CHAPTER 1

Introduction

Abstract This chapter contextualizes the book in terms of aims, methods, contents, and audience. It first discusses the impact of the COVID-19 pandemic on all aspects of society, and the crucial role that social networks played as a means to disseminate information and share feelings and ideas between users. Finally, a comprehensive summary of the most outstanding research related to this book is offered, focusing on those works that employ similar techniques to the ones used here.

Keywords Social media corpora · Social media analysis · COVID-19 pandemic · Corpus-based research methods

The general aim of this book is to offer a comprehensive overview of available techniques and approaches to explore large social media corpora in general, illustrating them with Chen's (2020) Coronavirus Twitter corpus. Thus, the book pursues a double objective. First, a fundamentally methodological one, in which I describe in detail a number of methods, strategies, and tools that can be used to access, manage, and explore large Twitter/X[1] corpora; these include both user-friendly applications, such

[1] In April 2023 Twitter's legal name was changed to X Corp. In this book, I will refer to the company as Twitter/X to avoid confusion, and because all corpora discussed or used in the book were compiled prior to this name change. For the same reason, I will refer to them as Twitter corpora or datasets.

A. Moreno-Ortiz, *Making Sense of Large Social Media Corpora*,
https://doi.org/10.1007/978-3-031-52719-7_1

as Sketch Engine (Kilgarriff et al. 2014) or Lingmotif (Moreno-Ortiz 2017), and more advanced methods and libraries that involve the use of data management skills and custom scripts. These tools and methods, on the other hand, are applied to explore one of the largest Twitter datasets on the COVID-19 pandemic publicly released, covering the two years when the pandemic had the strongest impact on society. Consequently, the second important objective is to seek out, identify, and describe this impact, and how it is reflected in the language on social media.

Therefore, this book is intended to be both a methodological guide for language researchers—understood as all those who use language and textual resources in general as a data source in their research, whichever their field of application—as well as a reference for researchers in other fields who are interested in the impact of the pandemic on society and its reflection on social networks across the English-speaking world.

The tools and methods discussed in this book are described with enough technical detail for readers to apply them to their own datasets, but not so much that the description obscures the practical applications. In order to facilitate the understanding and actual application of these techniques to other datasets, the text provides user-friendly descriptions of technicalities regarding data manipulation and algorithms. In addition, all datasets and data analysis results are made freely available as a companion online data repository.[2]

Given the significance and magnitude of the COVID-19 pandemic, a large amount of related research has been produced since it started, including work similar to the one in this book. Section 1.2 contains a review of a representative selection of such studies. This study differs from these in several key aspects. Firstly, the scope is significantly larger, as I tackle the analysis of two years of pandemic (2020–2021); also, given the extended time span covered by the data, this study attempts to provide both synchronic and diachronic analyses, as I will be using timestamped data over a period of two years. Consequently, the corpus is significantly larger than those used in most of the, often rushed (Hyland and Jiang 2021), studies that have been published on the topic, thus posing a number of methodological and technical issues. Secondly, although references to the medical, social, psychological, economic, and educational aspects of the pandemic are inevitable in such a significant event, the

[2] Moreno-Ortiz, A. (2024). LSMC Datasets. https://doi.org/10.17605/OSF.IO/H5Q4J.

main focus of this book is the communicative perspective of language as a vehicular instrument of all those aspects. Thirdly, this study is highly methodological in nature, as it attempts to compare tools, techniques, and methods that are available to the language researcher in order to extract linguistic and conceptual information from very large social media corpora. Because of the emphasis on language use, I will also offer a diatopic perspective, which is made possible by the geotagged subset of tweets available in the focus corpus.

To sum up, the goal of this work is to provide methodological and practical cues on how to manage and explore the contents of a large-scale social media corpus, such as Chen et al.'s (2020) Coronavirus Twitter corpus. The primary goal is to compare and evaluate, in terms of efficiency and efficacy, available methods to extract language-focused information from large-scale social media corpora, fundamentally keyword extraction, topic modelling, and sentiment analysis. Detailed descriptions of various approaches to completing these tasks are provided, and results are compared to provide the necessary criteria for determining the benefits and drawbacks of each, as well as their suitability to various research scenarios.

