

Research on Analysis of Social Phenomena in the Cyberspace by Utilizing Data-driven Approach

ガーシャ, ピユシュ

<https://hdl.handle.net/2324/4496073>

出版情報 : Kyushu University, 2021, 博士 (学術), 課程博士
バージョン :
権利関係 :



Research on Analysis of Social Phenomena in the Cyberspace by Utilizing Data-driven Approach

Piyush Ghasiya

September 2021

Doctoral Thesis

Department of Advanced Information Technology
Graduate School of Information Science and Electrical
Engineering (ISEE)

DEDICATED TO

my mom *Rajrani*, my dad *Tarachand*, and my elder sister *Divya*

Without their unconditional love, immense patience, and unreserved support, this
journey would be impossible

Abstract

The epoch of Big data offers several opportunities to conduct impactful research and because of that, data-driven research has risen to prominence during the past decade. To perform data-driven research on textual data or natural language data, we need a technique or techniques that can comprehend this textual data. NLP is a field in machine learning with the ability to understand, analyze, manipulate, and potentially generate human language. Because of this reason, NLP and ML techniques are indispensable when the data-driven research is applied on natural language data. The data in our research is textual data such as news headlines and articles, hence we utilized NLP and ML techniques. For our case study, we have chosen two social phenomenon – Cybersecurity and the COVID-19 pandemic. Cybersecurity is an important multi-disciplinary field of study, and it demands understanding from several perspectives. Similarly, the COVID-19 pandemic is a health crisis that impacted more than 200 countries or territories worldwide.

There are three main contributions of our research. The first two contributions are related to the two distinct NLP approaches: topic modeling and sentiment analysis. In topic modeling, the specification of optimal number of topics is essential, however usually it is unknown. We utilized TC-W2V metric to discover the optimal number of topics in our corpus. For supervised sentiment classification model, an appropriate labeled dataset is

required. As COVID-19 is new issue, labeled dataset was unavailable, so to fulfill this requirement, we created a COVID-19 related labeled dataset to train and test our supervised sentiment classification model using a novel method. The third contribution is related to overall research methodology. In our research methodology, we have combined topic modeling and sentiment analysis approaches. Benefit of applying this approach is that while topic modeling can help us find critical themes from large dataset. During the sentiment analysis step, the sentiment of these critical issues can be understood along with the sentiment of overall corpus.

Lastly, our research also presented several findings. The key finding of cybersecurity case study are - the US as the biggest influencer in the Cybersecurity field, and the US media's prioritization of domestically critical issues. Further, a countries' strategic interests also decide which news it would report and which it would not. The sentiment analysis results shows that Russian interference in the 2016 US election garnered high (48.6%) negative articles. This result shows that the Russian interference issue has high negative sentiment.

The key findings of the COVID-19 case study are - the US, Education, Sports, and Economic are some of the most widely reported topics/issues in all four countries in the corpus. This result corroborates with the real situation as these issues are the most impacted during the COVID-19 pandemic. The sentiment analysis of COVID-19 news headlines showed that the UK has a high (73.23%) percentage of negative headlines, and South Korea has a high (54.47%) percentage of positive headlines. These results also correspond to reality because, in terms of the impact of the COVID-19 pandemic, the UK is the worst, and South Korea is the most successful country in our dataset.

Acknowledgements

This Ph.D. is an amazing journey for me. It would not have been possible without the support and guidance that I received from many people. Here I would like to acknowledge those wonderful people.

I want to thank my advisor Prof. Koji Okamura. He accepted me as his student even though I came from a different field. Prof. Okamura was always supportive, gave me a free hand to do whatever I wanted to do, and encouraged me all the time.

I would like to thank my supervisor Dr. Tsunenori Mine, whose expertise and knowledge of the subject matter was invaluable for improving my Ph.D. thesis. Dr. Mine's insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would also like to acknowledge my deputy supervisors Prof. Sachio Hirokawa and Dr. Hiroshi Koide, for their valuable and constructive suggestions to improve my Ph.D. thesis.

I want to thank Alaa San, Tam San, Ariel San, and Ono San. I always bothered them with the problems (small or big) I faced, whether research-related or personal. I want to thank all the members (past and present) of Okamura Lab for their support. My life in Japan and university would be very boring and difficult if I did not have the companionship of Niisato San, Akiko San, Hashiguchi San, Mayumi San, and Yayoi San. For that, I will always be grateful to them. I would also like to thank Maeda San, Oiwa San, and all the SCICORP staff, ENPIT staff, and office staff of Kyushu

University.

A big thanks to all my Indian friends at Kyushu University; their presence changed my monotonous life into a fun ride.

I finish with India, where my source of energy resides: My Parent and elder sister. Without their unconditional love, sacrifices, and support, this Ph.D. would not be possible. Last but not least, I would like to thank my dear friend Rupali for her unreserved support.

Contents

Acknowledgements	iii
List of Figures	x
List of Tables	xiii
1 Introduction	2
1.1 Background	2
1.2 Focus and Scope	8
1.2.1 Cybersecurity	8
1.2.2 The COVID-19 Pandemic	11
1.3 Research Objectives	13
1.3.1 Exhibit the application of NLP	13
1.3.2 Understanding and analyzing Cybersecurity and COVID-19 pandemic News	13
1.4 Challenges	14
1.5 Limitations	15
1.6 Main Findings and Contributions	16
1.6.1 Findings	16
1.6.2 Contributions	17
1.7 Thesis Overview	19

2	Literature Survey	21
2.1	Literature Survey (Case Study – Cybersecurity)	21
2.2	Literature Survey (Case Study – The COVID-19 pandemic)	24
2.2.1	Literature survey from Data Source Perspective	24
2.2.2	Literature Survey based on Research Methods	26
2.2.2.1	Topic Modeling	26
2.2.2.2	Sentiment Analysis	27
2.2.2.3	Topic Modeling and Sentiment Analysis	28
3	Main Methods	30
3.1	Text Preprocessing	30
3.2	Topic Modeling	34
3.2.1	NMF	35
3.2.2	Top2Vec	39
3.3	Clustering	40
3.4	Sentiment Analysis	42
4	Understanding Cybersecurity News	46
4.1	Motivation	46
4.2	Research Methodology	47
4.2.1	Data Acquisition	47
4.2.2	Text Preprocessing	48
4.2.3	NMF Topic Modeling	50
4.2.4	Sentiment Analysis	51
4.3	Experiment, Results, and Discussion	52
4.3.1	NMF Topic Modeling on Preprocessed Cybersecurity Articles	52
4.3.1.1	NMF Topic Modeling for Canada	52
4.3.1.2	NMF Topic Modeling for India	54

4.3.1.3	NMF Topic Modeling for Japan	56
4.3.1.4	NMF Topic Modeling for South Korea	59
4.3.1.5	NMF Topic Modeling for the UK	61
4.3.1.6	NMF Topic Modeling for the US	63
4.3.1.7	Descriptive Analysis of Topic Models in Cybersecurity Corpus	65
4.3.2	Sentiment Analysis	67
4.3.2.1	Unsupervised (Lexicon-based) Sentiment Analysis . . .	68
4.3.2.2	Supervised Sentiment Analysis	70
4.3.2.3	Predicting Sentiments of Cybersecurity News Articles and Interpretation of Results	71
4.3.2.3.1	Sentiments of China and Huawei related cyber- security news articles and Discussion	72
4.3.2.3.2	Sentiments of Russia and Election-related cy- bersecurity news articles and Discussion . . .	74
4.3.2.3.3	Sentiments of Iran related cybersecurity news articles and Discussion	76
4.3.2.3.4	Sentiments of North Korea related cybersecu- rity news articles and Discussion	77
4.3.2.3.5	Sentiments of Vietnam related cybersecurity news articles and Discussion	79
4.4	Limitations	81
5	Topic Modeling and Clustering for COVID-19 News	82
5.1	Motivation	82
5.2	Research Methodology	83
5.2.1	Data Acquisition	84

5.2.2	Top2Vec and K-means Clustering	85
5.2.3	NMF Topic Modeling with TC-W2V Metric Measure	86
5.3	Experiment, Results, and Discussion	86
5.3.1	Top2Vec Topic Modeling	86
5.3.1.1	Describing India’s Top Topics	87
5.3.1.2	Describing Japan’s Top Topics	89
5.3.1.3	Describing South Korea’s Top Topics	90
5.3.1.4	Describing UK’s Top Topics	92
5.3.1.5	Comparative Analysis	94
5.3.2	K-means Clustering with Davies Bouldin Index	94
5.3.2.1	K-means Clustering on Indian Topics	95
5.3.2.2	K-means Clustering on Japan’s Topics	96
5.3.2.3	K-means Clustering on South Korea’s Topics	97
5.3.2.4	K-means Clustering on UK’s Topics	98
5.3.3	NMF Topic Modeling of COVID-19 Dataset and Comparison with Top2Vec	101
5.3.3.1	NMF Topic Modeling and Comparative Analysis with Top2Vec for Indian Dataset	104
5.3.3.2	NMF Topic Modeling and Comparative Analysis with Top2Vec for Japan’s Dataset	105
5.3.3.3	NMF Topic Modeling and Comparative Analysis with Top2Vec for South Korean Dataset	109
5.3.3.4	NMF Topic Modeling and Comparative Analysis with Top2Vec for the UK Dataset	110
5.3.4	Pros and Cons of NMF and Top2Vec	112
5.4	Limitations	114
5.5	Consideration	114

6	Sentiment Analysis of COVID-19 News	115
6.1	Motivation	115
6.2	Research Methodology	116
6.2.1	Data Acquisition and Text Preprocessing	116
6.2.2	Labeling Headlines for Sentiment Classification	118
6.2.3	Sentiment Classification with RoBERTa	120
6.3	Experiment, Results, and Discussion	120
6.3.1	Sentiment Classification	120
6.3.2	Sentiment Analysis	122
6.3.2.1	Sentiment Analysis of India’s COVID-19 Headlines	122
6.3.2.2	Sentiment Analysis of Japan’s COVID-19 Headlines	124
6.3.2.3	Sentiment Analysis of South Korea’s COVID-19 Headlines	125
6.3.2.4	Sentiment Analysis of UK’s COVID-19 Headlines	126
6.3.3	Comparative Analysis of All Four Nation’s Sentiments	127
6.4	Limitations	128
7	Conclusion	130
7.1	Cybersecurity	131
7.2	COVID-19 Topic Modeling	132
7.3	COVID-19 Sentiment Analysis	134
	Bibliography	136
	Publications	148

List of Figures

1.1	Flow of Data-driven Research	4
1.2	The Field of NLP	5
3.1	Text Preprocessing Steps	33
3.2	Representation of NMF equation	36
3.3	Four steps showing the working of the Top2Vec algorithm . . .	41
3.4	Various Approaches for Sentiment Analysis	44
4.1	Research Methodology for Cybersecurity Case Study	48
4.2	Mean Coherence and Number of Topics (Canada)	53
4.3	All Topics in MDS (Canada)	55
4.4	Mean Coherence and Number of Topics (India)	56
4.5	All Topics in MDS (India)	58
4.6	Mean Coherence and Number of Topics (Japan)	59
4.7	All Topics in MDS (Japan)	61
4.8	Mean Coherence and Number of Topics (South Korea)	62
4.9	All Topics in MDS (South Korea)	64
4.10	Mean Coherence and Number of Topics (UK)	65
4.11	All Topics in MDS (UK)	67
4.12	Mean Coherence and Number of Topics (US)	68

4.13	All Topics in MDS (US)	70
4.14	Sentiment Analysis results (in percentage) on our cybersecurity news articles dataset	72
4.15	China and Huawei related articles in our dataset	73
4.16	China and Huawei related articles sentiments	73
4.17	Russia and Election-related articles in our dataset	74
4.18	Russia and Election-related articles sentiments	75
4.19	Iran related articles in our dataset	76
4.20	Iran related articles sentiments	77
4.21	North Korea related articles in our dataset	78
4.22	North Korea related articles sentiments	78
4.23	Vietnam related articles in our dataset	80
4.24	Vietnam related articles sentiments	80
5.1	Research Methodology for COVID-19 Topic Modeling and Clustering	84
5.2	Topic 5 in India's Dataset - Vaccine Development related News	87
5.3	Topic 4 in Japan's Dataset - COVID-19 Cases in Different Prefectures	89
5.4	Topic 10 in South Korea's Dataset - Celltrion's Vaccine Development related topic	92
5.5	Topic 1 in UK's Dataset - Various Hospital news during the pandemic	93
5.6	Davies Bouldin Score and Number of Clusters (Indian Topics)	95
5.7	Indian Topics Classified in Cluster No. 4	96
5.8	Scatter Plot of Indian Topics	97
5.9	Davies Bouldin Score and Number of Clusters (Japan's Topics)	98
5.10	Japan Topics Classified in Cluster No. 2	98

5.11 Scatter Plot of Japan’s Topics	99
5.12 Davies Bouldin Score and Number of Clusters (South Korea’s Topics)	100
5.13 South Korea Topics Classified in Cluster No. 7 to 11	100
5.14 Scatter Plot of South Korea’s Topics	101
5.15 Davies Bouldin Score and Number of Clusters (UK’s Topics)	102
5.16 UK Topics Classified in Cluster No. 2	102
5.17 Scatter Plot of UK’s Topics	103
5.18 Mean Coherence and Number of Topics (India)	104
5.19 Mean Coherence and Number of Topics (Japan)	106
5.20 Mean Coherence and Number of Topics (South Korea)	108
5.21 Mean Coherence and Number of Topics (UK)	111
6.1 Research Methodology for Sentiment Analysis of COVID-19 News Headlines	117
6.2 Preprocessing Steps Used in Cleaning the Headlines	118
6.3 RoBERTa Sentiment Classification Model Summary	119
6.4 Overall and Various Topic’s Sentiment in India Dataset	123
6.5 Overall and Various Topic’s Sentiment in Japan’s Dataset	124
6.6 Overall and Various Topic’s Sentiment in South Korea’s Dataset	125
6.7 Overall and Various Topic’s Sentiment in UK’s Dataset	127
6.8 Comparison of Countries COVID-19 News Headlines Sentiments	128

List of Tables

3.1	Sample Dataset	37
3.2	Document-Term Matrix (V) in Sample Dataset	38
3.3	Document-Term (W) in Sample Dataset	38
3.4	Topic-Term Matrix (H) in Sample Dataset	38
4.1	Collected articles from different countries and Newspapers . .	49
4.2	Canada Top 10 Topics	54
4.3	India Top 10 Topics	57
4.4	Japan Top 10 Topics	60
4.5	South Korea Top 10 Topics	63
4.6	UK Top 10 Topics	66
4.7	US Top 10 Topics	69
4.8	Showing accuracy in different sentiment analysis libraries . . .	69
4.9	Evaluation Matrix for different classifiers and features	71
5.1	Data Collected for COVID-19 Topic Modeling and Clustering	85
5.2	No. of Topics Discovered from Top2Vec Model	87
5.3	India's Top 10 Topics with Topic Size, Topic words, and Topic Label	88

5.4	Japan's Top 10 Topics with Topic Size, Topic words, and Topic Label	90
5.5	South Korea's Top 10 Topics with Topic Size, Topic words, and Topic Label	91
5.6	UK's Top 10 Topics with Topic Size, Topic words, and Topic Label	93
5.7	Common Topics in top ten position	94
5.8	Comparing NMF and Top2Vec Approaches	103
5.9	Top Ten Topics from NMF from Indian Dataset	105
5.10	Comparing Top Ten Topics from NMF and Top2Vec (India) .	106
5.11	Top Ten Topics from NMF from Japan's Dataset	107
5.12	Comparing Top Ten Topics from NMF and Top2Vec (Japan) .	108
5.13	Top Ten Topics from NMF from South Korea's Dataset	109
5.14	Comparing Top Ten Topics from NMF and Top2Vec (South Korea)	110
5.15	Top Ten Topics from NMF from the UK's Dataset	112
5.16	Comparing Top Ten Topics from NMF and Top2Vec (UK) . .	113
6.1	Collected COVID-19 News Headlines	117
6.2	Sample Labeled Headlines with Sentiment	119
6.3	Comparison of RoBERTa with Other Classifiers	121
6.4	Example of RoBERTa and Other Classifier's Predictions	122

Chapter 1

Introduction

1.1 Background

News media is a part of mass media along with magazines, cinema, radio, and television. It plays an important role in society. News media's role in society can be broadly categorized into three points:

- (1) **Inform** – News media inform citizens about what is happening with/in their leaders, environment, government, and the world. For example, if there is any natural disaster such as earthquake, tsunami, or flood in one country, it is through news media that the people of the world know about it. 2011 Tōhoku earthquake and tsunami in Japan or 9/11 terrorist attack in the US are two well-known examples.
- (2) **Educate** – News media not just inform people but also assist them in understand and comprehend. There are several issues or policies that are difficult for common people to understand by themselves. For example, during the current COVID-19 crisis, various governments are making several policies such as vaccine policy, travel guidelines, prevention protocols, etc., that are important for common people

to understand. News media assist people in comprehending those policies.

- (3) **Watchdog** – News media is an important accountability mechanism. It highlights issues that might not be debated in public. For example, corruption by businesses or politicians, refugees/migration crisis, human rights violation by government or authority, etc. are few cases where news media worked as a watchdog.

Scholars from various disciplines such as social sciences, communications, psychology, political science, history, and language studies are performing news analysis (content analysis is a frequently used term for this). However, it is most widely used in social science and mass communication research. It has been used broadly to understand a wide range of themes such as social change, cultural symbols, changing trends in the theoretical content of different disciplines, verification of authorship, changes in the mass media content, nature of news coverage of social issues or social problems such as atrocities against women, dowry harassment, social movements, ascertaining trends in propaganda, election issues as reflected in the mass media content, and so on. One of the most important applications of news analysis has been to study the social phenomenon [1].

However, in present times, due to the internet revolution and digitalization, a large volume of news is being generated by news media every day. This situation becomes more complicated during public health emergencies such as COVID-19. According to a study published in the Lancet, at its peak in late March 2020, in one week, more than 60,000 news articles was being written by news media from 50 countries newspapers [2]. This huge volume of data is called Big data and it would not be wrong to say that we are living in the age of Big Data. It is becoming challenging for organizations and businesses to manage the data generated by various systems, processes, and transactions. However, it is not easy to quantify what data is Big Data. For some, it can be hundreds of thousands of records, while for others, it can be petabytes of data generated by social

media. If properly organized and investigated, this large amount of data can advance our understanding of society, business, and science. In this scenario, data-driven research – an exploratory approach that analyzes data to extract scientifically interesting insights by applying analytical techniques and modes of reasoning. Fig 1.1 shows the flow of data-driven research. This data-driven research is gaining popularity over traditional hypothesis-driven also known as theory-driven approach because it can utilize Big data. Some of the advantages or contributions of data-driven research over hypothesis-driven research is its ability to extract patterns from the analysis of large data and derivation of insights from these patterns.

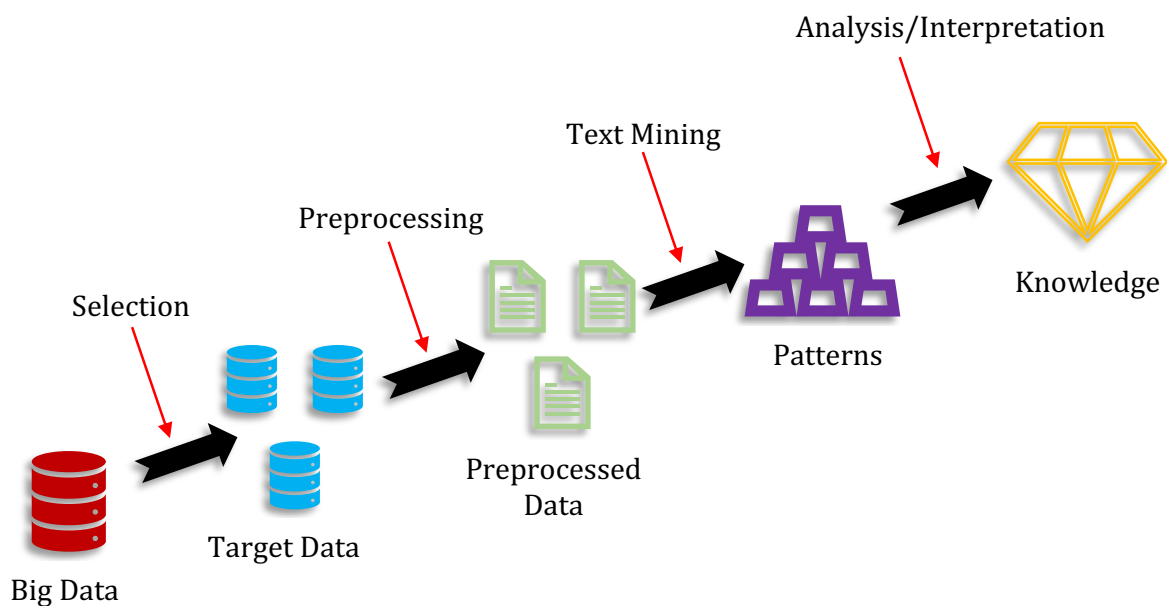


Figure 1.1: **Flow of Data-driven Research**

Out of the total generated big data, 80%¹ is unstructured and this unstructured data is typically text heavy. For example, articles, blogs, tweets, and Wikipedia pages. To perform a data driven research on textual data or natural language data, we need

¹<https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/>

a technique or techniques that can comprehend this textual data. NLP is a field in machine learning with the ability to understand, analyze, manipulate, and potentially generate human language. Because of this reason, NLP and ML techniques are indispensable when the data-driven research is applied on natural language data. The data in our research is textual data such as news headlines and articles, hence we utilized NLP and ML techniques. Through our research we will not only demonstrate the utility of data-driven research but also present the applicability and significance of various NLP and ML techniques in analyzing large natural language data.

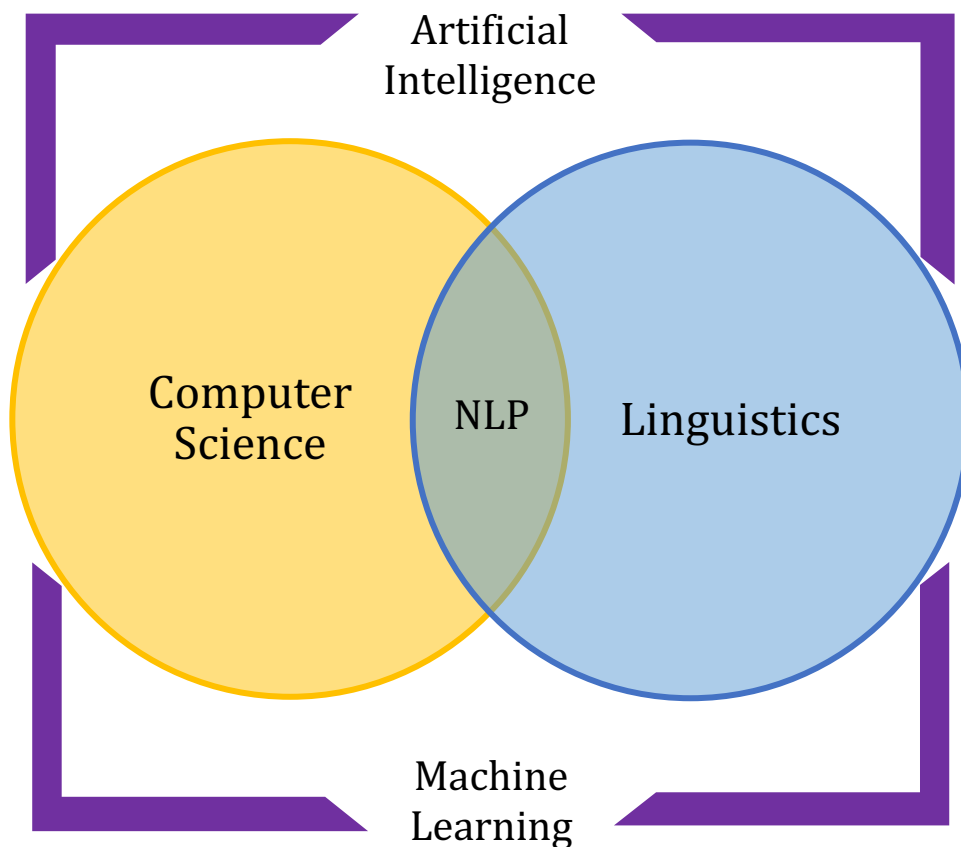


Figure 1.2: **The Field of NLP**

NLP combines techniques, algorithms, and tools to process and convert this un-

structured text data into information. Hence, the significance of NLP is twofold: 1) In the capability of interpreting human language, and 2) Making the almost impossible task of analyzing the large volume of data (text) possible.

