From September 2014 to October 2016, Abdul Sittar obtained his Master of Science degree in Computer Science from COMSATS University Islamabad, Lahore Pakistan. The master thesis was titled as *Semantic Methods for Detecting Mono- and Cross-language Paraphrased Text Reuse and Plagiarism.*
From October 2018 to October 2019, he worked as a lecturer at GIMS PMAS Arid Agriculture University, where he taught three courses named Programming Fundamental, Mobile Computing, and Analysis of Algorithms. Also, his duties include delivering course syllabus, conducting exams, and results compilation.
In October 2019, he enrolled as a Ph.D. student at Jožef Stefan International Postgraduate School and Jožef Stefan Institute, Ljubljana, Slovenia under an MSCA fellow working on CLEOPATRA ITN. His research work focus on Natural Language Processing (NLP) and machine learning (ML).