1.1 The Coronavirus Pandemic on Social Media

The COVID-19 pandemic has been a determining factor in the lives of all humans in the second decade of the twenty-first century, in all aspects of our existence, especially in regard to health, social relationships, politics, the economy, and, of course, language. Until the arrival of the pandemic, concepts and terms that were foreign or altogether unknown to most of us ('pandemic', 'variants', 'antigen test', 'community spread', 'contact tracing') became progressively commonplace in our mental lexicon and everyday language. Unlike previous pandemics, the coronavirus used the sky highways to spread across the globe at an unheard-of pace, becoming a global health issue in a very short time.[3] Likewise, the information highways quickly became flooded with news and data regarding the virus, the disease, and the social and economic impact that the event brought about. Also, given the widespread use of social networks by citizens all over

[3] As of September 2023, the pandemic has claimed over 6.8 million lives (https://www.arcgis.com/apps/dashboards/bda7594740fd40299423467b48e9ecf6) [Accessed 21 September 2023].

the world, these information exchange hubs quickly became the obvious choice of many to learn about the pandemic, share their reactions, opinions, and emotions, or just reach out to the world. The "community spread" of the virus was as fast as that of the perceptions about the virus, and the vast range of social implications it triggered.

Social networks have indeed revolutionized the way we communicate and disseminate information. They have enabled people to connect across geographical boundaries, share ideas, and exchange information in real time (Boyd and Ellison 2007). This has not only increased the efficiency of communication but also democratized access to information, allowing anyone with internet access to create and share content (Castells 2009). In the business world, social networks have generated new marketing and customer engagement opportunities. Businesses are able to promote their products, interact with consumers, and collect valuable data for market research (Kaplan and Haenlein 2010), thus contributing to the expansion of e-commerce, with platforms such as Facebook and Instagram incorporating capabilities that allow users to buy and sell products directly (Zhan et al. 2016).

These powerful and enabling features are not without drawbacks. Social networks have been used to mobilize political protests, influence public opinion, and even interfere with elections (Tufekci 2017). The term 'fake news' has crystallized the now common practice of spreading misinformation among the general public, in such a way that it is sometimes not easy to tell facts from fiction, truthful from false information. In turn, this situation has, among other things, contributed to the current climate of political polarization in many societies (Allcott and Gentzkow 2017).

The coronavirus pandemic significantly amplified the importance of social networks in contemporary society. As people around the world were forced into lockdowns and social distancing measures, social networks became a lifeline for many, serving multiple purposes, including communication, information dissemination, and emotional support. Again, the dark side of social networks reared its ugly face from the very beginning. Even at the onset of the pandemic researchers were able to identify certain trends; as Depoux et al. (2020) state,

> Within weeks of the emergence of the novel coronavirus disease 2019 (COVID-19) in China, misleading rumours and conspiracy theories about the origin circulated the globe paired with fearmongering, racism and mass

> purchase of face masks, all closely linked to the new 'infomedia' ecosystems of the 21st century marked by social media. A striking particularity of this crisis is the coincidence of virology and virality: not only did the virus itself spread very rapidly, but so did the information —and misinformation— about the outbreak and thus the panic that it created among the public. (p. 1)

Similarly, Rosenberg et al. (2020) note that social networks, specifically Twitter/X, have played a significant role in the medical world, a trend that was magnified during the pandemic. This includes both positive and negative aspects. On the positive side, it has become a forum where medical professionals exchange ideas, information, and commentary, facilitating fast spread of valuable information. However, unlike traditional medical educational resources, Twitter/X's free-flow of messages and ideas is not vetted or peer-reviewed, and therefore can pose a risk of harm. In particular, misinformation, information overload, and even hysteria are mentioned as the most immediate consequences.

It is debatable whether the negative effects of social networks during the pandemic outweighed the positive ones, but it is a fact that global use of social media sites soared. For example, according to a survey of social media users in the United States, 29.7% of respondents spent an additional 1–2 hours per day on social media. A further 20.5% utilized social media 30 to 60 minutes longer than usual per day.[4] Similarly, the share of TikTok users rose from 10% before the COVID-19 pandemic to 28% after it in users aged 15–25, and from 4 to 12% in the general population.[5]

The increase in the number of posts on Twitter was also significant. Haman (2020) reports a weekly growth rate of 1.5% in the number of followers experienced by the accounts of state leaders after March 9, 2020. Ahmed et al. (2021) also note that during the first year of the pandemic, social media participation and interaction increased dramatically, as it provided a forum for individuals to share their perceptions and perspectives on the medical, economic, and social crisis.