NLP can be defined in several ways. One of NLP definitions is - A field that combines linguistics and artificial intelligence (AI) to enable computers to understand human or natural language input [3]. Fig 1.2 shows how NLP is connected to other fields. NLP has its roots in the 1950s. Alan Turing published an article titled "Computing Machinery and Intelligence," where he proposed a criterion of intelligence, a task that involves the automated interpretation and generation of natural language. This is now called the Turing Test [4]. The development of NLP can be broadly divided into three phases: 1) from the 1950s to – early 1990s was the rule-based NLP where the rules are often developed by linguists and may operate at the lexical, syntactic, or semantic level, 2) from the 1990s – 2010s NLP was the statistical NLP where the introduction of machine learning algorithms revolutionized the NLP, and 3) from 2010s – present is called neural NLP as the development of representation learning and deep neural network style machine learning methods became widespread and helped in transforming this field. This present era of NLP typically integrates Machine Learning (ML) algorithms to process text automatically. These ML algorithms are trained to learn the language patterns from any text data such as social media posts, news articles, or other types of datasets. There are supervised, semi-supervised, and unsupervised ML algorithms. Supervised ML algorithms require labeled data that can be used to train and test the model. However, unsupervised ML algorithms do not need any labeled data, the researcher needs to input data, and these ML algorithms would provide results as output. In comparison, semi-supervised ML combines a small amount of labeled data (supervised) with a large amount of unlabeled data (unsupervised) during training. Topic modeling is an unsupervised ML approach. While sentiment analysis can be performed using both supervised (Traditional NLP classifiers, BERT, RoBERTa) and unsupervised (VADER,

SentiWordNet, Textblob) ML methods.

Today, NLP has become very powerful and capable of impacting our day-to-day life.

Below are some common tasks of NLP:

- (1) **Machine Translation** – Machine translation is the task of automatically converting source text in one language to text in another language. With the development and application of neural networks, machine translation accuracy has been improved a lot. The encode-decoder recurrent neural network architecture with attention is currently the state-of-the-art on some benchmark for machine translation. This architecture is at the heart of the Google Neural Machine Translation system, or GNMT, used in Google translate [5].
- (2) **Topic Modeling** – Topic models are a type of statistical model for discovering the abstract 'topics' that occur in a collection of documents (corpus). This technique has been used widely in computer science, focusing on text mining and information retrieval.
- (3) **Sentiment Analysis** – Sentiment analysis (also known as opinion mining) uses NLP, computational linguistics, and text analysis to systematically identify, extract, quantify, and study affective state and subjective information. It is useful to identify public opinion trends in social media for marketing and to understand the general mood towards a specific issue or event.
- (4) **Document Classification or Categorization** – This task is about classifying a document into one or more classes or categories. The document can be text, image, music, etc. Spam filtering, sentiment analysis, and humor detection [6] are examples of document classification applications.

Other than the techniques mentioned above, text summarization and question answering are also frequently used applications of NLP. The text on which NLP methods

are applied can be tweeter posts, consumer feedbacks, movie reviews, news headlines or articles, research papers, or political speeches. The chosen text for this research is news headlines and articles. News is the information about current events, and this information is delivered to the general public by news media – a form of mass media. News media’s practice of – gathering information about issues and events, wrapping them into discrete news stories and circulating them through various mediums such as print (newspapers, newsmagazines), broadcast (radio, television), and recently the internet (online newspapers, news videos, news blogs, and live news streaming) – can be referred to as – newsmaking. This newsmaking is informed by a variety of cultural beliefs and ideological assumptions [7]. Thus, the product of newsmaking – news – is a social text, symbolically incorporating and recirculating those assumptions and beliefs and thereby reproducing social reality. This justifies news headlines and articles’ usage and criticality as a resource to understand various events and social phenomenon.

1.2 Focus and Scope

This research utilized news headlines and articles to comprehend two social phenomena: 1) Cybersecurity and 2) COVID-19 pandemic. Below is a short introduction to these two issues.

1.2.1 Cybersecurity

Cybersecurity is not a new phenomenon, and it exists at least since the development of the internet. However, it was only in the late 2000s (2007) when large-scale cyber-attacks came over the entire nation (Estonia) that the subject was catapulted to the center of global attention². Nowadays, when the world is ushering into the Internet of Things (IoT) era, where more and more ‘things’ are embedded with sensors, soft-

²https://www.sciencespo.fr/cei/sites/sciencespo.fr.cei/files/art_htk.pdf

ware, and other technologies for connecting and exchanging data with other devices over the internet. Whether it is transportation, medical and healthcare, smart homes and appliances, agriculture, or weapons, everything is connected with the internet. According to Business Insider, “the Internet of Things Report 2019,” there will be 64 billion IoT devices by 2025³. Simultaneously, cybercrime, propaganda, surveillance, sabotage, and unrestricted exploitation of personal data threaten digital trust and security. Moreover, the borderless nature of cyberspace forever keeps all the countries on their toes. The global information security market is forecast to reach \$170.4 billion by 2022, and in 2019, the US Cybersecurity budget was \$15 billion^{4,5}. These figures indicate the ever-increasing necessity of Cybersecurity for both businesses and countries. However, increased spending does not ensure a safe and secure cyberspace. It is critical to understand numerous dimensions (organizational, political, financial, and human) of Cybersecurity to protect nations and businesses from cyber-attacks. Fredrick Chang, former director of Research at the National Security Agency in the US, beautifully described this multidisciplinary aspect of Cybersecurity:

“A science of Cybersecurity offers many opportunities for advances based on a multidisciplinary approach, because, after all, Cybersecurity is fundamentally about an adversarial engagement. Humans must defend machines that are attacked by other humans using machines. So, in addition to the critical traditional fields of computer science, electrical engineering, and mathematics, perspectives from other fields are needed⁶.”

Media and academia each play a crucial role in different aspects of Cybersecurity. Media develops widespread Cybersecurity awareness while academic-stricken activities are prerequisites of both technological advancement and investment. Thus, the two dis-

³<https://www.businessinsider.com/internet-of-things-report?IR=T>

⁴<https://cybersecurityventures.com/cybersecurity-market-report/>

⁵https://www.whitehouse.gov/wp-content/uploads/2018/02/ap_21_cyber_security-fy2019.pdf

⁶<https://www.nsa.gov/Portals/70/documents/resources/everyone/digital-media-center/publications/the-next-wave/TNW-19-4.pdf>

ciplines are the main sources of decision making and strategy planning in Cybersecurity fields. News agencies inform societies about Cybersecurity threats, vulnerabilities, and preventive solutions. From a business standpoint, the consequences of a data breach are heavily alleviated by the presence of a free press, even though the awareness about a given businesses' vulnerabilities could have negative ramifications [8].

Newspapers tend to report and focus on the issues that are contemporary and significant for nations and their citizens. We utilized several NLP techniques to understand the prominent cybersecurity-related events and issues that the nations are talking about during the selected period (from April 2018 to March 2019). This case study would also help us comprehend the similarity and difference between various nation's cybersecurity-related news and whether the strategic partnership among nations impacts the reporting of these events.

The research questions for which we would try to find answer are below:

- (1) **Research Question 1 [RQ 1]** - What are the most trending topics in the field of cybersecurity in the six countries we selected for study during 2018-19?
- (2) **Research Question 2 [RQ 2]** - Which topics are common in all countries?
- (3) **Research Question 3 [RQ 3]** - How a country's international relations impacted cybersecurity coverage in these countries?
- (4) **Research Question 4 [RQ 4]** - What is the overall sentiment (positive or negative) of cybersecurity news?
- (5) **Research Question 5 [RQ 5]** - How news media from different country covered (negatively or positively) the same issue?

1.2.2 The COVID-19 Pandemic

In December 2019, the outbreak of pneumonia with an unknown cause was discovered in Wuhan, Hubei Province, China. The virus that caused this outbreak was later found and named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). In February 2020, the World Health Organization (WHO) called the disease caused by SARS-CoV-2 “COVID-19” [9]. Since the first reported death on January 09, 2020, in Wuhan, China⁷, this virus has infected more than 117.52 million people and taken the lives of more than 2.6 million people as of March 10th, 2021⁸. The contagious disease caused by severe acute respiratory syndrome (SARS-CoV-2) is called Coronavirus disease 2019 (COVID-19). This pandemic is the biggest that the world is coping with since the Spanish flu of 1918. To stop this new and highly contagious COVID-19 disease, nations resort to lockdowns and emergency measures. The impact of these measures was felt on everyday life (social), the countries’ economy as most of the economic activities were closed (economic), and even political affairs of nations were influenced (political). This makes the COVID-19 pandemic not only a healthcare crisis but a social, economic, and political crisis also. Though many vaccines have been approved and many nations started conducting large-scale vaccination drives, it would still be too early to say that the COVID-19 pandemic is under control. F. Krammer in “SARS-CoV-2 vaccines in development” [10] and G. Forni and A. Mantovani in “COVID-19 vaccines: where we stand and challenges ahead” have discussed in detail the vaccines development process, different kind of vaccines, and the frontrunner vaccines candidates [11].

As the world is in the grip of the COVID-19 pandemic, Newspapers worldwide are extensively reporting about COVID-19 related news. This makes newspapers an incredible source to comprehend the social, economic, and political reality vis-a-vis this deadly pandemic in a particular society/country. Further, the practice of newsmaking is

⁷<https://www.nytimes.com/2020/01/23/world/asia/china-coronavirus.html>

⁸<https://coronavirus.jhu.edu/map.html>

informed by a variety of cultural beliefs and ideological assumptions. Thus, the product of newsmaking - news - is a social text, symbolically incorporating and recirculating those assumptions and beliefs and thereby reproducing social reality. Millions around the world lost their jobs due to the lockdown. India's complete lockdown led to the migrant crisis of a monumental scale [12], many top economies of the world clocked negative GDP growth⁹. More than 330 companies filed for bankruptcies in the US alone last year and blamed COVID-19 in part for their demise¹⁰.

Our COVID-19 case study would employ NLP techniques to analyze COVID-19 related news articles, which we collected from 1st January to 1st December 2020. This research would try to find and bring forward some interesting patterns from COVID-19 news of various nations. Further, this research can also provide a better understanding of COVID-19 related issues and events that affected a particular country.

The research questions for which we would try to find answer are below:

- (1) **Research Question 1 [RQ 1]** - What are the most trending topics in COVID-19 news India, Japan, South Korea, and UK?
- (2) **Research Question 2 [RQ 2]** - Which topics are common in all countries?
- (3) **Research Question 3 [RQ 3]** - How does a country's international relations affected COVID-19 reporting?
- (4) **Research Question 4 [RQ 4]** - What is the COVID-19 news sentiment (Negative or Positive)?
- (5) **Research Question 5 [RQ 5]** - Is there a correlation between a country's COVID-19 situation and news sentiment?
- (6) **Research Question 6 [RQ 6]** - What are the sentiments of common topics?

⁹<https://www.businesstoday.in/current/economy-politics/which-top-economies-have-suffered-worst-gdp-fall-due-to-covid-19/story/414683.html>

¹⁰<https://www.bloomberg.com/graphics/2020-us-bankruptcies-coronavirus/>

1.3 Research Objectives

This is an analytical study where NLP methods are employed to explain two critical issues – Cybersecurity and the COVID-19 pandemic. Hence, this study’s research objectives can be divided into two - 1) Showing the application of NLP methods and 2) Understanding and analyzing the Cybersecurity and the COVID-19 pandemic.

1.3.1 Exhibit the application of NLP

As mentioned before, NLP provides machines or computers the ability to understand and interpret human language. This capability becomes more significant when humans are generating millions of terabytes of data every day¹¹. When analyzed, this data can bring forward some crucial patterns and interesting facts that humans would not have discovered otherwise. Our research applies two popular NLP methods, namely topic modeling and sentiment analysis on the news related to two critical social phenomena, to show how the utilization of NLP can improve our understanding of these issues. Further, our research is also a practical example of how NLP makes the tedious task of analyzing an enormous amount of text data (4159 cybersecurity news articles and 102,278 COVID-19 news articles) which would have taken a large number of human hours and resources, can be accomplished by a single individual.

1.3.2 Understanding and analyzing Cybersecurity and COVID-19 pandemic News

Nobody can argue about the value or usefulness of understanding the Cybersecurity and the COVID-19 pandemic. There are various methods or approaches which researchers can apply to improve our understanding of these issues. For example, security experts and researchers can improve systems’ security, policy experts can analyze cybersecurity

¹¹<https://medium.com/@amoghvs/how-much-data-is-generated-every-day-b94af8bcef4b>

policies, and law experts can interpret cyber laws to provide better understanding. Similarly, research on the COVID-19 pandemic also has multiple facets. For example, scientists are trying to decipher the SARS-CoV-19, medical researchers are working on finding a cure, social scientists are trying to understand the impact of this pandemic on social life, and economists are talking about the effect of the COVID-19 pandemic on the economy and individual's economic well-being.

In this research, an attempt is made to present a completely different perspective to analyze these two issues. Almost all the facets, whether government policies, social life, economic situation, and scientific development related to a particular topic or issue, are present in the news. This is also the case for both Cybersecurity and COVID-19 pandemic. This research tried to tap this abundance of information with the help of NLP and bring forward interesting themes, patterns, and trends that are otherwise overlooked or deemed insignificant.

1.4 Challenges

In this research, we have utilized two popular NLP methods: 1) topic modeling and 2) sentiment analysis. The first challenge is related to topic modeling. In topic modeling, the specification of optimal number of topics is required up front to produce the best and coherent topic model. However, the optimal number of topics is often unknown specially in a big corpus that consist large documents such as news articles. To solve this challenge, we used Topic Coherence – Word2Vec (TC-W2V) metric to discover how many topics are present in corpus and then produced a topic model. TC-W2V metric measures the coherence between words assigned to a topic. The number of topics (k) with highest mean coherence is used to train a final NMF model. The first contribution is focused on resolving this challenge.

The second challenge is related to supervised sentiment analysis. For the supervised

sentiment analysis method, getting an appropriate labeled data for training and testing the model is prerequisite. As COVID-19 is the recent issue, the appropriate labeled dataset was unavailable. To solve this challenge, we created our own labeled dataset for COVID-19 case study by way of applying a novel method. The second contribution is motivated to solve this challenge.

1.5 Limitations

The limitations related to specific experiments and methods would be discussed separately in Chapters 4, 5, and 6. Here, we would mention two limitations that are part of this complete study.

Data is the backbone of every research, and collecting relevant data for particular research can be challenging. For this research, news headlines and articles from various newspaper websites were collected using the web-scraping technique. Some of the most widely circulated English language newspapers from Canada, India, Japan, South Korea, the UK, and the US were selected. Though the attempt to collect news articles from other countries such as China, Australia, and Brazil has also been made, due to the paywall restrictions, unavailability of newspaper's official website, and web-scraping proof websites, we were unable to acquire news articles. Because of that, our study is restricted to the six countries mentioned above only (four in the case of the COVID-19 case study as we could not collect from Canada and the US). This research has only focused on English language newspapers. For countries such as Japan and South Korea, where English is not an official language, English-language newspapers focus on the foreign nationals residing in those countries. Hence, it can be said that these newspapers may not provide a clear picture of a particular issue. Further, because of language limitations (only English language newspapers), we could not include some of the more prominent countries such as Germany, France, China, and Russia. The

inclusion of these countries can increase the scope of this research.

1.6 Main Findings and Contributions

1.6.1 Findings

This analytical research is an example of interdisciplinary study where advanced computational methods and tools (NLP) are used to analyze two critical social issues. Our investigation provides a different perspective to understand Cybersecurity and COVID-19 pandemic. The analysis and findings contribute to improving the overall understanding of these issues.

Experiment and results related to case study one (Cybersecurity-related research) are discussed in chapter 4. The main findings of this case study are:

- (1) A Cross-national Study of Cybersecurity-related news highlights that the US is the biggest influencer in Cybersecurity as all nation's newspapers extensively reported the US-related events and topics.
- (2) The geopolitical and strategic interests of nations can be understood from the analysis of a nation's news. For example, the presence of North Korea-related news in Japan's cybersecurity news.

As COVID-19 pandemic-related research is a large case study, we divided it into two chapters based on two methods. Implementation of topic modeling and clustering on COVID-19 news is discussed in chapter 5, and the main findings of this chapter are:

- (1) This chapter also highlighted that the news that is being reported in a specific country is influenced by that nation's partnership or relationship with other nations. For example, the presence of Australia and New Zealand-related news in UK news media points towards a close historical and cultural relationship among these nations.

- (2) The comparative analysis of the top ten topics for all the studied countries discovered that the US, Economic, Education, and Sports are the common issues discussed in these countries. Interestingly, these are some of the most severely impacted areas also.

Sentiment classification and analysis of COVID-19 news are performed in chapter 6. The key findings of this chapter are:

- (1) Using the state-of-the-art RoBERTa model, our sentiment classification model produced better results than the traditional Bag-of-words-based classifiers.
- (2) Our research showed that the UK (the worst affected country among the studied countries) has the highest percentage of negative news, and South Korea (the most successful country among the four nations that we studied) has the lowest percentage of negative news. This finding points towards the possible correlation between the negative news and the degree of a country's affectedness.

1.6.2 Contributions

In this research we have presented three key contributions. First contribution of this research is the utilization of TC-W2V metric to discover the optimal number of topics. TC-W2V means Topic Coherence – Word2Vec. For this first we train the word2vec model on our corpus, which would organize the words in an n-dimensional space where semantically similar words are close to each other. The TC-W2V for a topic will be the average similarity between all pairs of the top n- words describing the topic. We trained the NMF model for different values of the topic (k) and calculated the mean TC-W2V across all the topics. The k with the highest mean TC-W2V is used to train the final NMF model. As we applied NMF topic modeling in both cybersecurity and COVID-19 case studies, we first utilized TC-W2V to find the optimal number of topics than based on the results of that we produced final NMF topic models.

The second contribution of this research is to create COVID-19 news related labeled dataset for supervised sentiment analysis method. For this in the first step we used the three most popular python-based libraries 1) VADER, 2) Textblob, and 3) SentiWordNet on 102,124 COVID-19 headlines. In second step, we filter and keep the headlines that all three libraries categorized as either positive or negative. After the second step, we were left with only around 15% of the total headlines. In the third step, we manually confirmed all the headlines that we collected from the second step. Lastly, we used oversampling to balance the labeled data. Oversampling can lead to overfitting. One way to find whether the model is overfitting is to check the difference between training accuracy and validation accuracy. The training accuracy and validation accuracy of the 3rd (last) epoch of our classification model was 95.08% and 90% respectively. As the difference between training and validation accuracy is not very high, it can be said that our model is not overfitting. In this way we created a labeled dataset of 10,727 COVID-19 news headlines which has 5369 positive and 5358 negative headlines.

The third contribution of this research is the proposed model that combines topic modeling and sentiment analysis for the COVID-19 news dataset. While topic modeling and sentiment analysis are some of the most common NLP approaches, there has been very few research that combine both the approaches. The significance of combining both approaches is that by first applying topic modeling on large dataset, it become possible to find out several key issues and themes. For example, our COVID-19 dataset was consisting of more than 100,000 news articles and from this dataset we could discovered several critical common topics such education, economy, and sports that were present in all countries. In the next step of sentiment analysis, we used the results from topic modeling to investigate the sentiments associated with these common topics along with the overall dataset. Finding critical and common topics would not be possible with only sentiment analysis method. Hence, this combination of both approaches is very useful in finding sentiments of critical issues present in a large dataset.

Lastly, this research that involves both cybersecurity and COVID-19 case studies, we have performed sentiment classification on news media. Sentiment classification of news media is an underdeveloped or less explored area as most of the sentiment classification research are performed on social media data. Our research is a welcome addition that further the advancement of this area of research.

Other than these findings and contributions, the COVID-19 case study also contributed to the development “COVID-19 news articles” public dataset [13]. This dataset can be utilized by the researcher to further explore and comprehend the COVID-19 pandemic.

1.7 Thesis Overview

Chapter 1 – Introduction – has discussed the background, focus and scope, research objectives, limitations, findings, and contributions. The rest of this thesis is divided into six chapters.

Chapter 2 – Literature Survey – As this thesis consists of two different case studies, the literature survey is also divided into two sections. Section 2.1. would discuss Cybersecurity related existing literature that influenced this case study. Section 2.2. would talk about various existing COVID-19 pandemic-related research. This section can have two categories, one where COVID-19 research is surveyed from data-source perspective and second, where research approach is the basis of categorising the literature. Research approach based literature can be further divided into three subsections. For example, research that includes 1) topic modeling, 2) sentiment analysis and 3) both topic modeling and sentiment analysis.

Chapter 3 – Main Methods – This chapter reviews the methods or approaches common in both case studies in detail. Three main components are common; hence this chapter is divided into three subsections. Section 3.1. would examine the text

preprocessing method. Then section 3.2. would talk about topic modeling and lastly, section 3.3. would describe the sentiment analysis method in detail.

Chapter 4 – Understanding the Cybersecurity News by Utilizing Topic Modeling and Sentiment Analysis Methods – This chapter is the Cybersecurity case study’s representative chapter. We start the discussion with the introduction – section 4.1. – this talks about motivation of this case study. Then, the research methodology of this research is introduced in section 4.2. Section 4.3. is about experiments and results, and this is further divided into two subsections: section 4.3.1. topic modeling and 4.3.2. sentiment analysis. Then the limitations – section 4.4. – of this case study is presented.

Chapter 5 – Topic Modeling and Clustering Approach to Analyze COVID-19 News – This chapter is part of the COVID-19 case study where topic modeling and clustering methods are used to analyze COVID-19 news. The first section – 5.1. would discuss motivation behind COVID-19 case study. Next section – 5.2. present research methodology specific to this chapter. Section 5.3. would discuss in detail the experiment and results, which is further divided into two subsections – 5.3.1. – topic modeling and 5.3.2. – clustering. Limitations is discussed in section 5.4.

Chapter 6 – Sentiment Analysis of COVID-19 News – This chapter is the second part of the COVID-19 case study. This chapter starts with the vitality of sentiment analysis research in the COVID-19 context in section 6.1. The research methodology of this chapter is discussed in section 6.2. Section 6.3. would be about experiments and results of sentiment analysis. This section is further divided into sections 6.3.1 – sentiment classification and section 6.3.2 – sentiment analysis. Limitations specific to this research is being discussed in section 6.4.

Chapter 7 – Conclusion – This is the last chapter of this thesis. This chapter would bring together all the findings and contributions of both case studies. Further, we would also talk about future work in this chapter.

Chapter 2

Literature Survey

As we have two case studies, the related research literature is also divided into section 2.1. where research related to our case study of Cybersecurity is discussed, and section 2.2. would be about the COVID-19 case study. Based on the methods, the literature survey for this case study is further divided into three subcategories: 1) topic modeling, 2) sentiment analysis and 3) topic modeling and sentiment analysis.

2.1 Literature Survey (Case Study – Cybersecurity)

As the application of NLP is getting wider every day, the various text data related to the field of Cybersecurity are also being used to analyze the trends, discover patterns and themes. T. M. Georgescu performed cognitive analysis for Cybersecurity-related documents in “Natural Language Processing Model for Automatic Analysis of Cybersecurity-Related Documents” [14]. In this research, a domain ontology was developed using a two-step approach: 1) the symmetry stage and 2) the machine adjustment. A web application to integrates the model’s core components were also developed. C.L. Jones et

al., in their short paper “Towards a Relation Extraction Framework for Cyber-Security Concepts,” utilized semi-supervised ML and implemented a bootstrapping algorithm for extracting security entities and their relationships from a corpus of 62 news articles and blogs from a variety of security-related websites [15].

Similarly, H. Gasmi et al. also extracted Cybersecurity entities and the relationships between them from online textual resources. However, they utilized neural networks models, specifically the long short-term memory (LSTM), to accomplish this task. They presented this research in “Information Extraction of Cybersecurity Concepts: An LSTM Approach” [16]. They argued that previous feature-based models are time-consuming and labor-intensive, while their LSTM based approach achieved competitive performance with less feature-engineering work. O. Mendsaikhan et al. proposed an autonomous system for extracting cyber threat information from publicly available information sources in “Identification of Cybersecurity Specific Content Using the Doc2Vec Language Model” [17]. They collected more than 1,200,000 Cybersecurity-related texts from both informal conversations (Reddit discussions, Stack Exchange discussions, Hacker news comments) and formal news/bulletin (National Vulnerability Database, Slashdot news, Security news outlet RSS feeds).