[4] https://www.statista.com/statistics/1116148/more-time-spent-social-media-platforms-users-usa-coronavirus/#statisticContainer [Accessed 4 March 2023].

[5] https://www.statista.com/statistics/1207831/tiktok-usage-among-young-adults-during-covid-19-usa/ [Accessed 4 March 2023].

As of 2021, there were 4.26 billion social network users around the world, a number projected to increase to almost six billion by 2027.[6] Additionally, social networking sites generate the most user engagement (Chong and Park 2021). In terms of social network market share, Twitter/X is in third position (6.82%) in 2023, after Facebook (36.64%) and YouTube (27.01%).[7] However, Twitter/X has considerably more media coverage, as politicians worldwide use this network to communicate their messages, both as an amplifier of their party's ideology and a substitute that allows them to express a more individualised message (Silva and Proksch 2022). This interaction, in turn, causes the public to react and generate political content themselves.

In addition, the defining characteristic of Twitter/X versus other social media networks is the availability of its data by means of an API (Application Programming Interface), a set protocols provided by Twitter that developers can use to access and download its data. This is the reason why most of the research on social networks has utilized this Twitter as a data source.

1.2 Related Research

The COVID-19 pandemic and its effect on society has been studied using social media content, and user-generated content in particular, from a wide range of perspectives and fields of study, including keyword extraction, topic modelling, and sentiment analysis, the main methods that are described in this book. However, most of these studies are limited in the time span they considered, the geographical scope, and/or the type of methods and techniques employed, the likely reason being that many researchers rushed to publish results of studies during the first months of the pandemic given the significance of the event.

In this section, I aim to describe this body of research in order to shed light on what type of studies utilize the methods discussed in this book, while at the same time describe some results obtained by studies that use similar methods to the ones discussed here. In the literature review that

[6] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ [Accessed 4 March 2023].

[7] https://www.dreamgrow.com/top-10-social-networking-sites-market-share-of-visits/.

follows, a distinction is apparent according to the field of study that motivated the research and, consequently, the techniques that were employed. Whereas researchers in (corpus) linguistics fundamentally employ "off-the-shelf" tools and methods common in the field, including distant and close reading techniques (word frequency, keyword extraction, concordancing, multi-dimensional analysis, qualitative analysis), researchers in the social sciences skewed towards strictly distant reading, i.e. quantitative methods and tools developed within Natural Language Processing (NLP), such as topic modelling and sentiment analysis. Accordingly, the results obtained are of a different nature.

This section is intended to illustrate the kind of research that can be conducted employing the techniques and tools described in this book, as well as to contextualize its actual content. It begins with some relevant research works in the social sciences, followed by those in corpus linguistics.

As a typical piece of research in the social sciences, Boon-Itt and Skunkan's (2020) study aimed to determine the public's awareness of COVID-19 pandemic trends and to identify significant themes of concern expressed by Twitter/X users in English. They gathered a total of 107,990 tweets relating to COVID-19 between December 13, 2019, and March 9, 2020. Over the limited time of their study, they used keyword frequency, sentiment analysis, and topic modelling to identify and investigate discussion topics. The study concluded that sentiment analysis and topic modelling can produce useful information about the trends in the discussion of the COVID-19 pandemic on social media, which is in fact, applicable to any corpus.

In a similar study, but including geographical stratification, Dubey (2020) collected 50,000 tweets from 11 countries every 4 days, from March 11 to March 31, 2020, using several search keywords and performed sentiment and emotion analysis using the NRC Emotion Lexicon (Mohammad and Turney 2010, 2013). The study presents word clouds of tweets from each country, which visually represent the most used terms and primary topics of conversation in each country.

The study by Ahmed et al. (2021) provides evidence on how the pandemic has not only triggered a significant global public health crisis, but also other problems, such as economic crisis, job loss, and mental anxiety. They analysed the sentiment of Twitter users at various time intervals to identify trending topics, and generated sentiment-related word clusters from several conceptual categories.