M. R. Alagheband et al. utilized LDA topic modeling and the time-based gap analysis methods in “Time-based Gap Analysis of Cybersecurity Trends in Academic and Digital Media” [18]. To identify Cybersecurity trends and propose a conceptual framework to identify Cybersecurity topics of social interests, they used 3556 academic papers from top 10 highly reputable Cybersecurity academic conferences and 4163 news articles from the New York Times. Their experiment and analysis identified both convergences and divergences in two corpora which suggested a strong time-based correlation between these resources. F. Kolini and L. Janczewski’s paper “Clustering and Topic Modeling: A New Approach for Analysis of National Cyber security Strategies” [19] showed how national Cybersecurity strategies of various countries could be used as a resource. The

analysis of 60 strategies they point out stronger similarities between the strategies of the EU or NATO member countries. They argued that this kind of automated technique for analysis and understanding of text documents could be used by policymakers and governments during the development and reviewing of national Cybersecurity strategies and policies.

“Twitter Sentiment Analysis: An Examination of Cybersecurity Attitudes and Behavior” by B. Gupta et al. examines the Cybersecurity attitudes and actual behavior over time by using data from Twitter [20]. This research first performed sentiment analysis on 15000 Cybersecurity-related tweets to examine the Cybersecurity attitude. This step was followed by the qualitative text analysis to identify Cybersecurity behavior. K. Al-Rowaily et al. developed a “Bilingual Sentiment Analysis Lexicon - BiSAL” and presented it in the paper titled “BiSAL – A bilingual sentiment analysis lexicon to analyze Dark Web forums for cyber security” [21]. BiSAL consists of a Sentiment Lexicon for English (SentiLEN) and a Sentiment Lexicon for Arabic (SentiLAR).

The above literature survey shows the application of NLP in the field of Cybersecurity. The researches from [14] to [17] are focused on entities and relation extractions. In contrast, topic modeling [18], [19] and sentiment analysis [20], [21] are used in other researches. As our research combines both topic modeling and sentiment analysis to analyze cross-cultural Cybersecurity-related news articles, it can be seen as a combination of research from [18] to [21]. As per this researcher’s knowledge, this combination is performed the first time for Cybersecurity news articles. From this viewpoint, our Cybersecurity case study can be seen as a welcome addition to existing research in this area.

2.2 Literature Survey (Case Study – The COVID-19 pandemic)

2.2.1 Literature survey from Data Source Perspective

If we look from the perspective of data source, the literature survey for COVID-19 related research can be divided into two categories. One category of research that include news media as source and the other one where social media (Twitter, Facebook, and Reddit), press releases, and scientific research are used as source.

For this, we surveyed the first ten pages of google scholars (100 research papers) with “topic modeling and sentiment analysis of COVID-19 news”¹ as search query². We found out that out of these 100 research, only eight papers (including ours) belong to first category and rest 92 research papers came under the second category. As our research is belong in news media category, we only focus on news media research papers. Out of the seven research papers, one research paper was in Korean language [22] and the one research paper is focused on finding stock market sentiments during COVID-19 [23]. We did not consider these two research in our discussion of literature of news media category. The remaining five research papers we are going to discuss in little detail.

Krawczyk, K. et al. in “Quantifying the online news media coverage of the COVID-19 pandemic” [24] performed an extensive study where they collected 26 million news articles from 11 country’s 172 online news sources (6 languages). For analysis, they used topic detection and VADER sentiment analysis tool. They showed that COVID-19 accounted for 25% of all front page online news articles during January and October

¹As news research is published, the results on these pages are also updated, so it is possible to find newly published research that were not there at the time when we surveyed these pages

²https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5q=topic+modeling+and+sentiment+analysis+of+covid-19+news&btnG=

2020. Further, their results also showed that 16% of COVID-19 news articles can be classified as highly negative, citing issues such as death, crisis, and fear. This research is closest to our research as we have several common features such as news media, analysis of data from multiple countries, and usage of both topic modeling and sentiment analysis. In terms of data, the research by Krawczyk, K. et. al is better than our research as we only collected little more than 100,000 news articles from four countries and focused only on English language. However, in terms of methods, our research is better as we used neural network based topic modeling method – Top2Vec, and state-of-the-art RoBERTa for sentiment analysis. Further, our research also created very first labeled dataset of COVID-19 news headlines for sentiment analysis.

The other four research that used news media as source can be clubbed together as all of them focus only on single country. Poirier, W. et al. in “(Un) covering the COVID-19 pandemic: Framing analysis of the crisis in Canada” [25] collected news articles from 12 news media outlets from Canada in both French and English language. They found a noticeable difference in the use of the Health Crisis, Social Impact, and Chinese Outbreak between francophone and anglophone media. Similarly, Jo, W. and Chang, D in “Political consequences of COVID-19 and media framing in South Korea” [26] investigated COVID-19 news in South Korean media (11 national newspapers in Korea) and used structural topic modeling approach. They found that COVID-19 impact on everyday life, economy, sports, and international relations are highly discussed topics in Korean news media. South Korea is one of the country in our research, and our topic modeling results also suggested similar results as this research.

Liu, Q. et al. in “Health communication through news media during the early stage of the COVID-19 outbreak in China: digital topic modeling approach” [27] collected 7791 news articles from China and used LDA topic modeling approach. They found that prevention and control procedures, medical treatment and research, and global or local social and economic influences were top three topics that were discussed in Chinese

news media between January 1 and February 20, 2020. De Melo, T. and Figueiredo, C.M. in “Comparing News articles and tweets about COVID-19 in Brazil: sentiment analysis and topic modeling approach” [28] collected Portuguese language news articles (18,413) and tweets (1.5 million) from Brazil and used LDA topic modeling and VADER sentiment analysis. They compared both news media and social media and found that social media is more negative than news media.

2.2.2 Literature Survey based on Research Methods

Other than data source, COVID-19 research literature can also be surveyed from the perspective of research approach/methods. This category of research can be divided into three: 1) research where only topic modeling methods are used, 2) research where sentiment analysis methods are used, and 3) research where both topic modeling and sentiment analysis methods are used. We are going to discuss each category in detail.

2.2.2.1 Topic Modeling

COVID-19 disease is only a year old, and the world is still struggling to control it. However, natural language processing (NLP) based research on analyzing COVID-19 related material such as scientific articles, social media posts, and news is coming up quickly. Y. Bai et al. presented a topic evolution analysis of COVID-19 news articles from Canada in “Topic Evolution Analysis of COVID-19 News Articles” [29]. In this research, a dynamic topic analysis system to monitor the evolution of the large-scale text data topic is being developed. For this, they expanded the Dynamic Topic Model (DTM) with two modules: data sparsity computing and topic number selecting. Q. Liu et al. used a digital topic modeling approach to analyze news media during the COVID-19 outbreak in China in “Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China: Digital Topic Modeling Approach” [27]. Their research showed that prevention and control procedures, medical treatment and

research, and global or local social and economic influences are top three most popular themes from the collected news reports. E. De Santis et al. developed a system for detecting relevant topics from tweets in a paper titled “An Infoveillance System for Detecting and Tracking Relevant Topics From Italian Tweets During the COVID-19 Event” [30]. This study demonstrated the effectiveness of the topic tracking system in capturing the main socio-political events that occurred during their period of study in Italy. S. Noor et al. in “Analysis of public reactions to the novel Coronavirus (COVID-19) outbreak on Twitter,” performed thematic analysis to examine people’s reaction during the COVID-19 outbreak and also used sequential pattern mining techniques to find frequent words/patterns and their relationship in tweets [31].

2.2.2.2 Sentiment Analysis

J. Samuel et al. demonstrated insights into the progress of fear-sentiment as the COVID-19 approached peak levels in the United States in “Covid-19 public sentiment insights and machine learning for tweets classification” [32]. A. S. Imran et al. in “Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets” [33] analyzed tweets from six countries for polarity and emotion detection using the deep learning method. This research showed that despite being geographically close, countries reacted differently to one another. M. Huang et al. in “Lexicon-Based Sentiment Convolutional Neural Networks for Online Review Analysis” combined sentiment and contextual information from sentences by strengthening the connection between features of both sentences and their sentiment words [34]. S. Boon-Itt, and Y. Skunkan, in a paper titled “Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study” investigated twitter posts using sentiment analysis and topic modeling approach to find out the public perception [35]. G. Barkur et al. in “Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India” showed sentiments of Indians after the an-

nouncements were made [36]. They extracted two prominent hashtags: IndiaLockdown and IndiafightsCorona from March 25th to March 28th, 2020.

2.2.2.3 Topic Modeling and Sentiment Analysis

All the works cited from [29] to [36] either performed topic modeling or sentiment analysis on COVID-19 data. However, very few research has combined both topic modeling and sentiment analysis to investigate COVID-19 data. S. Das and A. Dutta characterized public emotions and sentiments in the COVID-19 environment in India by analyzing tweets in “Characterizing public emotions and sentiments in COVID-19 environment: A case study of India” [37]. First, the researchers performed sentiment analysis and then based on the sentiment polarity they divided the corpus into positive and negative. Then they utilized the topic modeling method to find out the positive and negative topics. R. Chandrasekaran et al. in “Topics, Trends, and Sentiments of Tweets About the COVID-19 Pandemic: Temporal Infoveillance Study” discussed topics, trends, and sentiments of tweets about the COVID-19 pandemic using tools such as LDA and VADER [38]. Their experiment yielded that the impact of COVID-19 on the economy and markets, spread and growth in cases, treatment and recovery, impact on the health care sector, and governments response are some of the top themes. Also, the topics of spread and growth of cases, symptoms, racism, source of the outbreak were found to be negative. J. Xue et al. analyzed 1.9 million coronavirus related tweets in “Public discourse and sentiment during the COVID-19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter” [39]. They found out that updates about confirmed cases, COVID-19 related death, cases outside China are some of the top themes. Further, sentiment analysis showed that fear for the unknown nature of the coronavirus is dominant in all the topics. Lastly, R. Xie et al. research “Exploring Public Response to COVID-19 on Weibo with LDA Topic Modeling and Sentiment Analysis” explored public response to COVID-19 on the Chinese microblogging site

Weibo [40]. They found out that support for frontline warriors, encourage each other about spiritually, preventive measures, concerns about economic and life restoration are some widely discussed topics. Further, sentiment analysis indicated that country media, influential individuals, and self-media together contribute to the information spread of positive sentiment.

Our research can be categorized along with the researches which combine topic modeling and sentiment classification methods to analyze COVID-19 data. However, the nature of our COVID-19 (news articles) and methods (top2vec for topic modeling and RoBERTa for sentiment classification) makes this research a welcome addition in comprehending the cross-cultural COVID-19 news. According to the authors' best knowledge, ours is the first study that analyzed COVID-19 news articles by applying topic modeling and sentiment analysis methods.

Chapter 3

Main Methods

This chapter presents in detail the main methods or techniques that are utilized in this thesis. This chapter aims to familiarize the readers with the background, relevance, and application of these techniques. We have applied text preprocessing, topic modeling, clustering, and sentiment analysis in Chapters 4, 5, and 6 either in combination or separately.

3.1 Text Preprocessing

As mentioned before, textual data is unstructured. The meaning of text data being unstructured is that it is not well-formatted and standardized. Text preprocessing is an umbrella term used to combine techniques that convert raw text into well-defined sequences of linguistic components that have standard structure and notation. Text preprocessing is also known as 'Text wrangling' or 'normalization.' The resultant cleaned and standardized textual data after text preprocessing can be consumed by other NLP and intelligent systems powered by ML and deep learning (DL). Raw text may contain components such as HTML tags, stopwords, etc., that might impact the analysis. Text preprocessing techniques varies according to the domain. As Twitter has limitations

on the number of characters in one post, people use a language that is not 'standard.' Some examples of social media language are LOL (laughing out loud), ROLF (rolling on the floor), Gr8 (great), OMG (Oh my god), etc. People also use emoji extensively in Twitter posts which might be useful or not. On the other hand, news reports are comparatively more structured as it is written formally. Hence, Twitter posts require an altogether different combination of text preprocessing steps than news articles or headlines. Below we present the techniques that we used in our research in detail. Though this list includes some of the most common preprocessing techniques, it is not a complete list.

- **Contraction** – Contractions are shortened versions of words or syllables. These words exist in both spoken and written form. For example, 'is not' to 'isn't', 'will not' to 'won't', 'cannot' to 'can't'. Usually, contractions are used extensively in informal communication (social media) but avoided in formal writing (news reports). These words can pose a problem for the NLP task, so it is important to expand contractions. As mentioned, it is very rare to have contractions in news articles; however, to be sure, we have performed this step in our research. We used Python file `contractions.py` for this step [41].
- **Remove HTML Tags** – Often, web scraping or screen scraping techniques can retrieve data from web pages, online repositories, and blogs. This data contains much noise in the form of HTML tags, Iframe tags, and JavaScript, and normally these tags do not add value to the analysis. Hence, it is better to remove these tags. Our research data are newspaper websites that heavily contain HTML tags, and because of this reason, the second step of our preprocessing pipeline is the removal of HTML tags.
- **Tokenization** – There are two types of tokenization – sentence tokenization and word tokenization. In this research, tokenization means word tokenization.

Word tokenization is a process of segmenting or splitting sentences into their constituent words. As a sentence is a collection of words, and with tokenization, we split a sentence into a list of words that can be used to reconstruct the sentence. This is an important step in preprocessing where stemming and lemmatization (next step) work on each word based on its respective stem and lemma. Natural Language Toolkit package (NLTK) [42] provides various useful tokenizers such as `word_tokenizer`, `TreebankWordTokenizer`, `RegexpTokenizer`, and `TokTokTokenizer`. We used `TokTokTokenizer` in our research.

- **Lemmatization** – Lemmatization is a process of removing word affixes to get a base form of the word – root word (lemma). Unlike root stem (output of stemming process), which may not always be a lexicographically correct word, lemma will always be present in the dictionary. For example, the lemma of 'running' is 'run,' 'saddest' is 'sad.' NLTK package has a robust lemmatization module `WordNetLemmatizer` where it uses WordNet [43]. For the lemmatization task in this research, we used this lemmatizer.
- **Remove Stopwords** – Stopwords are the words that have little or no significance and are usually removed from the text during preprocessing. After removing the stopwords, the remaining words are the words that have maximum significance and context. Words like 'a', 'the', 'an', 'and', 'of', 'to', and so on are stopwords. There is no exhaustive list of stopwords, and they differ from domain to domain. Further, each language has its own sets of stopwords. NLTK provides a standard English language stopwords list, and we used this in our research. We can see all the stopwords in NLTK's vocabulary by using `nlk.corpus.stopwords.words('english')`.
- **Remove Special Characters** – Special characters and symbols are non-alphanumeric characters or numeric characters present in the unstructured text and add noise.

Regular expressions (regexes) are used in removing these special characters.

- **Remove Accented Characters** – Some English language terms such as café, naïve, and other similar words have letters with diacritical marks. Most of these words are loanwords from French, and others are from German, Portuguese, Spanish, or other languages. For research that utilizes the English language, it is important to convert and standardized into ASCII characters.
- **Lowercase** – As the name suggests, this step converts all the tokens (words) into lowercase. This step brings uniformity to the text.

Figure 3.1 shows the text preprocessing steps (pipeline) used in this research.

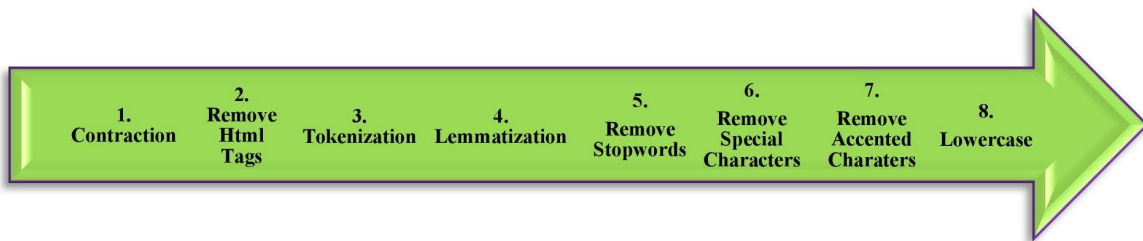


Figure 3.1: Text Preprocessing Steps

Below is the sample article, which shows the article in its original form and after preprocessing.

Sample Article

“DAKAR - Senegal on Tuesday inaugurated a cybersecurity school to strengthen West Africa’s defenses against computer hackers and use of the internet for terror funding and propaganda.’, ‘Senegalese Foreign Minister Sidiki Kaba and French counterpart Jean-Yves Le Drian gave the ceremonial start to the National Cyber-Security School (ENVR) on the sidelines of an annual regional security conference in Dakar.’ ‘The school will provide training in combating cybercrime for the security services, judiciary and private enterprises.’, ‘Backed by France, it will have a “regional vocational

role” in helping other countries in West Africa, French officials said.’, ’The ENVR, which was proposed at last year’s security conference, will initially be based in Dakar at the National School of Administration (ENA) before moving to Diamnadio, a new town being built around 30 km(20 miles) from the capital.”

Sample Article After Preprocessing

“dakar senegal tuesday inaugurate cybersecurity school strengthen west africa defenses computer hackers use internet terror fund propaganda senegalese foreign minister sidiki kaba french counterpart jeanyves le drian give ceremonial start national cybersecurity school envr sideline annual regional security conference dakar n school provide train combat cybercrime security service judiciary private enterprises back france regional vocational role help countries west africa french officials say envr propose last year security conference initially base dakar national school administration ena move diamnadio new town build around miles capital”

3.2 Topic Modeling

The ability to organize, search and summarize a large volume of text is a universal problem in NLP. As a text-mining tool, topic modeling is a type of statistical model for discovering the abstract ‘topics’ that occur in a collection of documents (corpus). Topic modeling is one of the most frequently used text mining tools to discover hidden semantic structures in a text. The earliest topic model was known as ‘Latent Semantic Indexing’ (LSI or LSA) and was developed by C.H. Papadimitriou et al. in 1998 [44]. T. Hoffmann in 1999 developed another method of topic modeling known as ‘Probabilistic Latent Semantic Indexing’ (PLSA) [45]. In 2002, D. Blei et al. introduced ‘Latent Dirichlet Allocation’ (LDA) [46], and it became one of the most widely used topic modeling approaches. LDA is a generative probabilistic model which describes each document as a mixture of topics and each topic as a distribution of words. LDA

generalized PLSA by adding a Dirichlet prior distribution over document-topic and topic-word distributions. The LDA model assumes that the words of each document arise from a mixture of topics, each of which is a distribution over the vocabulary. However, the inability to model topic correlation is the limitation of LDA [47]. Non-negative matrix factorization (NMF) is another popular topic modeling approach. The early work on NMF was performed in the early 1990s under the name positive matrix factorization by P. Paatero et al. [48]. Lee and Seung in 1999 popularized it as NMF by investigated the properties of the NMF algorithm and provided some simple and useful algorithms for two types of factorizations [49]. NMF as a topic modeling tool was proposed by S. Arora et al. in 2012 in "Learning Topic Models—Going Beyond SVD." According to this paper, previously existing approaches rely on Singular Value Decomposition (SVD) and thus have two limitations: 1) these works need to assume that each document contains only one topic and 2) can only recover the span of the topic vectors instead of topic vectors themselves. NMF approach is analogous to SVD, where all vectors are non-negative and produce topic models without the above two limitations [50]. The latest development in topic modeling is 'Top2vec' proposed by D. Angelov in 2020 in the paper "Top2Vec: Distribution Representation of Topics" [51]. In our research, NMF and Top2Vec topic modeling approaches have been utilized. It is imperative to discuss these two approaches in detail.

3.2.1 NMF

NMF is a group of algorithms in multivariate analysis and linear algebra where a matrix V is factorized into two matrices W , and H . NMF is also a matrix decomposition technique like Singular Value Decomposition (SVD) [52]; however, unlike SVD, NMF operates on non-negative matrices. Given a non-negative matrix V , the objective of NMF is to find two non-negative matrix factors, W and H , such that when they are multiplied, they can approximately reconstruct V . It can be represented as:

$$V \approx WH \quad (3.1)$$

All three matrices are non-negative. To get this approximation, a cost function such as L2 norm or Euclidean distance between two matrices or the Frobenius norm, which is a slight modification of the L2 norm, is used. The equation can be represented as follows:

$$\arg \min_{W,H} \frac{1}{2} \|V - WH\|^2 \quad (3.2)$$

Here we have three non-negative matrices – V, W, and H, and this can be further simplified as below:

$$\frac{1}{2} \sum_{i,j} (V_{ij} - WH_{ij})^2 \quad (3.3)$$

Figure 3.2 shows the representation of equation (3.1).

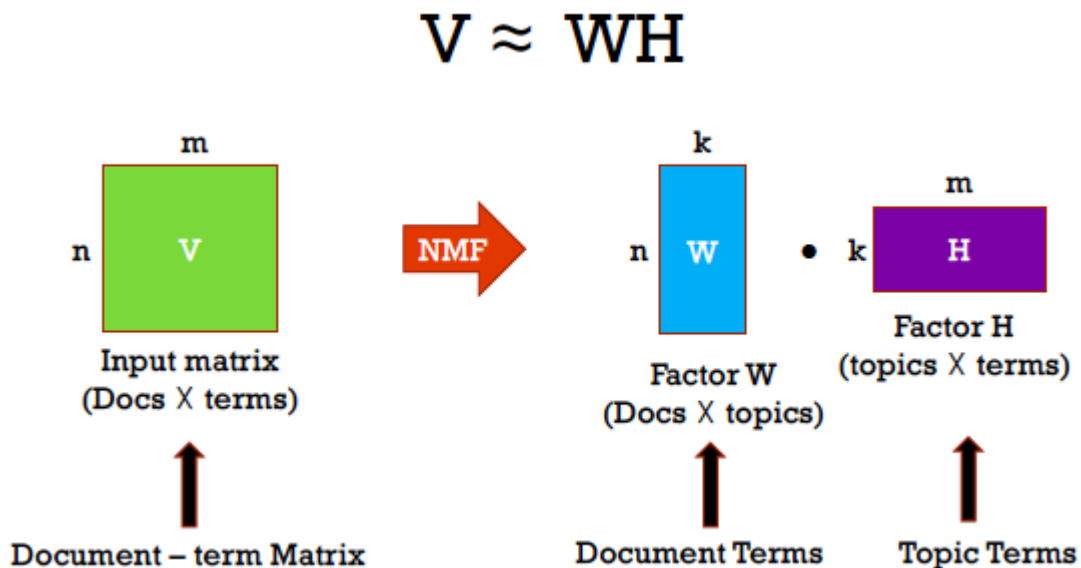


Figure 3.2: Representation of NMF equation

To understand the working of equation (3.1), a sample dataset of six articles is created. Table 3.1 shows the sample dataset.

Table 3.1: **Sample Dataset**

Document No.	Text
Doc_1	british cybersecurity inspectors find significant technical issue chinese telecom supplier huawei software
Doc_2	russian claim week country exonerate robert mueller final report make skin crawl highlight critical question
Doc_3	despite advance infrastructure connectivity heart southeast asia malaysia not able fully exploit capacities
Doc_4	park hot seat since group shoppers use online payment system fell prey cyberattack amid grow concern
Doc_5	yi open honorary consulate seoul april accord tunisian embassy korea last say choi president cybersecurity
Doc_6	southeast asia offer market business opportunities south korea young digitally connect grow middle class

We used TfidfVectorizer to convert text into vectors with $\text{max_df} = 0.8$ and $\text{min_df} = 2$ parameters, which gave us five terms – asia, Cybersecurity, grow, korea, southeast. Then these terms are used to create a document term matrix (V). Table 3.2 shows the document term matrix.

After V , now we need factors W and H . Factor W is a document term that shows weights for six documents in three topics (we opt for three topics for this sample data). Table 3.3 shows factor W weights for our sample data.

Factor H is the topic term that shows weights for five terms relative to three topics. Table 3.4 shows factor H weights for our sample data.