Kruspe et al. (2020) provide a cross-linguistic analysis of tweets posted in several European countries during 2020. The corpus consisted of approximately 4.6 million geotagged tweets in 60 different languages. The tweets were not filtered by subject, so many of them were unrelated to COVID-19. This was intentional, as the researchers were interested in the effect of the pandemic on people's mood in general, not just in relation to the outbreak. The study used an automatic method for sentiment analysis, training a neural network on the Sentiment140 dataset (Go et al. 2009), which contains around 1.5 million tweets collected through keyword search, and then annotated automatically by detecting emoticons. They found that there was a general downward trend in sentiment in the last few months corresponding to the pandemic, with clear dips at times of lockdown announcements and a slow recovery in the following weeks in most countries. Prior to February 2020, the use of pandemic-related keywords was uncommon, and the increase in Covid cases in each country correlates with an increased usage of those terms. The sentiment of tweets began as extremely negative at the onset of the pandemic, and then gradually became more positive. Nevertheless, it remained significantly below the average sentiment in most nations. They also found that there was a slight improvement in sentiment in the majority of countries towards the end of the period examined. As we will see in Chapter 6 these findings are in line with the sentiment analysis results presented in this book.

Mujahid et al.'s (2021) study combined sentiment analysis and topic modelling to investigate the efficacy of online education by analysing the sentiment of its stakeholders using Twitter data. It utilized machine learning techniques for annotation and topic modelling to identify e-learning issues, as expressed on Twitter by students, teachers, and other administrators. The dataset consists of 17,155 tweets, collected by searching for tags such as 'coronaeducation', 'covidneducation', 'distancelearning', and 'onlinelearning'. Tweets were classified using the sentiment lexicons provided by TextBlob (Keen et al. 2023), VADER (Hutto and Gilbert 2014), and SentiWordNet (Baccianella et al. 2010) and using several machine learning classification algorithms (LSTM, CNN, CNN-LSTM, and biLSTM), with and without data balancing with Synthetic Minority Over-sampling Technique (SMOTE). Topic modelling was used to identify the issues associated with e-learning, revealing that

the top three issues are the uncertainty of campus opening date, children's inability to comprehend online education, and the lack of efficient networks for online education.

Lyu et al (2021) carried out a study aimed at examining Twitter conversations about COVID-19 vaccines between March 11, 2020, and January 31, 2021, in order to identify the most prevalent topics and sentiment regarding vaccine-related issues and analyse the evolution of these topics and sentiment over time. The corpus consisted of approximately 1.5 million unique tweets collected from March 11, 2020 through January 31, 2021. The authors used the topic modelling implementation of the Latent Dirichlet Allocation (LDA) algorithm in the R *textmineR* package (Jones et al. 2023) to generate an initial labelling for the topics. After carefully reading through a sample of tweets from each topic, they refined the machine-generated labels to provide a more accurate, concise, and consistent description of each topic.[8] As in Dubey's study, sentiment and emotion analysis were performed using the NRC Emotion Lexicon (Mohammad and Turney 2013), thus assigning scores to various emotions, including anger, fear, anticipation, trust, surprise, sadness, happiness, and disgust. The study revealed that among the sixteen distinct topics, vaccination-related opinions were the most prevalent and remained so over time. As global vaccine development progressed, the dominant subjects also shifted. Instructions on how to obtain the vaccine became the most-discussed topic at the beginning of January 2021. Also, the discussion of COVID-19 vaccination on social media was largely influenced by significant news events. The increasing positivity and predominance of trust over time suggests that social media discussions may indicate greater acceptance of COVID-19 vaccines in comparison with previous vaccines.

Using data from Reddit, Melton et al. (2021) also carried out a study combining sentiment analysis and topic modelling to examine public opinions regarding the COVID-19 vaccine. The corpus consisted of approximately 9,000 Reddit posts, which collectively obtained over 600,000 upvotes. The researchers combined these two techniques in a novel way: first they classified posts from a sentiment perspective using the TextBlob toolkit (Keen et al. 2023), which offers both a subjectivity score

[8] In Chapter 5 of this book a more advanced method for labelling extracted topics is presented which provides high-quality titles employing state-of-the-art Large Language Models.

and sentiment classification. Then they used LDA-based topic modelling in two different ways: first, they used the global time-series dataset to extract topics over time, and then they extracted the topics from the sentiment-classified posts. The results indicate that the public sentiment in Reddit communities is generally positive regarding discussions about the experiences with receiving the vaccine, although keywords and topics were identified that indicate some reluctance among users. They did not find significant changes in sentiment over time, which they attributed to a potential bias in the Reddit communities and/or strict community guidelines that result in the removal of certain posts, thereby creating an echo chamber. Unsurprisingly, they found topic modelling hard to evaluate, as the quantification through the coherence and perplexity scores is not a good indicator of performance in topic extraction. This is something that we will revisit in Chapter 5, as it is an important issue in topic modelling in general. The results of the LDA analysis revealed a total of five optimal latent topics. The first four topics appear to be closely related to a more comprehensive discussion of the vaccine, safety concerns, efficacy, and potential side effects. Topic 5 appeared to be centred on much broader terms, information (e.g. news, source, question), and a direct mention of vaccination-related concerns. Autism was also identified as a topic, presumably in reference to the antivaccine movement's fixation on the myth that vaccines cause this disorder.

News articles is undoubtedly the other major source of data to analyse text using "distant reading" techniques, such as keyword analysis, sentiment analysis, and topic modelling. Ghasiya and Okamura (2021) used a corpus of over 100,000 COVID-19 news headlines and articles from January 1, 2020, to December 1, 2020, in order to examine the key topics, trends, and themes of English-language COVID-19 news articles across four countries (UK, India, Japan, South Korea). For topic modelling they used *top2vec* (Angelov 2020), which is able to jointly generate embedded topic, document, and word vectors. As we will see in Sect. 5.2, *top2vec* has a number of advantages over traditional topic modelling techniques, such as LDA, as it makes the task simpler by doing all pre-processing (stop-word removal, stemming, lemmatization) automatically, and, importantly, it does not require prior knowledge of existing topics to produce a good topic model (Le and Mikolov 2014). For sentiment analysis, they used a RoBERTa-based (Vaswani et al. 2017) sentiment classifier. Similarly to Melton et al.'s (2021) approach, they also used the output of the classified headlines to extract positive and

negative topics. They found that the economy, education, and sports were the sectors most affected by the COVID-19 pandemic. The United States topped every dataset for two reasons: first, it was the country most severely affected and, second, due to the global significance of the presidential elections. The study also revealed that the United Kingdom's media had strong negative attitudes regarding the pandemic and other related issues. In the Indian dataset, the negative headlines were only slightly more frequent than positive ones, whereas in Japan the difference was significantly bigger, with 57.38% negative and 42.61% positive headlines. South Korea had the most positive data of the four countries, with 54.47% positive and 45.52% negative headlines.

On the other hand, the body of corpus linguistics research exhibits some important differences as compared to the social sciences. Corpus linguistics researchers focus primarily on newspaper text rather than social media, the corpora tend to be smaller, and the methods more qualitative in nature. For example, in a special issue of the *International Journal of Corpus Linguistics*, Hyland and Jiang (2021) investigate the language used in COVID-19 scientific publications, given the deluge of papers that were hastily published immediately after the onset of the pandemic, which had certain negative consequences. The authors argue that the urgency and competition surrounding COVID-19 research led to an increase in the use of promotional language, or *hyping*, in scientific papers. They used a corpus of 1,000 COVID-19 research papers published in the first seven months of 2020, and compared it to a reference corpus of 1,000 papers from the same journals in 2015. They used the *AntConc* (Anthony 2023) concordancing software to analyse the frequency of certain features in the texts, such as boosting markers, affective markers, and self-mentions, that is, markers of what is usually referred to as *hyping language*. The results indicated that these markers were significantly more recurrent in the COVID-19 papers than in the reference corpus. The authors discovered that the former were more assertive and definitive in their presentation of results, with a steady increase in hyping features over time. They also discovered that scientists were more "present" in their texts, frequently highlighting the potential future value of their research and its potential contribution to the resolution of the pandemic.

In the same volume, Dong et al. (2021) explore the changes in the use of attitudinal markers in academic and media discourse on COVID-19, with a view to understanding how the use of these markers correlates with

the reported cases of the disease. They used two different methodologies; on the one hand, a discourse dynamics approach in order to analyse language according to the evolution of discourse over time and, on the other, they applied Appraisal Theory (Martin and White 2005) to describe and explain the way language is used to evaluate, adopt stances, construct textual personas, and manage interpersonal positioning and relationships. The authors also used LOESS regression, a non-parametric method that uses local weighted regression to fit a smooth curve through points in a scatter plot. The corpus for this study consisted of academic articles and media reports related to COVID-19; the former were sourced from the COVID-19 Open Research Dataset (CORD-19), while the latter were obtained from the BYU Coronavirus Corpus (Davies 2021); both these corpora are described in Chapter 2. The authors ensured the comparability of the two corpora at different observation points by segmenting the media corpus in accordance with the size of the CORD-19 academic corpus in the same period. They found a complex and intricate interaction between the use of affect markers in both corpora with regard to the four types of reported cases of COVID-19: total cases per million, new cases per million, total deaths per million, and new deaths per million. The use of the variable 'new cases per million' was found to strongly correlate with the occurrence of affect markers in the academic corpus at some observation points, whereas the variable 'new deaths per million' was also found to correlate strongly with appreciation markers in the academic and media corpus in some periods. They also identified a fluctuating correlation between judgement markers and the reported cases of COVID-19, in both the academic and media corpora.