In our research, a Python-based Scikit-learn decomposition module is used to apply

Table 3.2: Document-Term Matrix (V) in Sample Dataset

	Asia	Cybersecurity	Grow	Korea	Southeast
Doc_1	0.0	1.0	0.0	0.0	0.0
Doc_2	0.0	0.0	0.0	0.0	0.0
Doc_3	0.70	0.0	0.0	0.0	0.70
Doc_4	0.0	0.0	1.0	0.0	0.0
Doc_5	0.0	0.70	0.0	0.70	0.0
Doc_6	0.5	0.0	0.5	0.5	0.5

Table 3.3: Document-Term (W) in Sample Dataset

	Topic_1	Topic_2	Topic_3
Doc_1	0.0	0.88	0.0
Doc_2	0.0	0.0	0.0
Doc_3	0.77	0.0	0.0
Doc_4	0.0	0.0	1.07
Doc_5	0.05	0.88	0.05
Doc_6	0.61	0.09	0.58

Table 3.4: Topic-Term Matrix (H) in Sample Dataset

	Topic_1	Topic_2	Topic_3
Asia	0.87	0.0	0.0
Cybersecurity	0.0	0.95	0.0
Grow	0.0	0.0	0.90
Korea	0.26	0.40	0.12
Southeast	0.87	0.0	0.0

the NMF topic modeling algorithm.

3.2.2 Top2Vec

All the topic modeling approaches such as PLSA, LDA, and NMF works on the bag-of-words (BOW) representation of documents and, because of this, ignores word semantics. Further, these approaches also require the researcher to know the number of topics in advance, which is problematic. Lastly, these techniques also need preprocessing. All these limitations were taken into consideration in the newly proposed Top2Vec algorithm.

Distributed representation concept is central to deep learning. Distributed representation means each concept learned by the network is represented by many neurons. Therefore, each neuron participates in the representation of many concepts [53]. Distributed representations are often central to NLP machine learning techniques for learning vector representation of words and documents.

The continuous skip-gram and BOW model known as word2vec, introduced distributed word representations that capture syntactic and semantic word relationships [54]. Though word2vec produced state-of-the-art results on many linguistics tasks, however, it lacked the ability to scale to large corpora. To overcome this weakness, the distributed paragraph vector was proposed with doc2vec [55]. Doc2vec can learn distributed representations of varying text lengths, from sentences to documents. In addition to the context of words, a paragraph vector is also used to predict which adjacent words should be present. The newly proposed Top2Vec model takes the word2vec and doc2vec idea further. Top2vec works on the assumption that many semantically similar documents are indicative of the underlying topic. It produces jointly embedded topic, document, and word vectors such that the distance between them represents semantic similarity. Further, it does not require removing stop-words, stemming, and lemmatization of text, and there is no need to have prior knowledge of existing topics

to produce a good topic model.

Figure 3.3 shows the working of the Top2Vec algorithm. In the first step, jointly embedded document and word vectors are created using Doc2Vec. This step will place documents close to other similar documents and the most distinguished words. The next step will use UMAP [56] to create a lower-dimensional embedding of document vectors. As document vectors in high dimensional space are very sparse, dimension reduction helps find dense areas. In this, each point is a document vector. The third step would find dense areas of documents using HDBSCAN [57]. Here, the colored areas are the dense areas of documents, and red points are outliers that do not belong to a specific cluster. Finally, each dense area calculates the centroid of document vectors in the original dimension, the topic vector. The purple points are document vectors that belong to a dense area from which the topic vector is calculated (red points are not used for calculating the topic vector). Python-based Top2Vec module is used in our research .

3.3 Clustering

Clustering is also known as cluster analysis, is the task of grouping a set of objects in such a way that objects in the same group (a cluster) are more similar to each other than to those in other groups (clusters). Clustering is a common technique in many fields such as image analysis, bioinformatics, machine learning, and information retrieval. There are many clustering algorithms such as connectivity-based clustering (hierarchical clustering), centroid-based clustering (k-means), distribution-based clustering (Gaussian mixture models), density-based clustering (DBSCAN and OPTICS), and grid-based clustering (STING and CLIQUE) [58].

The k-means is a centroid-based clustering algorithm where clusters are represented by a central vector that may not necessarily be a member of the dataset. When the

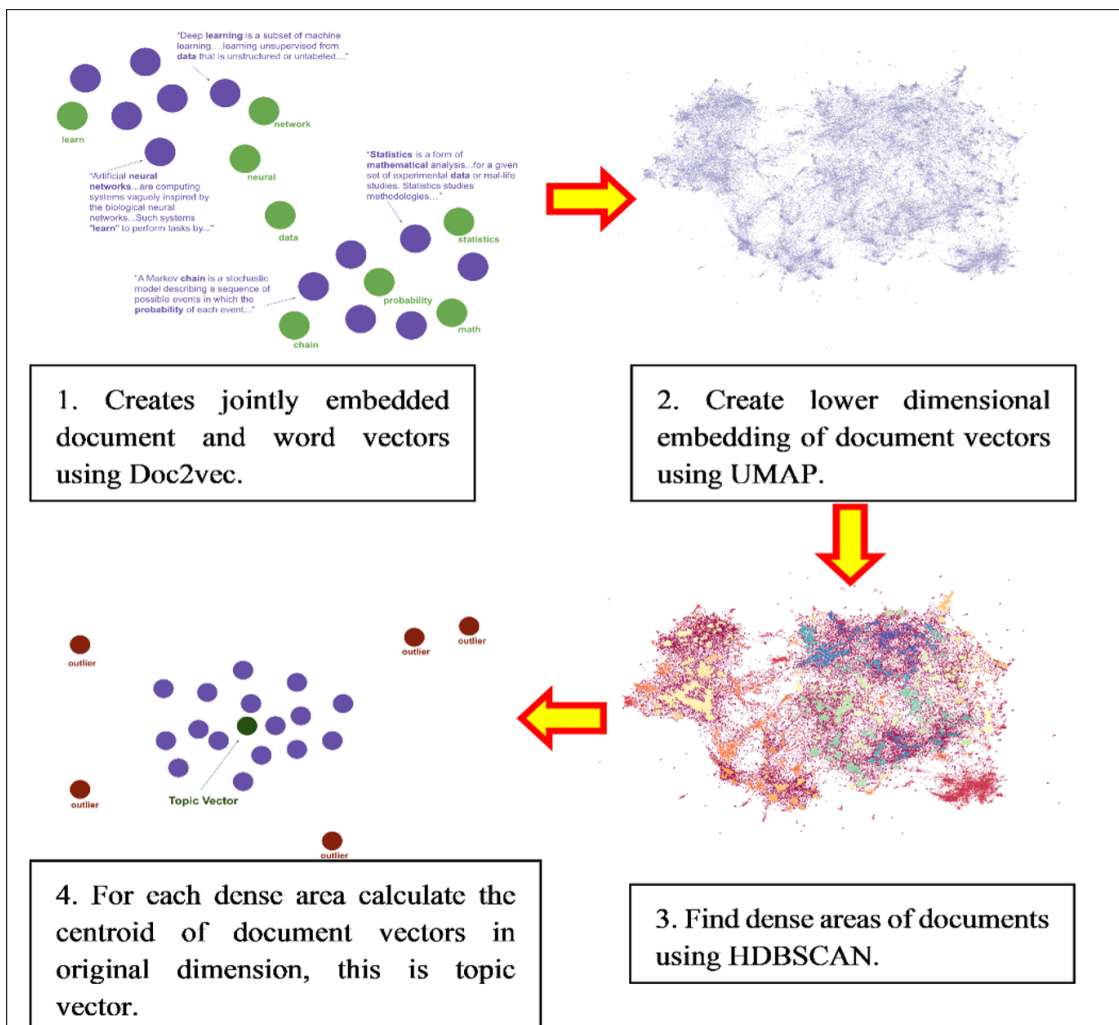


Figure 3.3: Four steps showing the working of the Top2Vec algorithm

number of clusters is fixed to k , k -means clustering gives a formal definition as an optimization problem: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized. This algorithm is one of the most popular clustering algorithms due to its ease of use and scale with a large amount of data. However, one of the main disadvantages of all centroid-based clustering models, including k -means, is that the number of clusters (k) needs to be specified in advance [41]. The k -means a type of hard clustering where each

data point either belongs to a cluster completely or not.

3.4 Sentiment Analysis

Sentiment analysis is also called opinion mining, is the field of study that analyzes people's opinions, sentiments, appraisals, attitudes, and emotions toward entities and their attributes expressed in a written text [59]. The entities can be products, organizations, individuals, services, events, topics, or issues. Sentiment analysis is an umbrella term used for all the related or slightly different tasks such as sentiment analysis, opinion analysis, review mining, sentiment mining, opinion mining, subjectivity analysis, emotion analysis, etc. Though sentiment analysis studies opinion text, there was almost no research on it before the year 2000 because of the non-existence of text in digital form. With the explosive growth of the web and social media in the past two decades, there is a constant flow of opinion data in digital forms.

Opinions are critical to businesses, organizations, and governments as they want to find out what the consumers or public is thinking about their products or policies. Thus, the application of sentiment analysis exists in almost every field, from businesses to international relations to politics to tourism to economy, etc. Some examples of sentiment analysis research-related to various fields are: Politics – C. S. G. Khoo et al. analyzed sentiment in political news articles about economic policies and political figures [60], A. Tumasjan et al. showed that even the mentions of the party on Twitter can be a good predictor of election results [61]. Stocks market – X. Zhang et al. identified negative and positive public moods on Twitter and used them to predict the movement of stock market indices such as SP 500, the Dow Jones, and NASDAQ [62]. Predict movie revenue – Y. Liu et al. [63] and S. Asur and B. A. Huberman [64] used sentiment analysis to predict box-office revenue. Sentiment analysis research can be carried out at three levels: document level, sentence level, and aspect level [65].

- **Document Level** – This type of sentiment analysis is performed on the document level to determine whether the whole document is positive or negative. For example, sentiment analysis of product review, movie review, or news articles are document-level sentiment analysis. This type of sentiment analysis shows whether the review (product or movie) or article expresses positive or negative sentiment. Our research has performed this type of sentiment analysis.
- **Sentence Level** – Sentence level sentiment analysis shows whether a sentence is positive, negative, or neutral (no opinion). This type of sentiment analysis is closely related to subjectivity classification, which distinguishes sentences that express factual information (objective sentences) from sentences that express subjective views (subjective sentences).
- **Aspect Level** – This type of sentiment analysis discovers what people like and dislike. It is neither document-level nor sentence-level. For example, "Tech companies are doing very well even when the economy is performing very badly in this COVID-19 situation." To classify this sentence as positive or negative will not make any sense because it is positive for tech companies but negative about the economy. For this, we have to go to aspect-level sentiment analysis. This was earlier called feature level or feature-based opinion mining.

As shown in Figure 3.4, sentiment analysis techniques can be divided into ML and Lexicon-based methods. ML sentiment analysis can further be divided into unsupervised and supervised approaches [66]. As Lexicon-based methods such as SentiWordNet [67], VADER [68], and Textblob [69] do not require any training or label dataset, these methods can be seen as an unsupervised approach. The below description also follows the same viewpoint. Below is a detailed explanation of unsupervised and supervised sentiment analysis methods.

- **Unsupervised Sentiment Analysis** – For this type of sentiment analysis, no

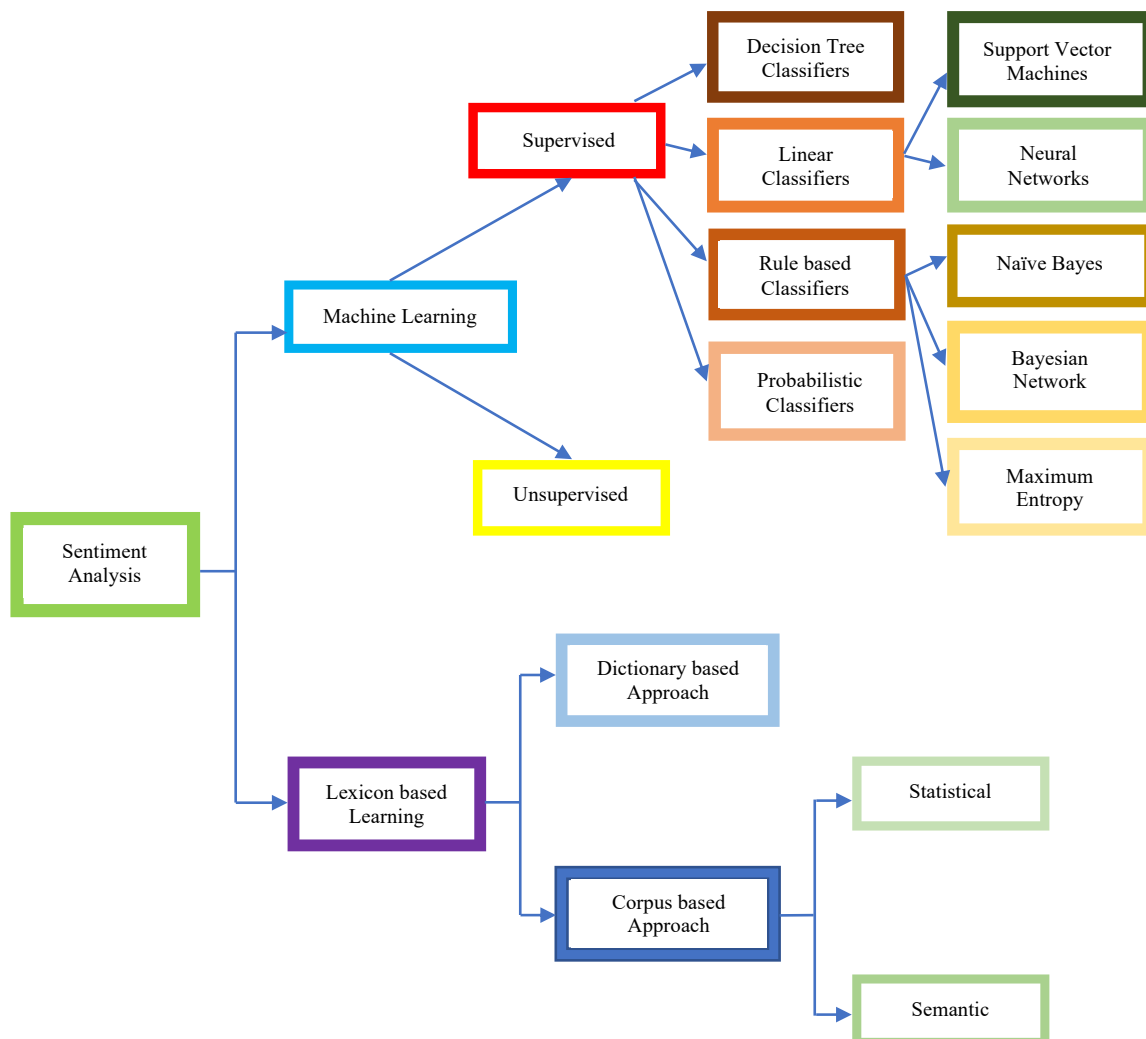


Figure 3.4: Various Approaches for Sentiment Analysis

training data or labeled data is required. Researchers need to input the text they want to classify, and output would tell the sentiment of that text. Various python-based modules (libraries) such as NLTK, VADER, IBM Watson, Textblob, and SentiWordNet can be used to perform unsupervised sentiment analysis. In our research, we have used VADER, Textblob, and SentiWordNet libraries.

- **Supervised Sentiment Analysis** – For any supervised machine learning ap-

proach, the first requirement is the availability of training or labeled data. This is also true for supervised sentiment analysis. This requirement also works as one of the main limitations for this kind of analysis. There is no 'one type fit for all' training dataset, which can be used for the text of various fields such as politics, businesses, and security. Thus, research either tries to find the most suitable training dataset from the available options or creates a training dataset manually. Some of the most popular examples of training datasets are – Stanford Sentiment Treebank [70], IMBD Movie Reviews Dataset [71], Twitter US Airline Sentiment, Sentiment140 [72], and Amazon Product Data. We have used the IMDB dataset for the Cybersecurity case study, and for the COVID-19 case study, we have developed our training dataset.

In this chapter, the main methods or approaches used in our research are introduced in detail. Subsequent chapters are case studies where these methods are utilized.

Chapter 4

Understanding Cybersecurity News by Utilizing Topic Modeling and Sentiment Analysis Methods

4.1 Motivation

As the importance of cybersecurity is increasing, it is no surprise that Cybersecurity coverage in the media has also increased steadily over the past decade. According to the study by M.R. Alagheband [18], the New York Times's (NYT) coverage of Cybersecurity-related events increased from 81 stories in 2008 to more than 900 in 2019. The significance of newspapers from the point of view of the analysis of Cybersecurity events is further increased as newspapers not only captures the multidimensional nature of Cybersecurity (politics, business, individual, technological advancement) but are also one of the most trustworthy sources of information in the times when fake news and disinformation are everywhere. However, as the coverage of Cybersecurity-related events increases, finding or discovering critical 'information' from a large amount of data (text) is challenging. NLP has made the analysis of a large volume of text pos-

sible and provided a new dimension for analytics. This case study's main objective is to understand Cybersecurity trends, popular themes, topics, and sentiments behind these issues in news media. To achieve this objective of analyzing Cybersecurity-related newspaper articles, advanced NLP and ML methods would be utilized.

Organization of the rest of the chapter: Section 4.2 would present the research methodology. Next, Section 4.3 would explain in detail the experiment, results, and analysis. Section 4.4 and 4.5 would talk about limitations and the conclusion for this case study.

4.2 Research Methodology

The research methodology of this case study is presented in Figure 4.1. The first step is to collect the Cybersecurity-related news articles, and then the next step is to clean those articles using text preprocessing techniques. After that, based on the methods, it is further divided into two main steps. One is topic modeling, and the other is sentiment analysis.

4.2.1 Data Acquisition

As the focus of this research is Cybersecurity, we scraped all the cybersecurity articles (articles with cybersecurity keyword) that published in English version websites of 18 newspapers from six countries during April 1, 2018, and March 31, 2019, and for that, we used Python-based BeautifulSoup library¹. We collected 4159 articles, with the US having the highest 1615 articles and South Korea has the lowest 87 articles. Below, Table 4.1 shows our data with countries (alphabetically arranged), newspapers, and the number of articles.

¹<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

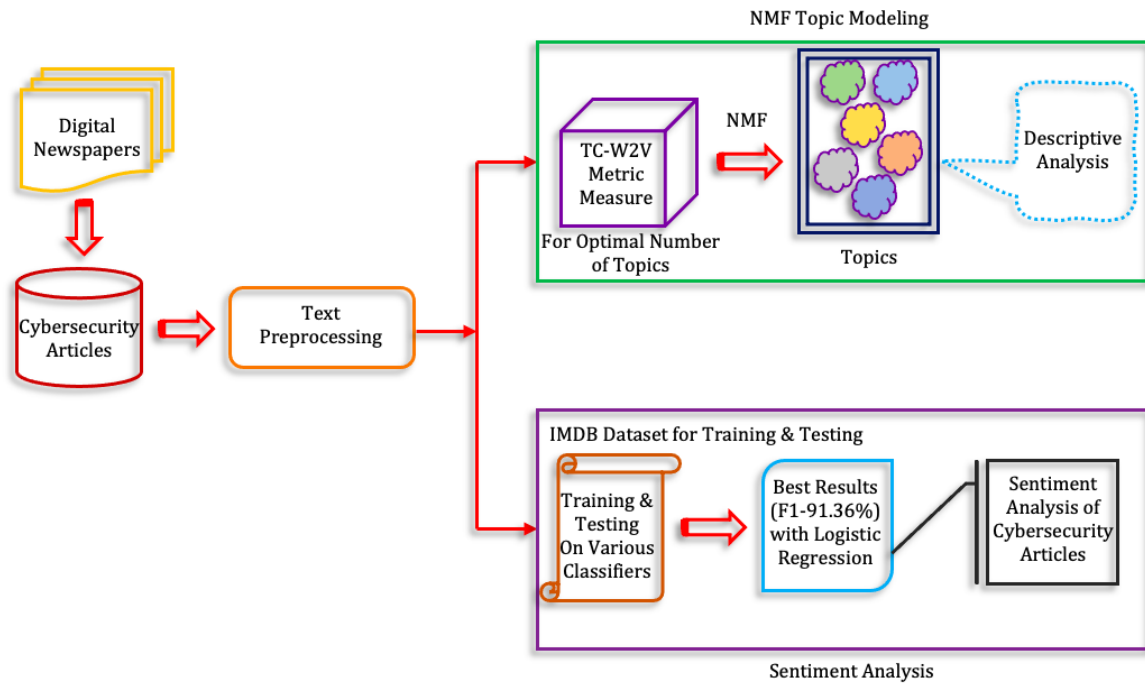


Figure 4.1: Research Methodology for Cybersecurity Case Study

4.2.2 Text Preprocessing

As explained in detail in Chapter 3 (section 3.1), text preprocessing is used to remove unnecessary sections from the text. Below is the orderly list of the text preprocessing steps.

- (1) Contractions are shortened versions of words such as don't or I'd. In the first step, we expanded all these words into their original form.
- (2) Next, we removed all the HTML tags.
- (3) Tokenization would convert each word into a token.
- (4) The root word is also known as lemma. With lemmatization, we converted all the tokens into their root word.

Table 4.1: Collected articles from different countries and Newspapers

Country	Newspaper	No. of Articles (total of all three newspapers)
Canada	Globe and Mail	749
	The Hamilton Spectator	
	The Star	
India	The Indian Express	473
	The Hindu	
	Hindustan Times	
Japan	The Japan Times	179
	Asahi Shimbun	
	Mainichi Shimbun	
South Korea (SK)	The Chosun Ilbo	87
	The Korea Herald	
	The Korea Times	
United Kingdom (UK)	Daily Mail	1056
	Metro	
	The Sun	
United States of America (US)	The New York Times	1615
	Chicago Tribune	
	Houston Chronicle	
Total		4159

- (5) Then we removed stopwords such as a, the, an, and this by using a custom stopwords list.
- (6) Special characters are usually non-alphanumeric characters, which can add noise to the text. So, we removed the special characters.
- (7) As our corpus is the English language, we need to make sure all the accented characters are converted and standardized into ASCII characters.

- (8) In the last step, we converted all the words or tokens into lowercase.

4.2.3 NMF Topic Modeling

Topic modeling is an unsupervised machine learning approach, and to achieve the best results, the number of topics should be known in advance. Because of this reason, topic modeling can be further divided into two steps: 1) Find out the optimal number of topics and 2) NMF topic modeling.

To find out the optimal number of topics, we applied Topic Coherence – Word2Vec (TC-W2V) metric [73]. This metric, measures the coherence between words assigned to a topic. For example: how semantically close are the words that describe a topic. We train the word2vec model [54] on our corpus (articles), which would organize the words in an n-dimensional space where semantically similar words are close to each other. The TC-W2V for a topic will be the average similarity between all pairs of the top n- words describing the topic. We trained the NMF model for different values of the topic ($k = 5$ to 30) and calculated the average TC-W2V across all the topics.

The k with the highest average TC-W2V is used to train the final NMF model. The results (topics) of NMF topic modeling are visualized using the PyLDAvis package (Python-based). PyLDAvis is based on LDAvis [74], a visualization tool made for R. PyLDAvis uses Multidimensional Scaling (MDS) [75] to visualize the topics in a two-dimensional plane. MDS is a means of visualizing the level of similarity of individual cases of a dataset. Given a distance matrix with the distances between each pair of objects in a set and a chosen number of dimensions, N , MDS algorithm places each object into N -dimensional space such that the between-object distances are preserved as well as possible. For $N=1, 2$, and 3 , the resulting points can be visualized on a scatter plot [76].

4.2.4 Sentiment Analysis

As explained in section 3.4 of chapter 3, sentiment analysis has two main methods: 1) Unsupervised or Lexicon-based and 2) Supervised (ML). To predict correct sentiments, it is critical to choose a method that provides the best result. Keeping this in mind, we first experimented with different libraries (unsupervised), features, and classifiers (supervised) to see which method is the most accurate. Next, we will use that method to predict the sentiments. After that, we will discuss and try to interpret our results. It is also essential here to understand the nature of news articles. News articles are different from tweets (only 140 characters) in length and written from the subjective perspective. Tweets are very subjective, and while the news articles (even editorials/opinions are fact-oriented) are objectively written. Moreover, it is paramount to recognize that news (good/bad or negative/positive) is different from sentiments (negative/positive) in news articles' sentiment analysis. For example, a news article is about a cyber-attack (bad/negative news) does not necessarily mean it has negative sentiment. Below is a sample (not a complete article) to prove our point.