As an example of research that focuses on government communication, Gallardo-Pauls (2021) carried out an analysis of risk communication in the context of emergencies. She proposes a discursive model for risk communication, focusing on the Spanish context during the COVID-19 pandemic. The paper examines the communication strategies used by the Spanish government during the pandemic, focusing on the discursive elements that contribute to the perception and understanding of risk. The corpus used in the study consists of the communication materials and strategies used by the Spanish government during the pandemic, including press conferences, official statements, and other forms of public communication. The research identified several dangers associated with risk communication, including complexity of the risk's description, ambiguity in data interpretation, and the domino effects of risk. A distinction

is made between 'old' and 'new' risks, with the former being natural risks and the latter being technological or human-made. The paper proposes a discursive model for risk communication that takes into account these risks and aims to improve the effectiveness of communication during emergencies.

Within Critical Discourse Analysis, Florea and Woelfel (2022) analysed a corpus of news reports from four major global TV news providers—CNN, BBC, DW, and RT—covering the COVID-19 pandemic from its outbreak to mid-crisis in 2020; they analysed a total of 12 dataset reports consisting of approximately two million words, to which they applied multi-level content analysis and Proximization Theory (Cap 2013). They used the *Catpac Pony* software package to identify the occurrence of concepts and the semantic relationship among the highly frequent clusters. The results suggest that the news texts surrounding the COVID-19 pandemic formulate a particular type of discourse on suffering that individualizes the sufferer, sets out the course of action, and turns the pandemic into a global cause for action. The negative values of the pandemic, among which they highlight the devastating economic impact, were found to legitimize the proximal discourses of suffering and safety.

As an example of a corpus-based study that focuses on gender and political communication, Power and Crosthwaite (2022) aimed to investigate the crisis communication during the pandemic of two political leaders, Jacinda Ardern, Prime Minister of New Zealand, and Scott Morrison, Prime Minister of Australia. The authors aim to understand how these two leaders' communication styles differed and how these differences might be related to their gender identities. The corpus consists of statements published by these two PMs during 2020, focusing on those that made reference to the pandemic by using the search terms "covid*", "coronavirus", and "pandemic" at least once. They focused on monological genres, as the proportion of dialogical genres was much higher in one of the leaders (Morrison). The corpus was divided into subcorpora reflecting the leader's identity and the relative status of the epidemic curve in each country. The reduced size of the corpus (24,083 tokens in total) allowed the authors to use the *Scattertext* (Kessler 2017) visualizer to compare the keywords associated with each PM. Thus, the results of the quantitative analysis are presented in the form of Scattertext's keyword comparisons charts for the entire corpus and for specific periods of the pandemic (initial case period, steep curve-rising periods, curve flattening periods, and flat curve periods). Guided by

the quantitative results, they then produced a qualitative analysis based on Stokoe's (1998) gender conceptualization framework and transitivity analysis (Halliday and Matthiessen 2014) in order to ascertain whether the observed differences were gender-motivated.

This study is a good example of how extraction and visualization of quantitative data can be an entry point to detailed qualitative analysis, as it provides critical cues on what to focus on. Keyword extraction is probably the most useful quantitative method to quickly obtain insights from a corpus since keywords somehow condense the corpus' contents. Chapter 4 of this book focuses on this topic, exploring the different concepts, approaches, methods, and tools.

References

Ahmed, Md Shoaib, Tanjim Taharat Aurpa, and Md Musfique Anwar. 2021. Detecting Sentiment Dynamics and Clusters of Twitter Users for Trending Topics in COVID-19 Pandemic. *PLOS ONE* 16: e0253300. Public Library of Science. https://doi.org/10.1371/journal.pone.0253300.

Allcott, Hunt, and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31: 211–236. https://doi.org/10.1257/jep.31.2.211.

Angelov, Dimo. 2020. Top2Vec: Distributed Representations of Topics.