“The number of cybercrimes confirmed by police nationwide in 2018, including many cases of child pornography and fraud, stood at 9,040, rising slightly to reach a record high for the third consecutive year, National Police Agency data showed Thursday., The figure was up by just 26 from a year before, but it marked an increase of more than 1,000 from 2014.”

This sample is about the increasing numbers of cybercrimes (negative). After pre-processing this sample, VADER (which shows how much (percentage) of a text is negative, positive, and neutral) is used, and it shows that 75.2% of this text is neutral. Whereas negative and positive portion accounts, only 14.4% and 10.4% respectively. This example shows that the percentage of sentiment-related content (words) is small in news articles, and because of this reason, our sentiment analysis discussion would focus on the number (percentage) of negative articles.

4.3 Experiment, Results, and Discussion

The preprocessed Cybersecurity newspaper articles are first utilized to produce topics by NMF topic modeling and then used for sentiment analysis.

4.3.1 NMF Topic Modeling on Preprocessed Cybersecurity Articles

The Cybersecurity corpus has six countries, and for each country, there would be two steps as mentioned in research methodology section 4.2.3. In the first part (A) TC-W2V is used to find the optimal number of topics, and the second part (B) is about producing topics (NMF algorithm), visualization (MDS), and analysis. This section would help us find answers to RQ 1, RQ 2, and RQ 3 for cybersecurity.

4.3.1.1 NMF Topic Modeling for Canada

- (A) **Optimal Number of Topics** – There are 749 Cybersecurity articles that we collected from Canada. NMF model was trained from $k_{min} = 5$ to $k_{max} = 25$, and then a 500 dimensions w2v model was built. The resulting w2v model had 10,860 terms. Based on this w2v model, the coherence value is calculated. The highest coherence (69.23%) is achieved by a model with 21 topics. Hence, $k = 21$ is the optimal number of topics in the Canada dataset. Figure 4.2 is a scatter plot that shows the mean coherence for each topic.
- (B) **Topics, Visualization, and Analysis** – After finding the best k ($k=21$) for Canada data, we now discuss these topics. We would present the top 10 topics with topic weightage (%), top keywords, and topic label. The total percentage of these top 10 topics is 66.8% which shows that majority of the news is covered by these topics. Table 4.2 shows the top ten topics in Canada dataset.

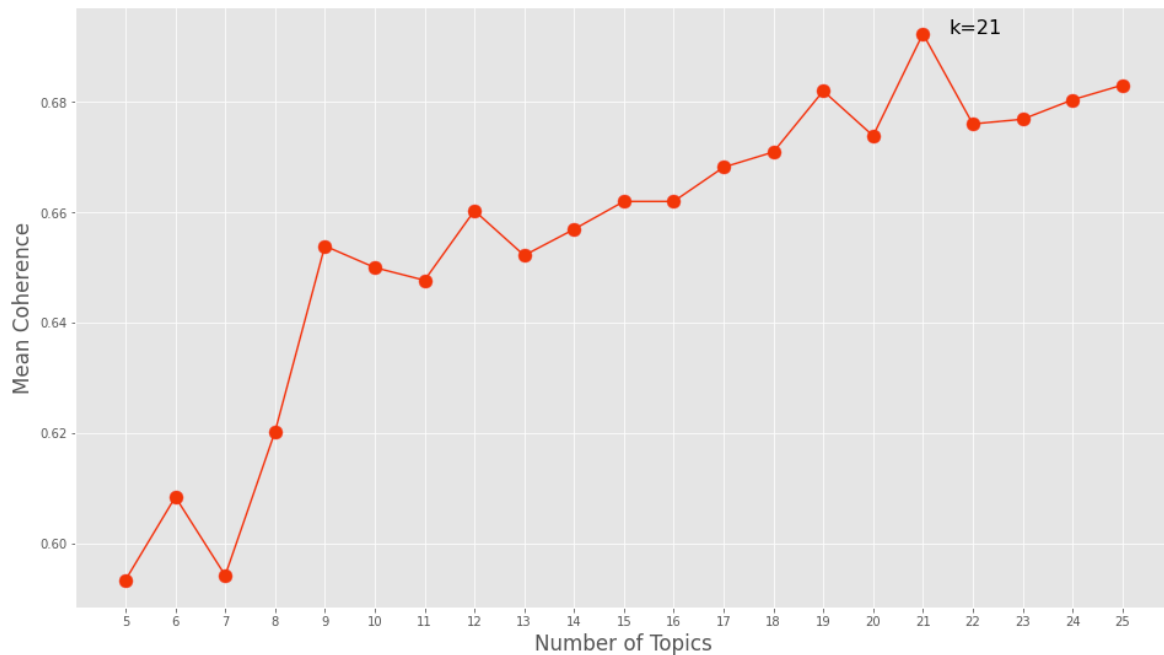


Figure 4.2: Mean Coherence and Number of Topics (Canada)

The Canada dataset’s biggest topic is about market and investment, with 9.3% tokens belonging to this topic. US-China trade dispute is the next biggest topic with 8.1% tokens. There are two topics related to the Huawei issue. Topic 4 (7.6% tokens) is about the arrest of Huawei CFO Meng Wanzhou by Canadian authorities, and topic 5 (7.5% tokens) about the Huawei 5G issue. Other topics such as jobs and education (topic 3), Mueller investigation (topic 9), fake page removal from social media (topic 10) also find a place in the top 10 position. Figure 4.3 present the visualization of Canada topics in MDS.

In the Figure 4.3, bubbles on the left side show that the topics and bars on the right side show the 30 most salient terms (tokens). As it is visible, topic 4 (arrest of Huawei CFO) and 5 (Huawei 5G issue) are intersecting with topic 2 (US-China trade dispute). These topics are related to each other or part of the same theme. Further, as the topics are fairly distributed, this topic modeling is a coherent one. The right side of Fig 4.3

Table 4.2: Canada Top 10 Topics

No.	Topic Weightage (%)	Keywords	Topic Label
1.	9.3%	Company, Board, Investors, SEC, Fund, Story, Firm, Capital, Venture, Subscribers	Market and Investment
2.	8.1%	China, Chinese, Trade, Beijing, Tariff, Xi, United, Deal, States, Trump	US-China Trade Dispute
3.	7.7%	Job, Workers, Program, Technology, Students, Car, Talent, Skills, Robots, Tech	Jobs and Education in Tech Industry
4.	7.6%	Canada, Huawei, Meng, Canadian, China, Arrest, Chinese, Ottawa, Trudeau, Universities	Arrest of Meng Wanzhou by Canadian Authorities
5.	7.5%	Huawei, Equipment, Network, Company, Chinese, Security, Telecom, Government, Officials, Ally	Huawei 5G Issue
6.	6.8%	Privacy, Data, Information, Personal, Commissioner, Breach, Users, Facebook, Law, Protection	Data breach and Privacy
7.	5.2%	Attack, Hackers, Target, Security, Hack, Cyber, Company, Systems, Email, Network	Hacking and Cyber Attacks
8.	5.1%	Cent, Stock, Price, Market, Trade, Growth, Share, Rise, Index, Globe	Stock and Market
9.	4.9%	Trump, Giuliani, President, Mueller, Cohen, White, House, Clinton, Russian, Putin	Mueller Investigation
10.	4.6%	Facebook, Account, Page, Users, Social, Remove, Fake, Media, Company, Iranian	Fake Page Removal from Facebook

also shows that the most frequently used term in Canada dataset is 'Huawei.' This also verifies that, during this study, the Huawei 5G issue is the most widely reported Cybersecurity topic in Canadian newspapers.

4.3.1.2 NMF Topic Modeling for India

- (A) **Optimal Number of Topics** – We collected 473 Cybersecurity related articles from Indian newspapers. After preprocessing, we trained the NMF model from $k_{min} = 5$ to $k_{max} = 25$ and built a 500 dimensions w2v model. The resulting w2v model had 7403 terms. Then, to find out the best k, we calculated coherence for each value of k. The best model turns out to be $k = 22$ with a mean coherence

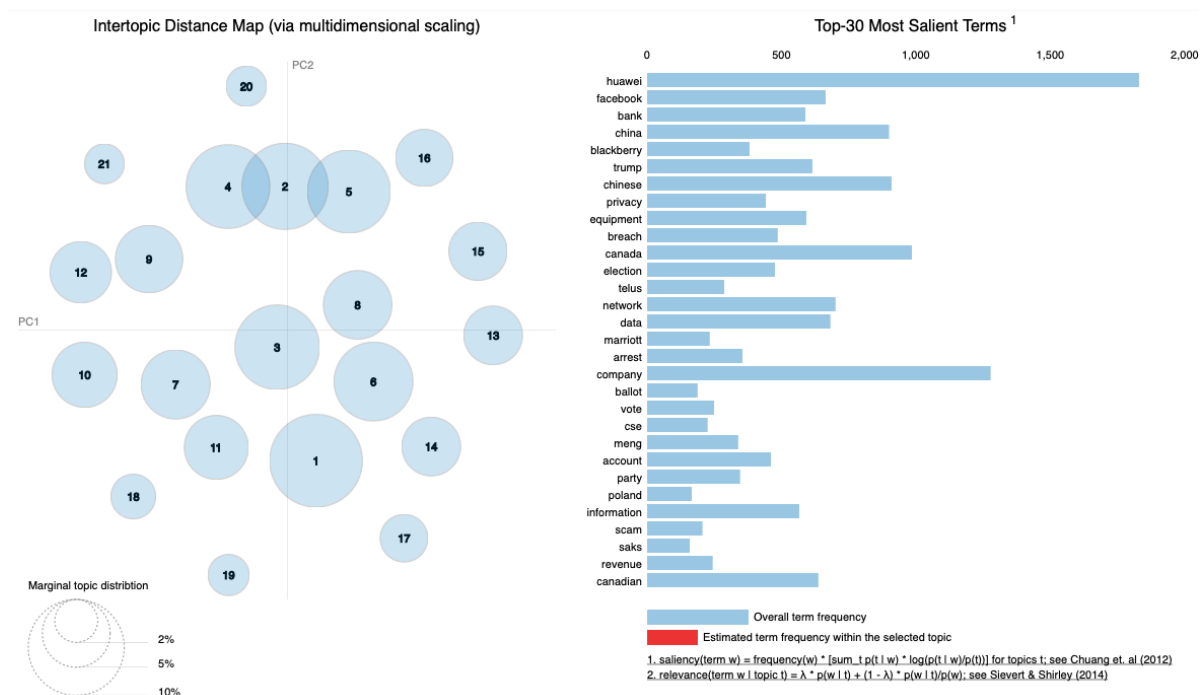


Figure 4.3: All Topics in MDS (Canada)

score of 87.84%. Figure 4.4 shows the mean coherence and number of topics for the Indian dataset.

- (B) **Topics, Visualization, and Analysis** – Figure 4.4 shows that $k = 22$ is the optimal number of topics for our Indian Cybersecurity dataset. Using this number, next, we created the NMF model to find what are these topics. Table 4.3 shows the top 10 topics with topic weightage, keywords, and topic labels.

The biggest topic in the Indian dataset is 'Mueller Investigation' with 9.3% tokens. Data breach-related topic is the second biggest topic with 8.9% tokens. There are two topics (topics 7 and 8) that are related to Huawei 5G issue. Fake pages removal from Facebook was at the fourth position. Other topics present in the Indian dataset: Jobs and Skills, Cyberthreat report, Aadhaar data leak, and Google's AI project. MDS visualization of the Indian dataset is presented in Figure 4.5.

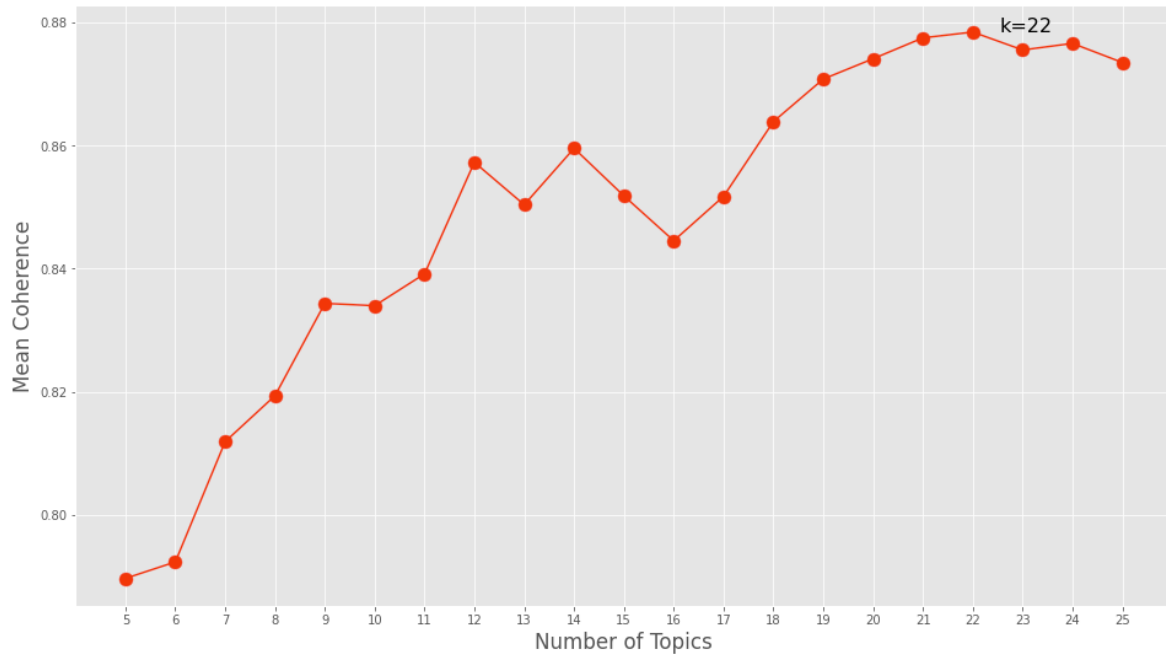


Figure 4.4: Mean Coherence and Number of Topics (India)

Topic 1 and 7 are intersecting, which shows that they are related. This relation is verified by the Table 4.3, where it shows that both topics are US-related (topic one is about the Mueller investigation, and topic seven is about the US-China trade dispute). Topic 8 (Huawei 5G issue) is also visible near both topics. In the Indian dataset, 'Company' and 'Facebook' are the two most frequent terms (token).

4.3.1.3 NMF Topic Modeling for Japan

- (A) **Optimal Number of Topics** – From Japan's English newspapers, we collected 179 Cybersecurity related articles. After preprocessing, we trained the NMF model from $k_{min} = 5$ to $k_{max} = 27$ and built a 500 dimensions w2v model. The resulting w2v model had 4388 terms. Then, to find out the best k , we calculated coherence for each value of k . The best model turns out to be $k = 22$ with a mean coherence score of 98.01%. Figure 4.6 shows the mean coherence and number of

Table 4.3: India Top 10 Topics

No.	Topic Weightage (%)	Keywords	Topic Label
1.	9.3%	Trump, Russian, Russia, Election, Putin, Mueller, Campaign, Democratic, President, Clinton	Mueller Investigation
2.	8.9%	Data, Users, Breach, Information, Law, Access, Privacy, Facebook, Security, Email	Social Media, Privacy, and Data Breach
3.	7.6%	India, Modi, Bilateral, Prime, Cooperation, Trade, Minister, Indian, Countries, Bangladesh	India's Bilateral Cooperation
4.	7.5%	Facebook, Account, Page, Post, Remove, Media, Social, Fake, Instagram, Political	Fake Page removal from Facebook
5.	6.7%	Attack, Malware, Security, Cyber, Target, Hackers, Report, Ransomware, Threat, McAfee	Cyber threat Report
6.	5.8%	Job, India, Skills, Sector, Cloud, Data, Digital, Organisations, Need, Cent	Jobs and Skills
7.	5.7%	China, Chinese, Trade, United, Espionage, Trump, States, Beijing, Economic, Xi	US-China Trade Dispute
8.	5.6%	Huawei, Equipment, Chinese, Company, Meng, Telecom, Network, Ban, Security, Concern	Huawei 5G Issue
9.	5.3%	Google, Maven, Contract, Company, Project, AI, Pichai, Cloud, Military, Greene	Google's AI Project
10.	4.2%	Aadhaar, Card, Bank, UIDAI, Account, Varsity, Leak, Number, Employees, Patel	Aadhaar Number Data Leak

topics for the Indian dataset.

- (B) **Topics, Visualization, and Analysis** – Above figure shows that $k = 22$ is the optimal number of topics for the collected Cybersecurity dataset from Japan. Using this number, next, we created the NMF model to understand what these topics are. Table 4.4 shows the top 10 topics with topic weightage, keywords, and topic labels.

62.4% of all topics are covered by the top ten topics in Japan's dataset. 'Huawei 5G issue' with 11% tokens is the biggest topic. Topic 2 and 3 are both related to the US. While topic two, with 9.8% tokens, is about 'Russian interference in 2016 US presidential elections,' topic three is about the US-China trade dispute. Other topics

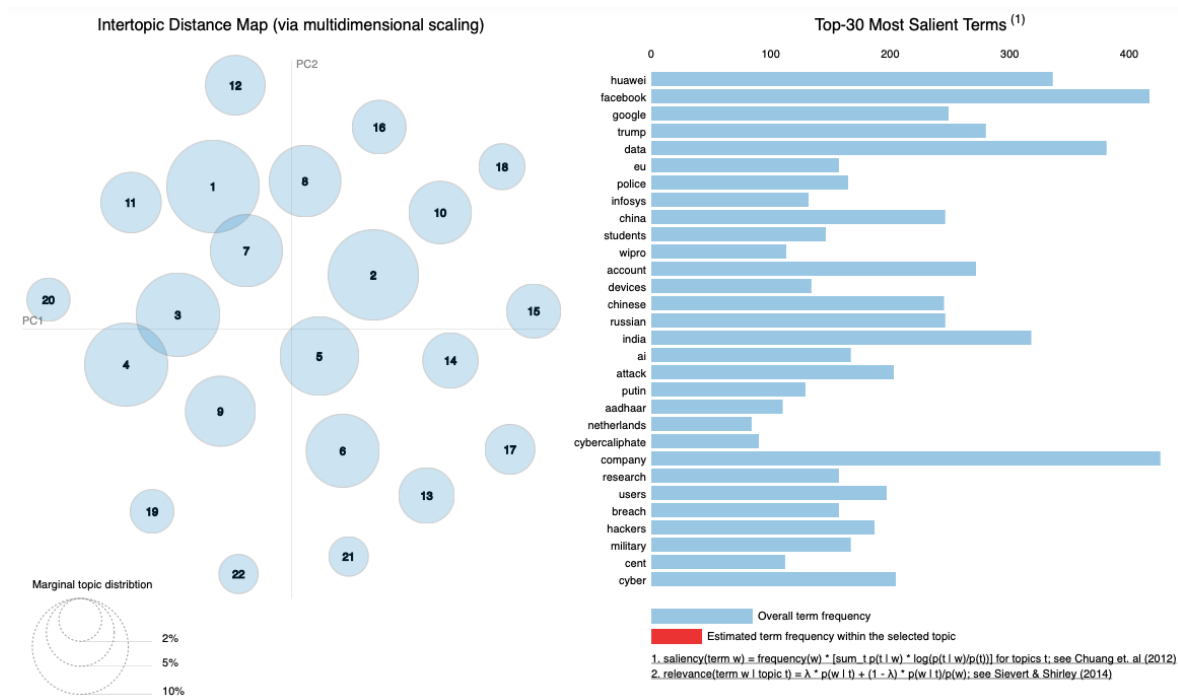


Figure 4.5: All Topics in MDS (India)

such as 'Sony Hack,' 'Cryptocurrency hacking,' 'Vietnam's New Cybersecurity Law,' are also found a place in the top ten. Figure 4.7 presents the distribution of all 22 topics in Japan's dataset.

Topic 1 and 3 are intersecting as both are about the US and China. Topic 4 and 7 are very close to each other as both involved 'Five Eyes'² discussion. On the right side, the biggest bar is in front of 'Huawei,' which signifies that Huawei is the most frequent term in Japan's dataset.

²Five Eyes is an intelligence alliance comprising Australia, Canada, New Zealand, the UK, and the US. Though not a member, Japan is along with Israel, Singapore, and South Korea, is collaborating with Five Eyes countries. <https://asia.nikkei.com/Politics/International-relations/Japan-lends-its-vision-to-Five-Eyes-intelligence-alliance>

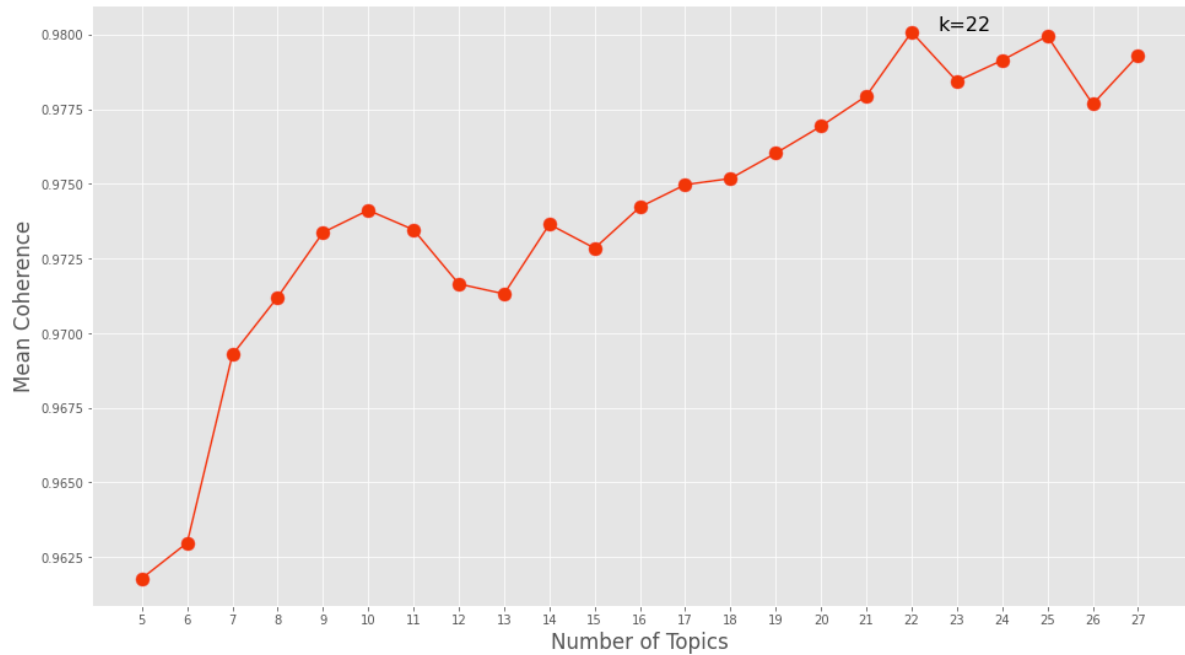


Figure 4.6: Mean Coherence and Number of Topics (Japan)

4.3.1.4 NMF Topic Modeling for South Korea

- (A) **Optimal Number of Topics** – Out of the six countries in our Cybersecurity corpus, we could collect the least number of articles (87) from South Korea’s newspapers. After preprocessing, we trained the NMF model from $k_{min} = 5$ to $k_{max} = 27$ and built a 500 dimensions w2v model. The resulting w2v model had 2617 terms. Then, to find out the best k , we calculated coherence for each value of k . The best model turns out to be $k = 26$ with a mean coherence score of 99.97%. Figure 4.8 shows the mean coherence and number of topics for South Korea’s dataset.
- (B) **Topics, Visualization, and Analysis** – Above figure shows that $k = 26$ is the optimal number of topics for the collected Cybersecurity dataset from South Korea. Using this number, next, we created the NMF model to understand what

Table 4.4: Japan Top 10 Topics

No.	Topic Weightage (%)	Keywords	Topic Label
1.	11%	Huawei, Network, Company, Equipment, Telecom, Chinese, Security, Mobile, Risk, Guo	Huawei 5G Issue
2.	9.8%	Trump, Russian, Russia, Campaign, Election, Elections, Intelligence, Hack, Cohen, Clinton	Russian Interference in 2016 US Election
3.	7.9%	European, Trade, EU, China, Chinese, Trump, Beijing, Li, President, Union	US-China Trade Dispute
4.	7.5%	Japanese, Japan, Government, Cyberattacks, Keidanren, Information, Eyes, Products, Share, Computer	Countermeasures against Cyberattacks
5.	4.8%	Cryptocurrency, Currency, Exchange, Virtual, Mine, Police, Currencies, Engineer, Boy, Digital	Cryptocurrency Hack
6.	4.8%	Vietnam, Law, Google, Data, Content, Internet, Bill, Facebook, Draft, Country	Vietnam's New Cybersecurity (2018)
7.	4.7%	Australia, Australian, Huawei, Chinese, Pacific, Cable, Guinea, Lord, Broadband, Government	Australia's action against Huawei 5G
8.	4%	Hackers, Chinese, Indictment, Case, Secrets, Steal, Department, Justice, Allege, China	Cyberespionage and China
9.	4%	North, Sony, Korea, Korean, Malware, Kim, Cobra, Hidden, Hack, FBI	Sony Hack involving North Korea
10.	3.9%	Nielsen, Trump, Kelly, Border, Immigration, Secretary, Cross, Families, President, Children	US Immigration Issue

these topics are. Table 4.5 shows the top 10 topics with topic weightage, keywords, and topic labels.