Anthony, Laurence. 2023. AntConc (Version 4.2.0). Tokyo, Japan: Waseda University.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the International Conference on Language Resources and Evaluation*, 2200–2204. Valletta, Malta.

Boon-Itt, Sakun, and Yukolpat Skunkan. 2020. Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study. *JMIR Public Health and Surveillance* 6: e21978. https://doi.org/10.2196/21978.

Boyd, Danah m., and Nicole B. Ellison. 2007. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication* 13: 210–230.https://doi.org/10.1111/j.1083-6101.2007.00393.x

Cap, Piotr. 2013. *Proximization. pbns.232.* John Benjamins Publishing Company.

Castells, Manuel. 2009. *Communication Power*. Oxford University Press.

Chen, Emily, Kristina Lerman, and Emilio Ferrara. 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health and Surveillance* 6: e19273. https://doi.org/10.2196/19273.

Chong, Miyoung, and Han Woo Park. 2021. COVID-19 in the Twitterverse, from Epidemic to Pandemic: Information-Sharing Behavior and Twitter as an Information Carrier. *Scientometrics* 126: 6479–6503. https://doi.org/10.1007/s11192-021-04054-2.

Davies, Mark. 2021. The Coronavirus Corpus: Design, Construction, and Use. *International Journal of Corpus Linguistics* 26: 583–598. John Benjamins Publishing Company. https://doi.org/10.1075/ijcl.21044.dav

Depoux, Anneliese, Sam Martin, Emilie Karafillakis, Raman Preet, Annelies Wilder-Smith, and Heidi Larson. 2020. The Pandemic of Social Media Panic Travels Faster than the COVID-19 Outbreak. *Journal of Travel Medicine* 27: taaa031. https://doi.org/10.1093/jtm/taaa031.

Dong, Jihua, Louisa Buckingham, and Hao Wu. 2021. A Discourse Dynamics Exploration of Attitudinal Responses Towards COVID-19 in Academia and Media. *International Journal of Corpus Linguistics* 26: 532–556. John Benjamins. https://doi.org/10.1075/ijcl.21103.don.

Dubey, Akash Dutt. 2020. Twitter Sentiment Analysis during COVID-19 Outbreak. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.3572023.

Florea, Silvia, and Joseph Woelfel. 2022. Proximal versus Distant Suffering in TV News Discourses on COVID-19 Pandemic. *Text & Talk* 42: 327–345. De Gruyter Mouton. https://doi.org/10.1515/text-2020-0083.

Gallardo-Pauls, Beatriz. 2021. Riesgos de la comunicación de riesgo: un modelo discursivo para la comunicación de riesgo en emergencias. *Círculo de Lingüística Aplicada a la Comunicación* 88: 135–154. https://doi.org/10.5209/clac.77761.

Go, Alec, Richa Bhayani, and Lei Huang. 2009. Twitter Sentiment Classification Using Distant Supervision. *CS224N Project Report, Stanford*.

Ghasiya, Piyush, and Koji Okamura. 2021. Investigating COVID-19 News Across Four Nations: A Topic Modeling and Sentiment Analysis Approach. *IEEE Access: Practical Innovations, Open Solutions* 9: 36645–36656. https://doi.org/10.1109/ACCESS.2021.3062875.

Halliday, M. a. K., and Christian M. I. M. Matthiessen. 2014. *An Introduction to Functional Grammar*. Routledge.https://doi.org/10.4324/9780203783771.

Haman, Micahel. 2020. The Use of Twitter by State Leaders and Its Impact on the Public During the COVID-19 Pandemic. *Heliyon* 6: e05540. Elsevier. https://doi.org/10.1016/j.heliyon.2020.e05540.

Hutto, C., and E. Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the International AAAI Conference on Web and Social Media*, 216–225.

Hyland, Ken, and Feng (Kevin) Jiang. 2021. The Covid Infodemic: Competition and the Hyping of Virus Research. *International Journal of Corpus Linguistics* 26: 444–468.https://doi.org/10.1075/ijcl.20160.hyl.

Jiang, Julie, Xiang Ren, and Emilio Ferrara. 2021. Social Media Polarization and Echo Chambers in the Context of COVID-19: Case Study. *Jmirx Med* 2: e29570. https://doi.org/10.2196/29570.

Jones, Tommy, William Doane, and Mattias Attbom. 2023. textmineR: Functions for Text Mining and Topic Modeling (version 3.0.5).