'Huawei 5G Issue' is the most widely reported topic in the collected South Korean articles with 8.5% tokens. 'Fake pages removal from Facebook' comes at second position with 7% tokens. North Korea-related topics such as 'Sony Hack' and 'Trump-Kim summit meet' present at third and eighth positions. This shows the focus of South Korean media on any news from North Korea. US-related topics such as 'Russian interference in 2016 US election' (topic 3) and 'US-China Trade dispute' (topic 4) also found a place in the top ten. The presence of ASEAN-related topics (topics 6 and 10) shows the region's economic importance for South Korea. Figure 4.9 shows the topic distribution of all 26 topics present in the South Korean dataset.

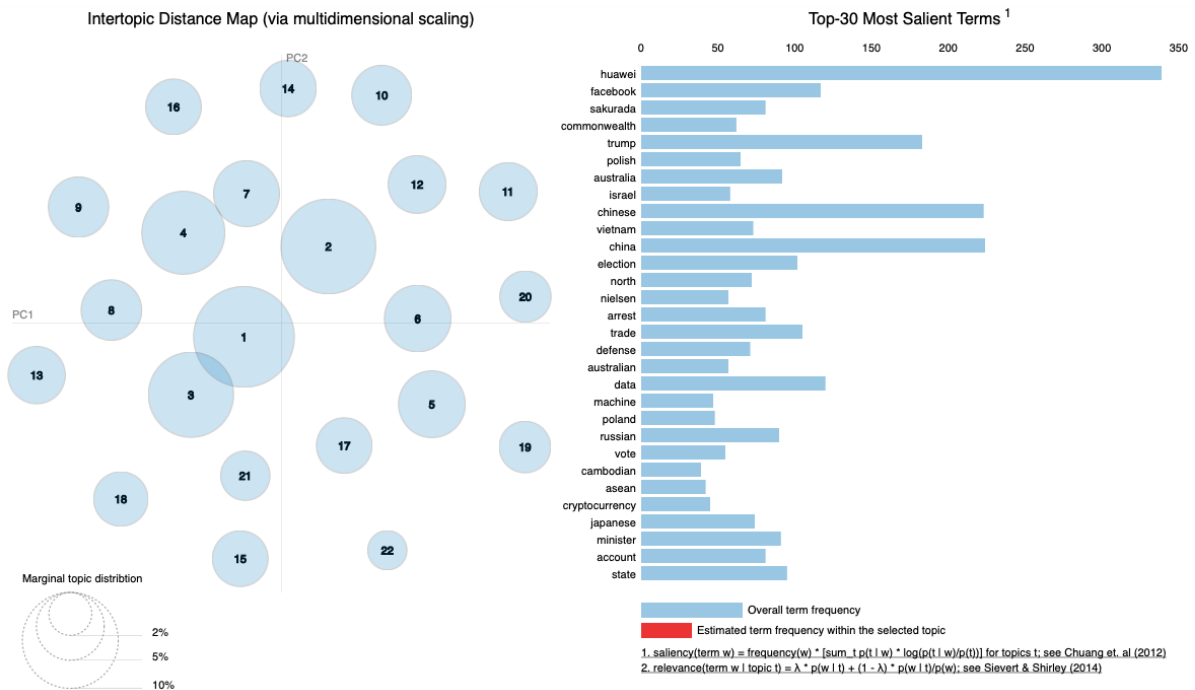


Figure 4.7: All Topics in MDS (Japan)

In both topics 2 and 4, Facebook is the main actor; hence, we see these two topics' intersecting. Similarly, topics 6 and 7 are also intersecting as the common point between them is trade. 'Huawei' and 'North Korea' are the two most frequent terms in the South Korean dataset. This shows the importance of these actors for South Korea.

4.3.1.5 NMF Topic Modeling for the UK

- (A) **Optimal Number of Topics** – We collected the second-highest number of articles (1056) from UK Newspapers. Using the same parameters ($k_{\min} = 5$ to $k_{\max} = 27$), the w2v model (500 dimensions) created by this dataset is also the second-largest with 12,564 terms. With this w2v model, $k = 19$ with the coherence score of 65.64% turn out to be the best fit. Figure 4.10 shows the mean coherence and number of topics for the UK dataset.

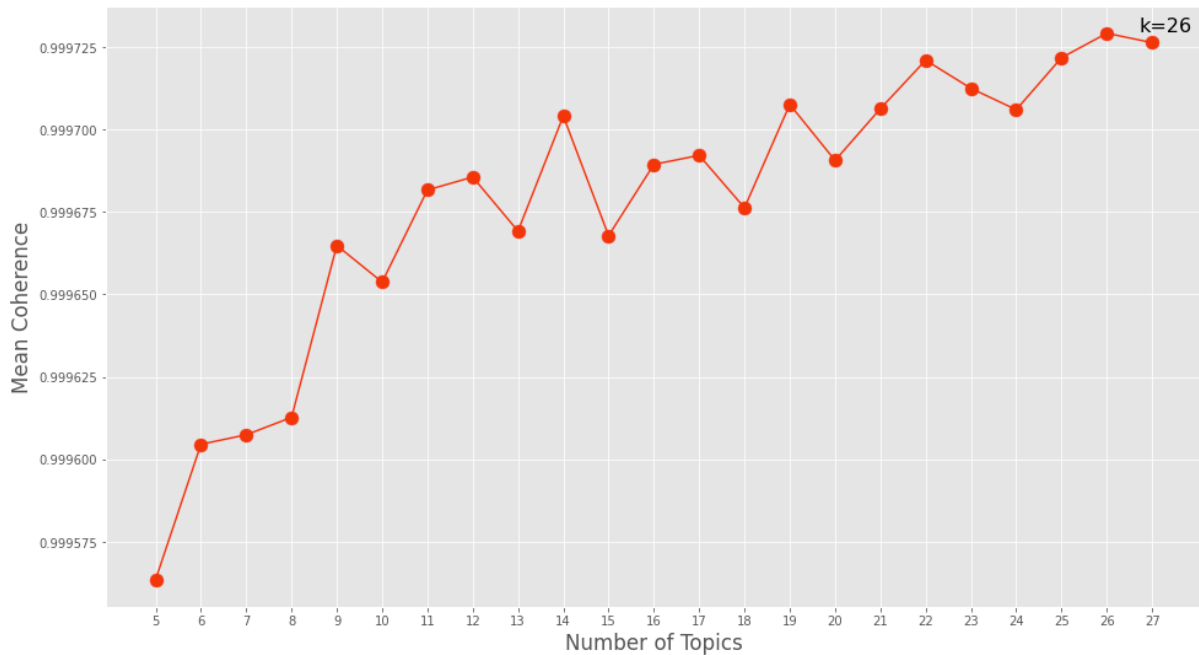


Figure 4.8: Mean Coherence and Number of Topics (South Korea)

- (B) **Topics, Visualization, and Analysis** – Above figure shows that $k = 19$ is the optimal number of topics for the UK’s collected Cybersecurity dataset. Using this number, next, we created the NMF model to understand what these topics are. Table 4.6 shows the top 10 topics with topic weightage, keywords, and topic labels.

‘Mueller Investigation’ with 9.1% tokens and ‘Cyberattack involving North Korea’ with 8.5% tokens are the two biggest topics in the UK dataset. Topic 3 is about ‘Russian interference in the 2016 US election’, which is also part of topic one. Topic 5 (Huawei 5G issue) is part of the broader topic 6 (the US-China trade dispute). Other topics such as ‘Various data breach,’ ‘US Voting Machine Security,’ and ‘Various Online Scams,’ are also placed in the top ten position. The total percentage these top ten topics cover is 68.6%. Figure 4.11 shows the topic distribution of all the topics and salient terms in MDS in the UK dataset.

Table 4.5: South Korea Top 10 Topics

No.	Topic Weightage (%)	Keywords	Topic Label
1.	8.5%	Huawei, Chinese, Company, Telecom, Network, Security, Government, Ban, Equipment, Czech	Huawei 5G Issue
2.	7%	Facebook, Account, Page, Gleicher, Iran, Link, Remove, Group, Fake, Instagram	Fake Page Removal from Facebook
3.	6.8%	North, Sony, Charge, Korean, Korea, Department, Complaint, Cyber, Pictures, Park	Sony Hack by North Korea
4.	6.3%	Russia, Russian, Trump, Moscow, Putin, Cyber, Russians, Dialogue, Mueller, UN	Russian Interference in 2016 US Election
5.	5.2%	China, Chinese, Trade, Google, Foreign, Bank, Financial, Company, Benefit, Clinton	US-China Trade Dispute
6.	5%	ASEAN, Malaysia, Trade, Investment, Economic, Korea, Malaysian, Development, State, Opportunities	South Korea-ASEAN Trade related
7.	4.4%	European, Bulgaria, Europe, EU, Bulgarian, Presidency, Digital, Innovation, Sofia, Andonov	EU Tech News
8.	4.4%	North, Trump, Kim, Sanction, Denuclearization, Korea, President, House, Korean, White	Trump-Kim Summit Meet
9.	4%	Cloud, Microsoft, AI, Compute, Samsung, Data, Service, Technology, Build, Rank	New R&D in Tech
10.	4%	Cooperation, Korea, Philippines, South, Korean, Moon, Minister, Peace, Trade, Seoul	South Korea-Philippines Cooperation

There is no overlapping or intersecting in UK topics. As visible on the above figure's right side, 'Huawei' is the most frequent term during our study. This confirms that 'the Huawei 5G issue' is the most extensively reported topic in the UK media in the Cybersecurity area.

4.3.1.6 NMF Topic Modeling for the US

- (A) **Optimal Number of Topics** – With 1615 articles, the US dataset is the largest in our corpus of six countries. When the w2v model (500 dimensions) is created using the parameters $k_{min} = 5$ to $k_{max} = 30$, the resultant model has the largest terms (17,966). With this w2v model, $k = 25$ with the coherence score of 57.82% turn out to the best fit. Figure 4.12 shows the mean coherence and number of

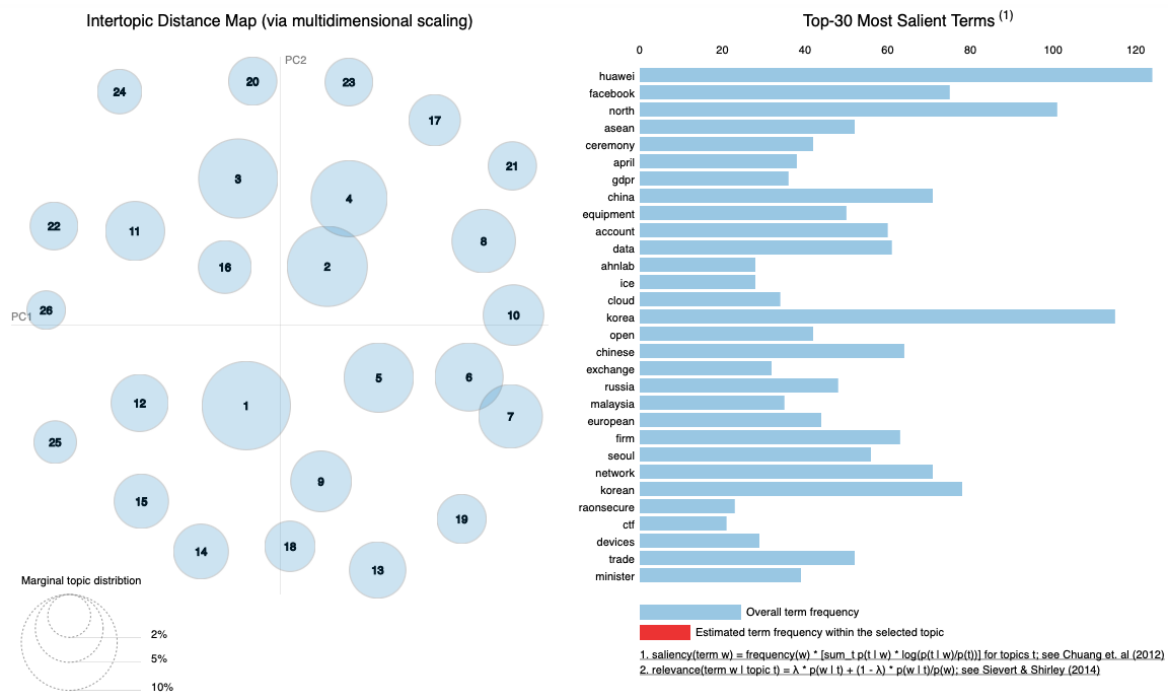


Figure 4.9: All Topics in MDS (South Korea)

topics for the US dataset.

- (B) **Topics, Visualization, and Analysis** – Above figure shows that $k = 25$ is the optimal number of topics for the collected Cybersecurity dataset from the US. Using this number, next, we created the NMF model to understand what these topics are. Table 4.7 shows the top 10 topics with topic weightage, keywords, and topic label.

Four topics in the US dataset are related to elections in one way or another. These topics 4, 5, 6, and 7. While topics 4 (US election campaigns) and 5 (Election and Voting) are about election campaigns and voting-related, topics 6 (Russian Interference in 2016 US election) and 7 (Mueller Investigation) are specifically about the 2016 US presidential election. 'Cambridge Analytica' and 'Stocks and Market' related topic came in top two positions with 7.3% and 6.6% tokens. Other top topics are: 'Disinformation

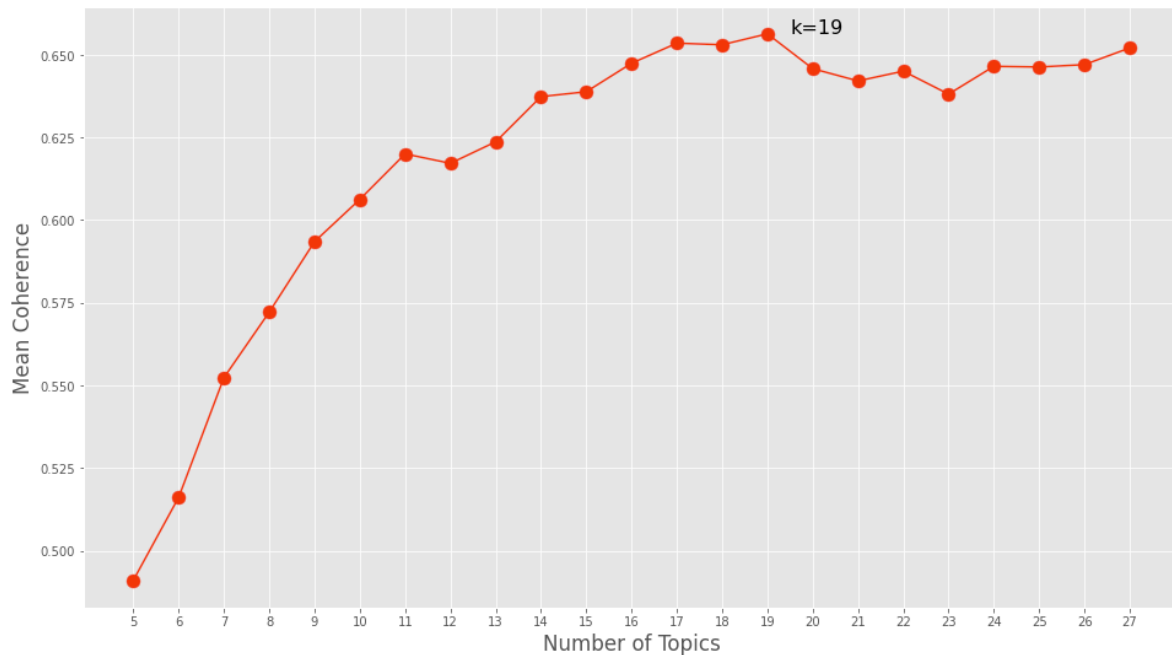


Figure 4.10: Mean Coherence and Number of Topics (UK)

and Fake news on Facebook,’ ‘US-China Trade Dispute,’ and ‘Cyberattack on Target.’ One thing that stands out in the US dataset is the absence of ‘the Huawei 5G issue’ in the top ten topics. Figure 4.13 shows the topic distribution of US topic modeling.

The right side of the above figure shows that ‘Company,’ ‘Facebook,’ and ‘Trump’ are some of the most frequent terms in the US dataset. In comparison, the term ‘Huawei’ is comparatively less in number. This is a change from the previous five countries where ‘Huawei’ is the most frequent term.

4.3.1.7 Descriptive Analysis of Topic Models in Cybersecurity Corpus

Topic models in section 5.1 highlighted the issues or themes that are present in our corpus. Two issues are common in all six countries: 1) Russian interference in the 2016 US election and related topics such as the Mueller investigation and 2) Huawei 5G issue and related topics such as the US-China Trade dispute. Other issues, such as North

Table 4.6: UK Top 10 Topics

No.	Topic Weightage (%)	Keywords	Topic Label
1.	9.1%	Trump, President, White, House, Nielsen, Giuliani, Donald, Administration, Mueller, Washington	Mueller Investigation
2.	8.5%	Attack, North, Cyber, Hackers, Korea, Target, Korean, Security, Group, Hack	Cyberattack involving North Korea
3.	8.2%	Russian, Election, Campaign, Russia, Officials, Elections, Intelligence, Clinton, Hackers, Email	Russian Interference in 2016 US Election
4.	7.7%	Data, Breach, Passwords, Information, Card, Email, Personal, Company, Customers, Access	Various Data Breach incidents
5.	7.5%	Huawei, Chinese, Equipment, Network, Company, Security, Mobile, Telecom, Government, Telecoms	Huawei 5G Issue
6.	6.2%	China, Chinese, Trade, Beijing, Australia, United, Xi, Foreign, States, SCO	US-China Trade Dispute
7.	6.1%	Vote, Ballot, Machine, State, Election, Paper, Georgia, Voter, Voters, Elections	US Voting Machine Security
8.	5.8%	Scam, Apps, Whatsapp, App, Message, Users, Online, Click, Pay, Email	Various Online Scams
9.	5%	GRU, Russia, Russian, Attack, Cyber, Chemical, OPCW, Salisbury, Weapons, Dutch	Cyberattack involving Russia
10.	4.5%	Google, AI, Search, Project, Engine, Pichai, China, Dragonfly, Military, Employees	Google's AI Project

Korea-related topics, Fake page removal from Facebook, also exist in most countries' datasets. Another significant discovery from the topic models is that all the countries except the US extensively reported on the issues that are related to the US. For example, topics such as the US-China trade dispute, Russian interference in the 2016 US election, and US elections and security are found a place in the top ten position. However, US media only focused on domestic issues. No topic is discussing other countries' issues in the top ten topics in the US dataset. This points towards the importance of the US or US's influence globally (generally) and in the cybersecurity area (specifically).

Further, the right side of MDS visualization also points towards an important trend. For countries other than the US, the term 'Huawei' is the most frequent (the longest

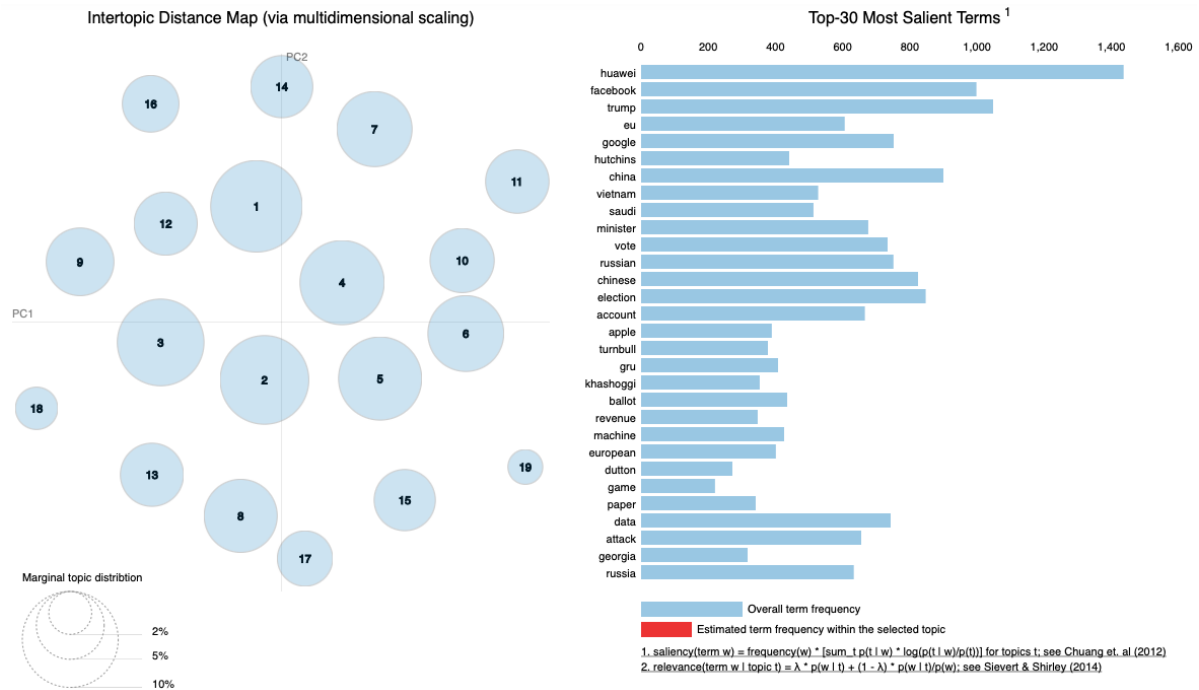


Figure 4.11: All Topics in MDS (UK)

blue bar). However, this is not true in the US's case even though it was the main party in the 'Huawei 5G issue' as the opposition was started and spearheaded by them. For the US, this issue was part of the bigger US-China trade dispute. While US media focused on other topics that are critical for them, such as election security, Russian interference, and hacking incidents, the foreign media paid too much attention to 'the Huawei 5G issue.' This observation also points towards the influence of the US globally.

4.3.2 Sentiment Analysis

As mentioned in section 4.2.4 of Chapter 4, we first perform both unsupervised and supervised sentiment analysis methods. Then, compare which produces the best results and the method which generates the best results, we implement that method to carry out the sentiment analysis on the Cybersecurity corpus. The sentiment analysis section

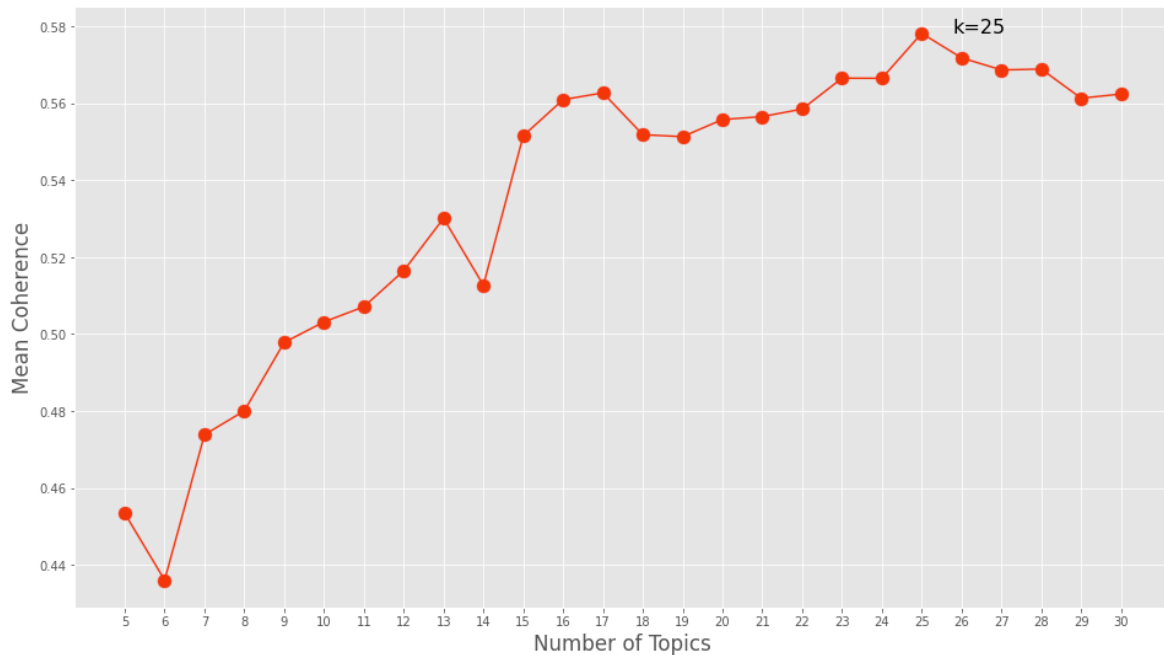


Figure 4.12: Mean Coherence and Number of Topics (US)

help us find answer to RQ 4 and RQ 5 for cybersecurity.

4.3.2.1 Unsupervised (Lexicon-based) Sentiment Analysis

Unsupervised methods are useful when data is not labeled. Also, it is not a prerequisite to have training data for these methods. However, as there is no training data, it is impossible to predict the accuracy of the results, which works as one of the most significant limitations of unsupervised methods. Since one of the objectives of this section is to find out the method that produces the best result, we would use the IMDB movie review dataset to predict the accuracy. As mentioned, we used three of the most popular Python-based libraries Textblob, VADER, and SWN. Table 4.8 shows the results from these libraries.

Textblob gave the best overall accuracy (69.44%) out of all three, but VADER and SWN were also not far with 68.63% and 68.05%, respectively. However, if we look

Table 4.7: US Top 10 Topics

No.	Topic Weightage (%)	Keywords	Topic Label
1.	7.3%	Privacy, Data, Company, Facebook, Users, Tech, GDPR, Law, Analytica, Cambridge	Cambridge Analytica
2.	6.6%	NYT, Company, Investors, Bank, market, Stock, Fund, WSJ, Trade, Bloomberg	Stock Market News
3.	6.3%	Women, Times, People, Reid, Get, Men, Write, York, Work, Go	Me too Movement
4.	5.5%	Democrats, Republican, Committee, Campaign, Democratic, Republicans, District, House, Congress, Party	US Election Campaigns
5.	5.5%	Election, State, Vote, Ballot, Machine, Elections, Voter, Paper, Officials, Voters	US Elections and Voting
6.	5.5%	Russian, Russia, Election, Campaign, Intelligence, Interference, Russians, Putin, Influence, American	Russian Interference in 2016 US Election
7.	5.4%	Trump, Putin, President, Meet, Mueller, Clinton, Cohen, Comey, Russia, Summit	Mueller Investigation
8.	5.2%	Attack, Hackers, Target, Hack, Malware, Ransomware, Systems, Microsoft, Computer, Email	Cyberattack on Target
9.	5.2%	Facebook, Account, Page, Social, Post, Fake, Media, Disinformation, Twitter, Instagram	Disinformation and Fake news on Facebook
10.	5%	China, Chinese, Trade, Beijing, United, States, Taiwan, Xi, Economic, Administration, American	US-China Trade Dispute

Table 4.8: Showing accuracy in different sentiment analysis libraries

Library	True Positive (25000)	True Negative (25000)	Overall Accuracy (%)
Textblob	23180 (92.7%)	13198 (52.8%)	69.44%
VADER	21117 (84.4%)	11543 (46.2%)	68.63%
SWN	16694 (66.7%)	17330 (69.3%)	68.05%

at the true positive and true negative, we find a different picture. Textblob produced extremely good results (92.7% accuracy) with positive reviews, and SWN gave 69.3% accuracy (highest in all three) with negative reviews. This experiment also showed that some of these libraries are good at predicting positive sentiments while others are good in negative sentiments.

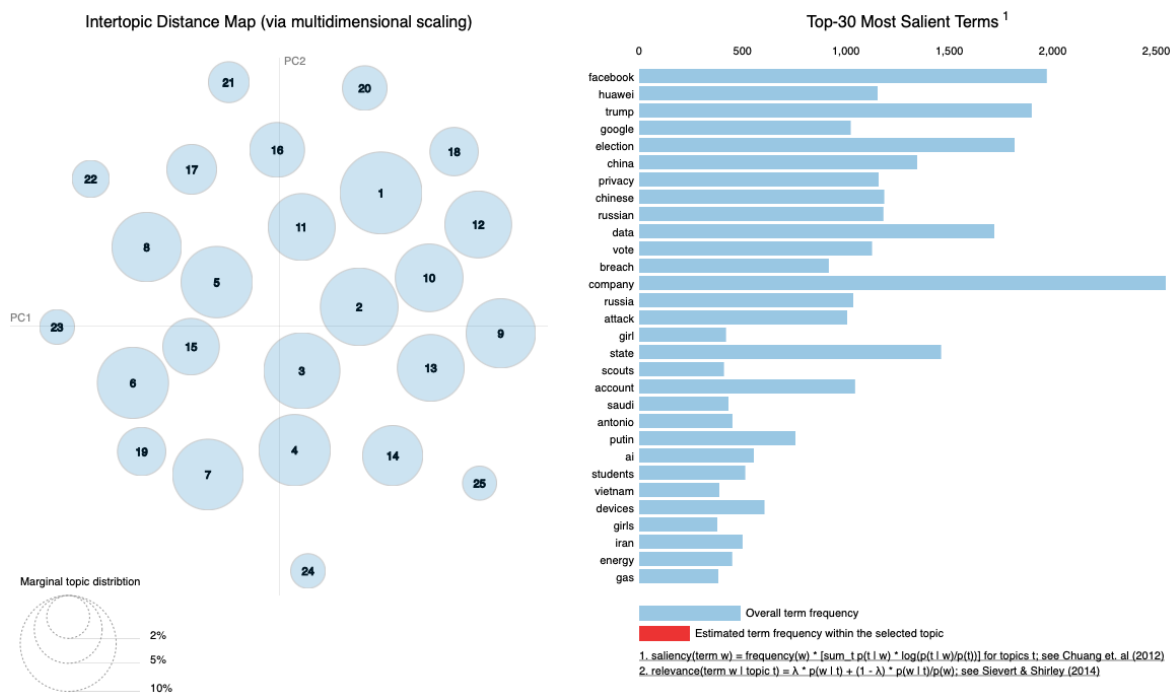


Figure 4.13: All Topics in MDS (US)

4.3.2.2 Supervised Sentiment Analysis

For supervised sentiment analysis, we have used a Python-based scikit-learn ML library [77]. To apply the ML algorithms on any dataset to successfully predict correct sentiments, it is critical to extract essential features. We have used N-gram (unigram, bigram, and combination of unigram and bigram) feature extractions algorithm in this research. N-gram can capture (to some extent) the context of textual data. As for classifiers, there are numerous options available, and out of that, we experimented with four classifiers: Multinomial Naïve Bayes (MNB), LinearSVC, Logistic Regression (LR), and Passive Aggressive (PA). Table 4.9 shows the performance of different classifiers with different features on the IMDB dataset.

From table 4.9, it is clear that LR, with the combination of unigram and bigram feature, gave the best results with an F1 score of 91.36%, and LinearSVC came a

Table 4.9: Evaluation Matrix for different classifiers and features

ML Classifier	Dataset	IMDB		
	Feature Extraction	Unigram	Bigram	Unigram & Bigram
MNB	Accuracy	86.21	88.43	88.82
	Precision	87.22	91.40	90.17
	Recall	84.79	84.79	87.09
	F - Score	85.99	87.97	88.60
LinearSVC	Accuracy	89.11	88.30	91.13
	Precision	88.49	86.73	89.97
	Recall	89.86	90.38	92.54
	F - Score	89.17	88.52	91.24
LR	Accuracy	89.45	88.45	91.26
	Precision	88.43	87.04	90.15
	Recall	90.72	90.30	92.60
	F - Score	89.56	88.64	91.36
PA	Accuracy	86.70	88.51	91.11
	Precision	86.57	87.30	90.12
	Recall	86.81	90.08	92.30
	F - Score	86.69	88.67	91.20

close second with 91.24%. When compared with the unsupervised methods, supervised methods showed very high performance. Based on these results, we choose LR with the combination of unigram and bigram features to predict sentiments of our dataset (cybersecurity news articles).

4.3.2.3 Predicting Sentiments of Cybersecurity News Articles and Interpretation of Results

For sentiments analysis, we have considered 1) results of topic modeling, 2) our understanding of cybersecurity scenarios worldwide, and 3) the time of the collected dataset.

On that basis, we have chosen five broad themes. These themes are China (Huawei), Russia (Election interference), Iran (One of the major players in the cyber field, especially cyber warfare), North Korea (Notorious for many state-backed cyber hacking incidents), and Vietnam (New Cybersecurity Law).

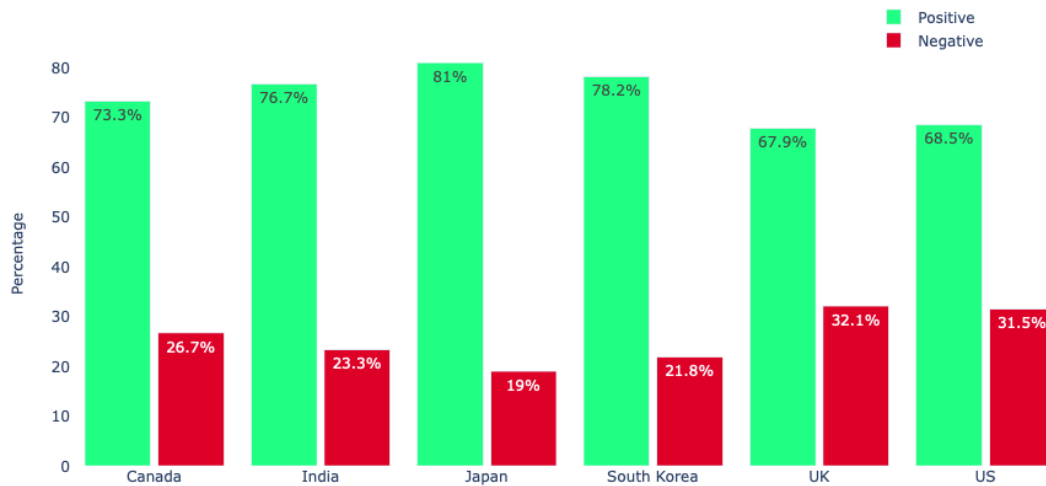


Figure 4.14: **Sentiment Analysis results (in percentage) on our cybersecurity news articles dataset**

Figure 4.14 shows that while the UK and the US have the highest percentage of negative articles, 32.1% and 31.5%, respectively, Japan, with 19%, comes last in this category. On average, 74% of articles are positive, and 26% are negative. It also confirms that negative news does not always mean negative sentiments.

4.3.2.3.1 Sentiments of China and Huawei related cybersecurity news articles and Discussion

Before analyzing sentiments of China and Huawei related articles, it would be beneficial to know how many articles belong to this theme in our dataset.

Figure 4.15 shows that Japan and Canada have the highest percentage of articles

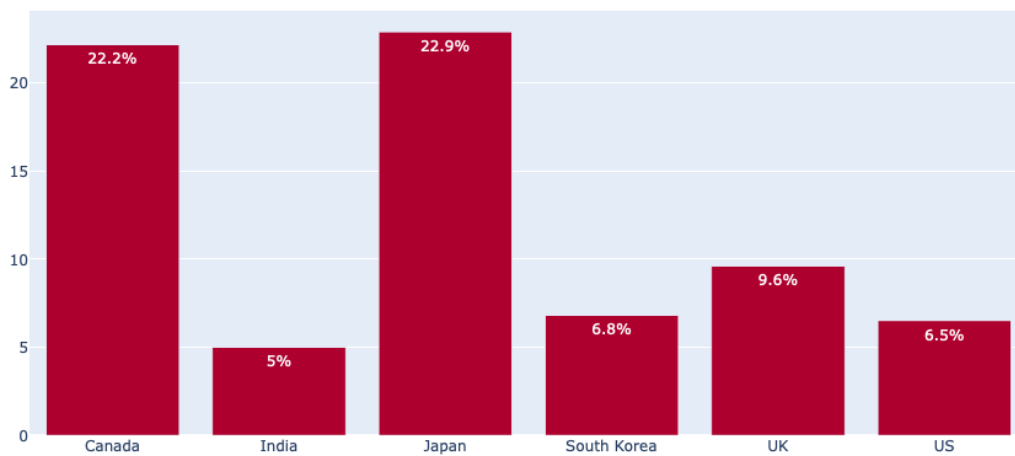


Figure 4.15: **China and Huawei related articles in our dataset**

related to this theme. Numbers from Canada are understandable because Canada arrested Huawei's CFO, Meng Wanzhou. However, finding fewer (only 6.5%) articles in the US is very surprising as the US is one of the main parties in the US-China trade dispute and banning Huawei 5G technology.

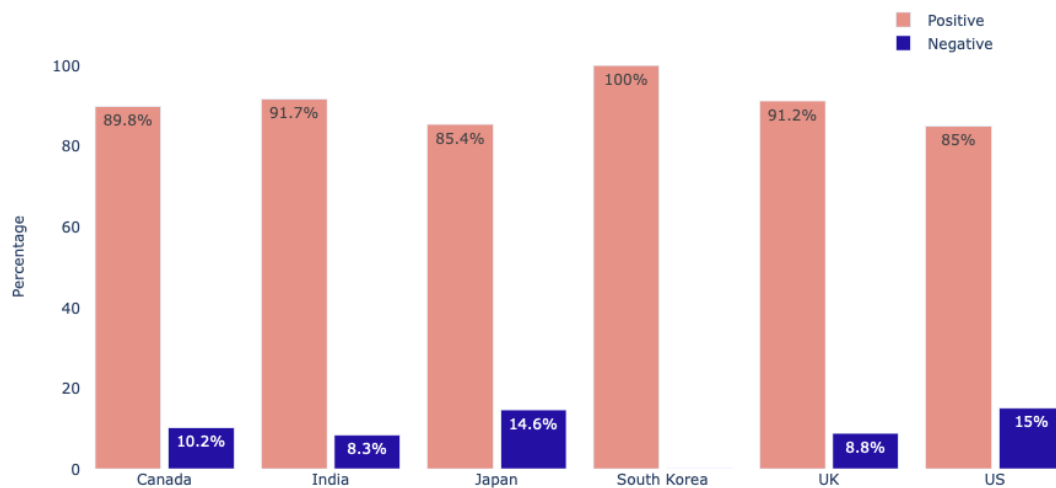


Figure 4.16: **China and Huawei related articles sentiments**

Figure 4.16 US and Japan have the highest percentage (15% and 14.4% respectively)

of negative articles. As South Korea has a smaller number of articles, 100% positive articles are not surprising. The average of negative articles in this theme (9%) is less than the overall average of 26%. The above results are unexpected, especially when the US (particularly Donald Trump) tried their best to play with the national sentiments by portraying Huawei and its 5G technology as a threat to national security. However, as 5G technology is the future and Huawei is one of the leaders in this field, many countries (especially EU countries) defy the US pressure (blackmail) and take decisions based on their national interest. Moreover, the US intelligence failed to provide substantial evidence to back their claims about the danger of Huawei technology ³, ⁴. The results showed that Donald Trump's negative campaign against Huawei was a failure, as it did not even garner negative sentiments from the US media.

4.3.2.3.2 Sentiments of Russia and Election-related cybersecurity news articles and Discussion

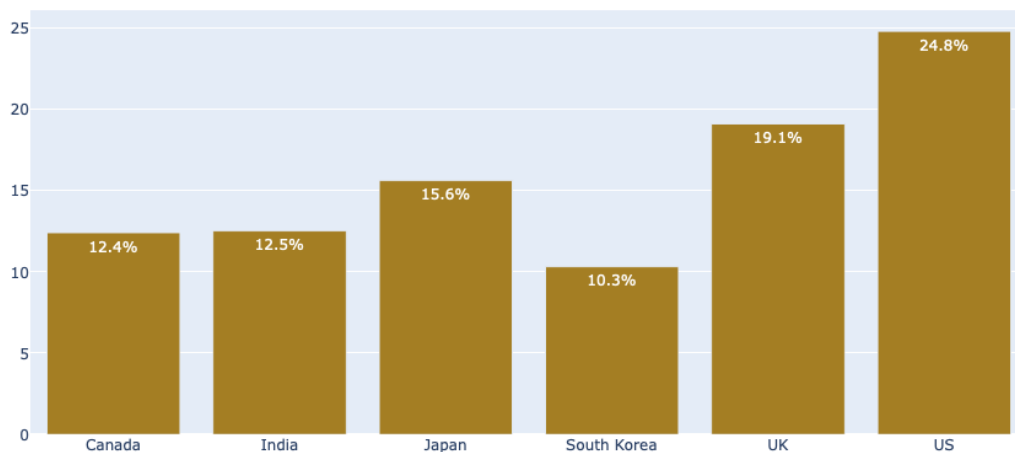


Figure 4.17: **Russia and Election-related articles in our dataset**

³<https://www.latimes.com/projects/la-fg-huawei-timeline/>

⁴<https://www.tomsguide.com/us/us-huawei-ban-op-ed,news-30132.html>

US dataset, with 24.8%, has the largest percentage of articles related to this theme. This shows that the US media prioritize news related to the Russian interference in the 2016 US election. This is interesting, as though the Mueller investigation was going on during this time. Compared to the Huawei issue that came into prominence only during the specific time of our study, this issue is almost two years old. Other than the US, the UK and Japan had the second and third highest percentage of articles.

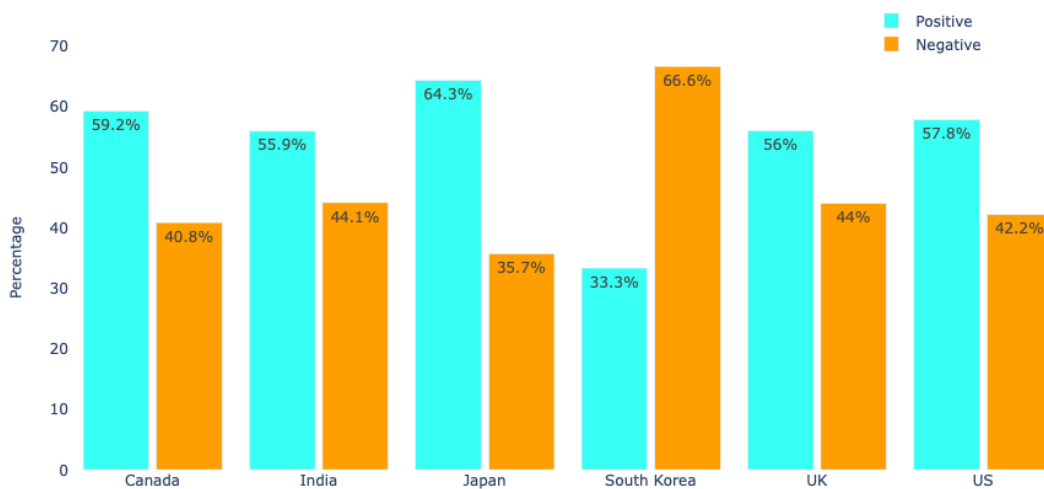


Figure 4.18: **Russia and Election-related articles sentiments**

Figure 4.18 shows that this theme attracted negative sentiment in all countries, with India and the UK showed the highest percentage of negative articles, 44.1% and 44%, respectively. The US, with 42.2% and Canada with 40.8%, was also not too far behind. The average of negative articles for this theme is astoundingly high (48.7%) than the overall average of 26%. All these countries are democracies, and for democratic nations, the election is one of the most valuable assets. However, the growing web of fake news and disinformation campaign used by political parties to defame their opponent has seen an unprecedented increase. A recent Oxford study showed that at least 70 countries have such disinformation campaigns [78]. Russian interference received wide attention because it happened to the US – the most powerful and the oldest democracy globally

– and nations, especially citizens across borders, wonder about the validity and security of elections in their own countries.

4.3.2.3.3 Sentiments of Iran related cybersecurity news articles and Discussion

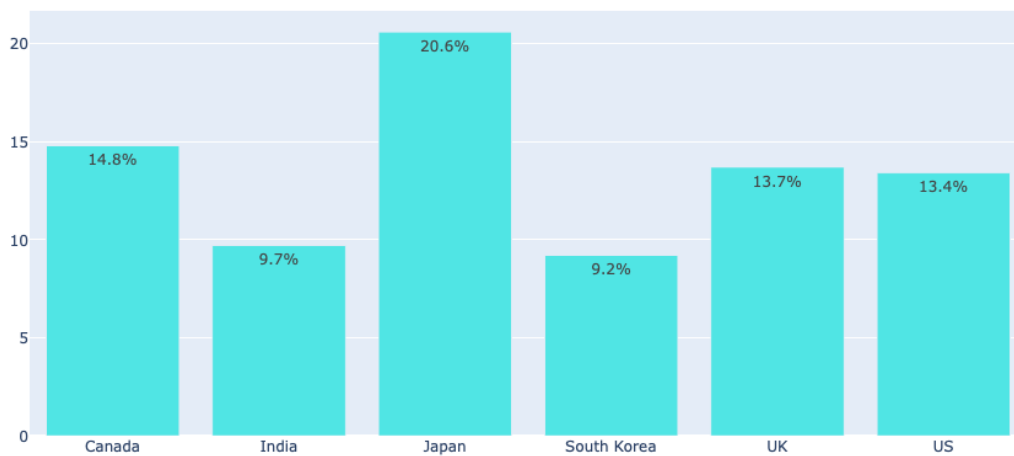


Figure 4.19: Iran related articles in our dataset

Figure 4.19 shows that Iran acquired considerable space during the research period, with Japan having the largest percentage (20.6%) of articles, and the US, Canada, and the UK obtained around 14% of articles. Iran is a major player in cyberwarfare capabilities and is always in the news for cyber-related issues, along with North Korea. The Middle East is an extremely significant region for Japan because of its crude oil requirement, and because of that, Japan keeps a close eye on the happenings in that region. In May 2018, the US withdrew from the Nuclear deal with Iran and imposed fresh sanctions; it became a piece of concerning news. That can be cited as the probable reason for a high percentage of Iran-related articles in Japan’s dataset.

Figure 4.20 shows that South Korea, with 37.5%, has the highest negative articles percentage. After that, Britain, the US, and India all have around 27% of articles with

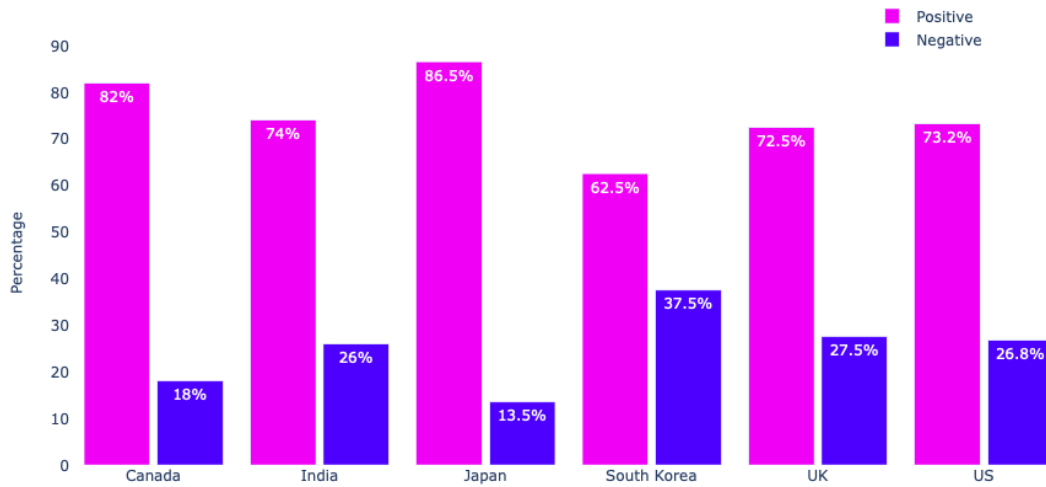


Figure 4.20: Iran related articles sentiments

negative sentiments. The negative articles average for this theme is 25%, almost equal to the overall negative average of 26%. As mentioned earlier, in May 2018, the US, under Donald Trump’s leadership, withdrew from Iran Nuclear Deal⁵. After this, the security experts warned that Iran might retaliate with cyber-attacks as a similar pattern was seen in the past when Barack Obama imposed sanctions in 2012 (increased cyber-attacks) or concluded the nuclear deal in 2015 (Iran restrained from state-sponsored digital attacks)⁶. These reports might explain the negative sentiments against Iran during the chosen period of this study.