Kaplan, Andreas M., and Michael Haenlein. 2010. Users of the World, Unite! The Challenges and Opportunities of Social Media. *Business Horizons* 53: 59–68.

Keen, Peter, Matthew Honnibal, Roman Yankovsky, David Karesh, and Evan Dempsey. 2023. TextBlob: Simplified Text Processing.

Kessler, Jason. 2017. Scattertext: A Browser-Based Tool for Visualizing How Corpora Differ. In *Proceedings of ACL 2017, System Demonstrations*, 85–90. Vancouver, Canada: Association for Computational Linguistics.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten Years On. *Lexicography*: 7–36.

Kruspe, Anna, Matthias Häberle, Iona Kuhn, and Xiao Xiang Zhu. 2020. Cross-Language Sentiment Analysis of European Twitter Messages During the COVID-19 Pandemic. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online: Association for Computational Linguistics.

Le, Quoc, and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning—Volume 32*, II-1188-II–1196. ICML'14. Beijing, China: JMLR.org.

Lyu, Joanne Chen, Eileen Le Han, and Garving K. Luli. 2021. COVID-19 Vaccine-Related Discussion on Twitter: Topic Modeling and Sentiment Analysis. *Journal of Medical Internet Research* 23: e24435. https://doi.org/10.2196/24435.

Martin, James R., and Peter R. R. White. 2005. *The Language of Evaluation: Appraisal in English*. Basingstoke: Palgrave Macmillan.

Melton, Chad A., Olufunto A. Olusanya, Nariman Ammar, and Arash Shaban-Nejad. 2021. Public Sentiment Analysis and Topic Modeling Regarding COVID-19 Vaccines on the Reddit Social Media Platform: A Call to Action for Strengthening Vaccine Confidence. *Journal of Infection and Public Health* 14. Special Issue on COVID-19—Vaccine, Variants and New Waves: 1505–1512. https://doi.org/10.1016/j.jiph.2021.08.010.

Mohammad, Saif M., and Peter D. Turney. 2010. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches*

to Analysis and Generation of Emotion in Text, 26–34. Association for Computational Linguistics.

Mohammad, Saif M., and Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence* 29: 436–465.

Moreno-Ortiz, Antonio. 2017. Lingmotif: Sentiment Analysis for the Digital Humanities. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 73–76. Valencia, Spain: Association for Computational Linguistics.

Mujahid, Muhammad, Ernesto Lee, Furqan Rustam, Patrick Bernard Washington, Saleem Ullah, Aijaz Ahmad Reshi, and Imran Ashraf. 2021. Sentiment Analysis and Topic Modeling on Tweets about Online Education during COVID-19. *Applied Sciences* 11. Multidisciplinary Digital Publishing Institute: 8438. https://doi.org/10.3390/app11188438.

Power, Kate, and Peter Crosthwaite. 2022. Constructing COVID-19: A Corpus-Informed Analysis of Prime Ministerial Crisis Response Communication by Gender. *Discourse & Society* 33: 411–437. Sage Publications Ltd.

Rosenberg, Hans, Shahbaz Syed, and Salim Rezaie. 2020. The Twitter Pandemic: The Critical Role of Twitter in the Dissemination of Medical Information and Misinformation During the COVID-19 Pandemic. *Canadian Journal of Emergency Medicine* 22: 418–421. Cambridge University Press. https://doi.org/10.1017/cem.2020.361.

Silva, Bruno Castanho, and Sven-Oliver Proksch. 2022. Politicians Unleashed? Political Communication on Twitter and in Parliament in Western Europe. *Political Science Research and Methods* 10: 776–792. Cambridge University Press. https://doi.org/10.1017/psrm.2021.36.

Stokoe, Elizabeth H. 1998. Talking about Gender: The Conversational Construction of Gender Categories in Academic Discourse. *Discourse & Society* 9: 217–240. Sage Publications Ltd. https://doi.org/10.1177/0957926598009002005.

Tufekci, Zeynep. 2017. *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. Yale University Press.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010. NIPS'17. Long Beach, CA, USA: Curran Associates Inc.

Zhan, Liuhan, Yongqiang Sun, Nan Wang, and Xi Zhang. 2016. Understanding the Influence of Social Media on People's Life Satisfaction Through Two Competing Explanatory Mechanisms. *Aslib Journal of Information Management* 68: 347–361. Emerald Group Publishing Limited. https://doi.org/10.1108/AJIM-12-2015-0195.