4.3.2.3.4 Sentiments of North Korea related cybersecurity news articles and Discussion

Figure 4.21 shows that South Korea (13.8%) and Japan (13.4%) have the highest percentage of articles related to North Korea. The US, with 9%, came at the third position. North Korea is notorious for its state-sponsored cyber-attacks. As mentioned

⁵<https://www.nytimes.com/2018/05/08/world/middleeast/trump-iran-nuclear-deal.html>

⁶<https://www.wired.com/story/iran-nuclear-deal-cyberattacks/>

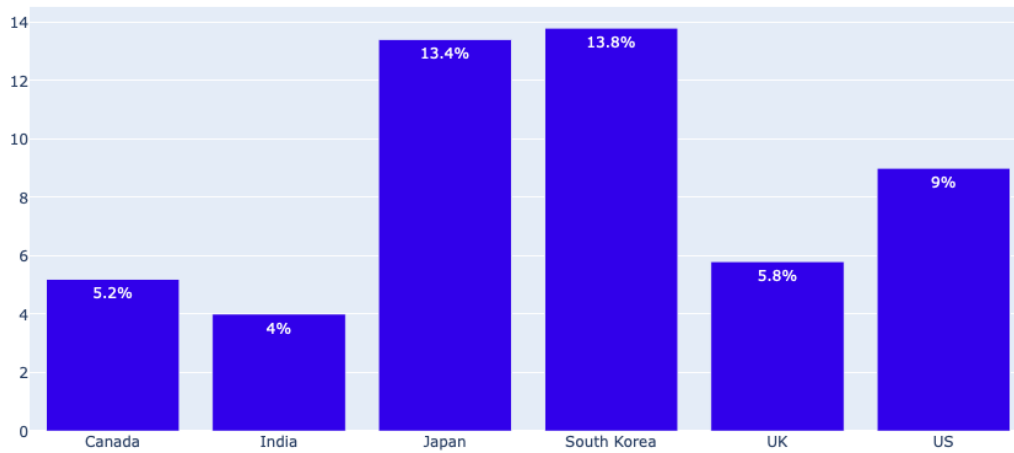


Figure 4.21: North Korea related articles in our dataset

earlier in this research, South Korea is at the receiving hand of these attacks more than any other country. Sony (Japanese company) hack of December 2014 also alerted Japan that it (Japan's government or companies) can also be the target of North Korea-sponsored cyber-attacks in the future⁷.

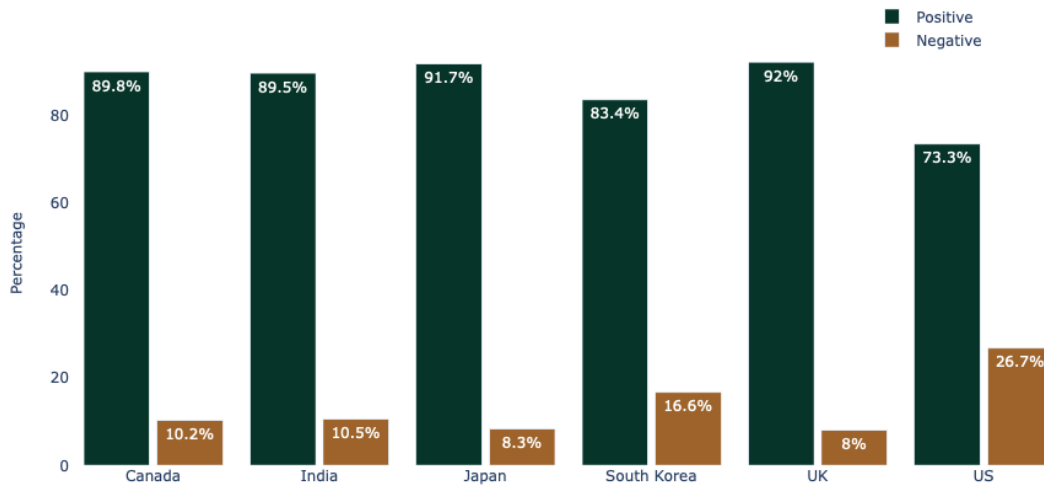


Figure 4.22: North Korea related articles sentiments

⁷<https://www.telegraph.co.uk/news/2019/05/23/north-korea-hits-japan-cyber-arms-race-heats/>

Figure 4.22 shows that with 26.7%, the US has by far the highest negative percentage in this theme, and South Korea, with 16.6%, came at the second position. The average 13.6% negative articles are lesser than the overall average of 26%. In April 2018, top leaders from South and North Korea met and signed the "Panmunjom Declaration," in which North Korea pledged to "cease all hostile acts... in every domain."⁸ Just two months later, in June 2018, the first-ever summit meeting between North Korea (Kim Jong-Un) and the US (Donald Trump) was held in Singapore. This meeting gave a cautious hope that this might led North Korea to stop its rogue behavior and would open the path of the possible peace deal⁹. However, at the same time, there were reports that Pyongyang continued to engage in cyber-attacks against South Korea¹⁰. Earlier mentioned Fireeye report on APT38 and another report titled "Kim Jong Un's "All Purpose Sword" North Korean Cyber-Enabled Economic Warfare" talked in detail about how heavily sanctioned, and cash-strapped North Korea uses cyber-attacks to generate illicit funds¹¹. It is understandable from this explanation about the high negative sentiments in South Korea and the US against North Korea.

4.3.2.3.5 Sentiments of Vietnam related cybersecurity news articles and Discussion

As visible in Figure 4.23, this theme has the lowest percentage of articles out of the other four themes. Japan and UK have the highest (7.3%) and the second-highest (6%) percentage. Vietnam is neither an advanced nation in the field of Cybersecurity nor a notorious country like North Korea. As mentioned, the reason that Vietnam garnered

⁸<https://www.mofa.go.kr/viewer/skin/doc.html?fn=2018091804122336rs=/viewer/result/202002>

⁹<https://abcnews.go.com/International/world-watched-trump-kim-summit-drew-global-reaction/story?id=55835926>

¹⁰<https://www.wsj.com/articles/north-korea-while-professing-peace-escalated-cyberattacks-on-south-1527239057?mod=e2twatesla=y>

¹¹https://www.fdd.org/wp-content/uploads/2018/09/REPORT_NorthKorea_CEEW.pdf

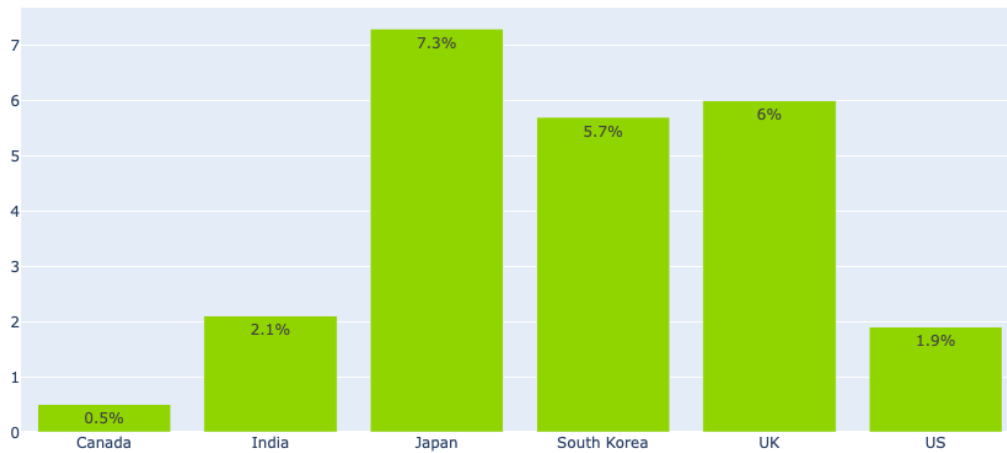


Figure 4.23: Vietnam related articles in our dataset

some attention during this period is due to its passing of the new cybersecurity bill.

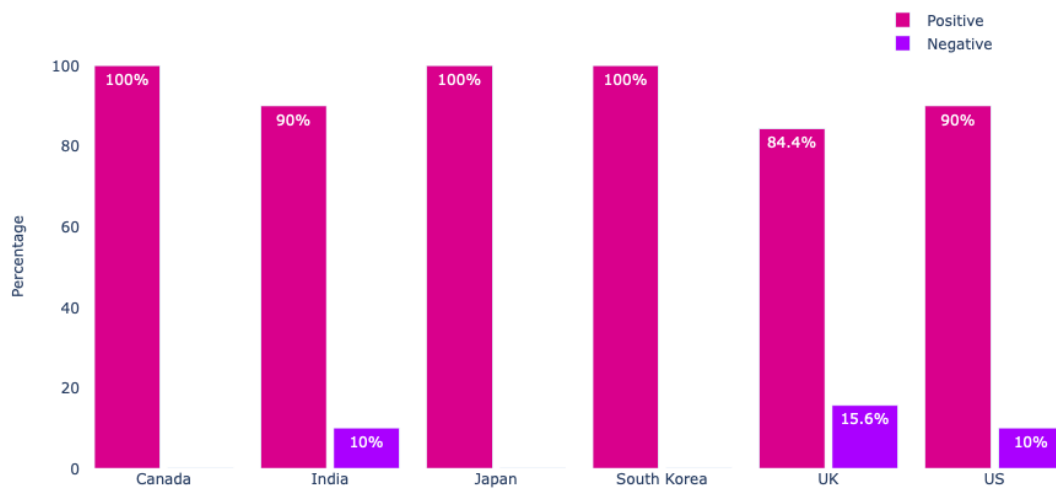


Figure 4.24: Vietnam related articles sentiments

The negative sentiment in Vietnam-related cybersecurity articles is the lowest among the five themes. Britain (15.6%) highest, and the US and India with 10% have the second-highest negative percentage. These pockets of negative sentiments are only

when there is a criticism of the new cybersecurity bill¹².

We compared unsupervised and supervised methods and found out that supervised methods achieved high accuracy. Then our sentiment classification showed that the general sentiment of Cybersecurity news articles is positive. One reason for that can be the type of chosen text (articles). As news articles are objectively written, they have less sentiment content. However, even after that, Russian election interference received exceptionally high negative sentiments compare to other issues or themes. This result shows that world media, particularly in democratic nations, is on the same page on this issue, undermining one of the most sacred processes in democracies - elections.

4.4 Limitations

Data acquisition is the most important step in this research as it is unthinkable to start this research without getting enough credible data. It is exceedingly difficult to get the latest news articles from different newspaper websites. The issue of limited and paid access, changing the structure of websites and language are some of the problems we faced in data collection. Further, as the data (news articles) is comparatively more objective than tweets or other social media posts, the neutral content is higher. For this reason, multi-class classification (positive, negative, and neutral) could have produce better results. However, because of the unavailability of a suitable labeled dataset, this research used the 'IMDB dataset' as a training and testing dataset that is binary classified. Hence, the results are either positive or negative. This point is also one of the limitations of this research.

¹²<https://www.cyberdb.co/vietnam-new-cyber-security-law/>

Chapter 5

Topic Modeling and Clustering

Approach to Analyze COVID-19

News

5.1 Motivation

Natural Language Processing (NLP) and its various techniques have gained prominence for processing and analyzing a large amount of natural language data. NLP can be defined as a field that combines linguistics and artificial intelligence (AI) to enable computers to understand human or natural language. In present times, the criticality of NLP is further increased from the very fact that we are generating an enormous amount of unstructured text data every day. Some of the most common and popular NLP techniques include named entity recognition [79], sentiment analysis, machine translation, topic modeling, and text summarization [80]. In the current COVID-19 crisis, one critical asset is to identify as much information about the problem as possible. Any available information, which can be located, amassed, and understood, will facilitate the right decisions to be taken. NLP enables researchers to access information

on topics such as global and regional spread, the socio-economic impact of the disease, vaccine development, patient demographics and co-morbidities, and the national and international politics from sources such as scientific literature, policy documents, social media, and news¹.

Newspapers (both print and digital) worldwide are full of COVID-19/Coronavirus-related news, articles, and stories. Making sense of this data is a challenge but a challenge worth doing. As mentioned in detail in chapter 3, Topic Modeling is an approach used for automatic comprehension and classification of data in various settings. We would use the recently proposed Top2Vec algorithm - in understanding COVID-19 related news in India, Japan, South Korea, and the UK. Next, the results of the top2vec model are used in clustering (K-means) to classify them further. We would also compare the results of the top2vec model with the more established topic modeling algorithm – NMF. This research collected more than 100,000 COVID-19/Coronavirus-related news articles from January 1, 2020, till December 1, 2020) from major newspapers (digital version) from four countries and tried to analyze them using the topic modeling approach. Our research aims to identify various topics (hence issues) in the COVID-19/Coronavirus-themed news articles, helping us decluttering this large dataset.

5.2 Research Methodology

The first step to begin the COVID-19 case study is to collect the relevant data. Hence, data acquisition is the first step in the research methodology for this chapter. Then based on the requirement text preprocessing² would be performed (top2vec does not require, but NMF needs preprocessing). The next steps would be divided based on the three main methods: 1) Top2Vec, 2) NMF, and 3) K-means clustering. Figure 5.1

¹<http://www.copyright.com/blog/natural-language-processing-information-covid-19/>

²The steps for text preprocessing are similar to what we used previously; hence we will not discuss them here. Please refer to section 4.2.2 of chapter 4 for a detailed explanation.

shows the outline of the research methodology for this chapter.

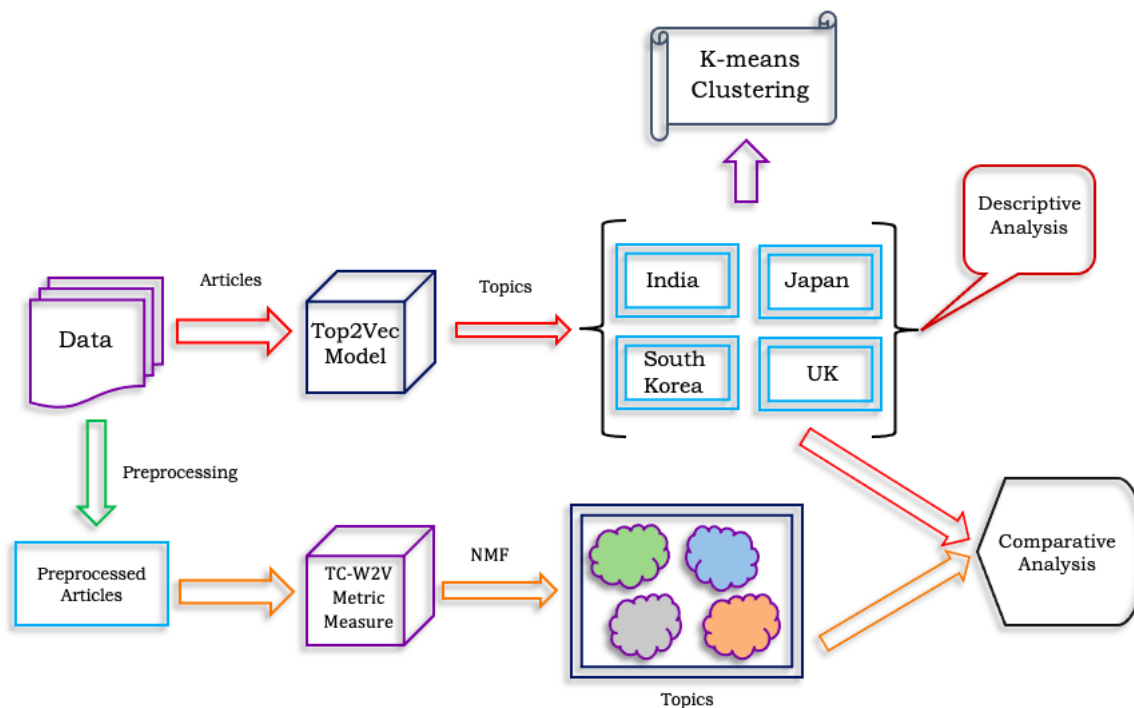


Figure 5.1: **Research Methodology for COVID-19 Topic Modeling and Clustering**

5.2.1 Data Acquisition

We searched and scraped articles with COVID-19 or Coronavirus keywords in the English language websites of eight major newspapers from four countries. The period for collecting both news headlines and articles is from January 1st, 2020, to December 1st, 2020. For web-scraping, we used the BeautifulSoup Python library. After removing the duplicate articles, the remaining dataset consisted of 102,278 articles. In this dataset, India has the largest number of articles, 47,342, and South Korea has the lowest number of articles (10,076). We collected 23,821 and 21,039 articles from UK

and Japan, respectively. Table 5.1 shows the newspapers and the collected COVID-19/Coronavirus-related articles.

Table 5.1: **Data Collected for COVID-19 Topic Modeling and Clustering**

Country	Newspapers	No. of Articles
India	Hindustan Times, The Indian Express	47,342
Japan	The Japan Times, Asahi Shimbun, Mainichi Shimbun	21,039
South Korea	Korea Herald, Korea Times	10,076
UK	The Daily Mail	23,821
Total		102,278

5.2.2 Top2Vec and K-means Clustering

Though `top2vec` provides an option to use pre-trained embedding models such as universal-sentence-encoder to generate joint word and document embeddings, we use our data to create topic models. We used the 'deep-learn' parameter, which generates the best quality vectors but takes significant time to train. We train the `top2vec` model for each country dataset separately. For example, it took around 16 hours on our system to train the `top2vec` model on the Indian dataset (largest dataset). Once the model is produced, it provides various information such as topic size, topic words, topic number, topic score, etc. Topic size shows the number of documents most similar to each topic, while for each topic, the top 50 words (we have only presented the top 10 words in Table 5.3, 5.4, 5.5, and 5.6) are returned, in order of semantic similarity to the topic. Topic score – for each topic is the cosine similarity to the search keywords. The higher the topic score, the most representative that topic would be for the searched keyword. This feature of `top2vec` is used in this research to find the most representative topic using keywords. The word cloud in Figure 5.2, 5.3, 5.4, and 5.5 are also generated using the `top2vec` semantic search feature. Lastly, based on our understanding of the topic's

specific keywords (all 50 words), we labeled these topics.

Once we finish explaining the top ten topics, to further our understanding of these topics' nature and see how closely or sparsely these topics, we would apply the k-means clustering algorithm on the topic vectors produced by the top2vec model. Like other unsupervised machine learning clustering methods, k-means clustering also requires the researcher to know the optimal number of clusters previously. For finding the optimal number of clusters, we used the Davies Bouldin index [81].

5.2.3 NMF Topic Modeling with TC-W2V Metric Measure

For NMF topic modeling, we first find the optimal number of topics using the TC-W2V metric measure. After discovering the best number of topics, the NMF topic model is produced. Then we would analyze the top ten topics and compare the topics produced from NMF with top2vec.

5.3 Experiment, Results, and Discussion

5.3.1 Top2Vec Topic Modeling

As mentioned before, the top2vec model does not require preprocessing and provides the option to either train a top2vec model on your dataset or use a pre-trained model universal-sentence-encoder. For this research, we trained the top2vec model on our collected dataset of each country individually. Further, top2vec has three parameters to determine the speed - fast-learn, learn, and deep-learn. Fast-learn is the fastest but provides the lowest quality results, while deep-learn takes the longest time to train but provides the best results. We trained our top2vec model on deep-learn parameters. Table ?? shows the results of the top2vec model on our dataset. This section help us find answer to RQ 1, RQ2, and RQ3 for COVID-19 pandemic.

Table 5.3: India's Top 10 Topics with Topic Size, Topic words, and Topic Label

No.	Topic Size	Top 10 Words	Topic Label
1	1278	Cricket, ODI's, ODI, Batsman, ESPNcrinfo, Cricketing, IPL, Cricketers, Pacer, Overs	Cricket and IPL
2	658	Trump, Democrat, Americans, Republican, Biden, Arizona, Delaware, Carolina, Florida, Donald	US related
3	623	Ludhiana, Jalandhar, Faridkot, Gurdaspur, Hoshiarpur, Amritsar, Sangrur, Bagga, Patiala, Moga	Punjab related news
4	610	Accused, Complainant, Murder, Allegedly, SHO, Arrested, IPC, FIR, Complaint, Incident	Law and Order during Pandemic
5	605	Vaccine, Astrazeneca, Vaccines, Doses, Oxford, Trails, AZD, Moderna, Pfizer, mRNA	Vaccine Development
6	500	Classes, Classroom, Teachers, Learning, Worksheets, Teaching, Curriculum, Schools, Syllabus, School	School Education during Pandemic
7	431	RBI, Contraction, GDP, Liquidity, Monetary, Economy, Shaktikanta, Growth, Outlook, Fiscal	RBI and Economy related
8	423	Apr, Myself, Pdt, Chores, Enjoying, Adds, My, Me, Things, Mom	Celebrities on Social Media During Lockdown
9	388	Semester, UGC, Varsity, Universities, Examinations, Semesters, Academic, Varsities, Chancellors, Exams	University Education during Pandemic
10	380	ICMR, PCR, RT, Polymerase, Confirmatory, Rapid, Diagnostic, Antigen, Laboratories, Testing	Covid-19 testing

related on topic nine), Vaccine development, Reserve Bank of India (RBI) economic policies, and COVID-19 testing. US presidential election was planned for November 2020, and along with the COVID-19 pandemic, the election campaign was also going on. The US is the leading global power, and its election results impact the whole world. Because of this reason, global media keeps a close eye on the US election, and Indian media is no exception to that. Indian drug companies are major manufacturers of vaccines distributed worldwide and supply more than 60% of the vaccines to the developing world [82]. Hence, it is understandable to see vaccine development-related

news in the top ten topics. Table 5.3 presents the top 10 topics with the number of articles (topic size), top 10 words, and topic label for India. Figure 5.2 shows the word cloud for the 5th biggest topic – Vaccine development.

5.3.1.2 Describing Japan’s Top Topics

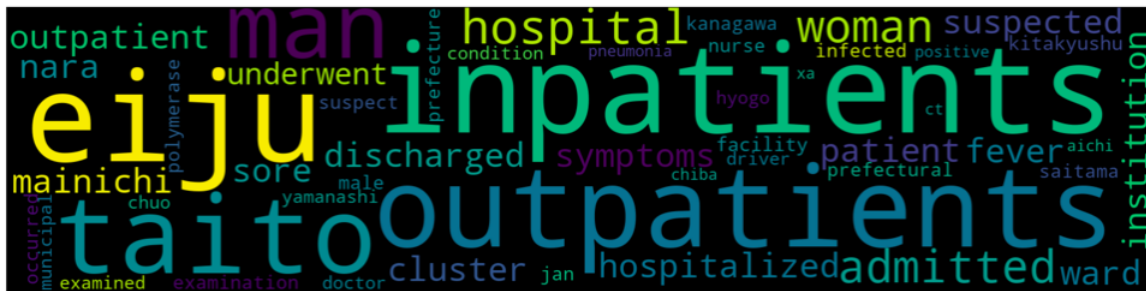


Figure 5.3: Topic 4 in Japan’s Dataset - COVID-19 Cases in Different Prefectures

Our top2vec model came up with 255 topics in Japan’s dataset. News about global stock exchanges and markets such as Dow Jones and NASDAQ was widely reported in Japanese media during this pandemic. It especially came upon the top spot with 409 articles. Other top topics include Nikkei (Japan’s stock exchange) News, postponement of Tokyo 2020 games, North Korea (topic nine), South Korea (topic five), US election, and education-related news. The presence of stock markets news (global market in the first position and Nikkei in the second position) shows Japan’s English language media is paying too much attention to markets and trade even during the pandemic. Japan has invested a huge sum of money preparing for Tokyo 2020 games and expected considerable economic benefits. However, due to the pandemic, the Tokyo 2020 games were postponed; this was a huge setback for Japan. Tokyo 2020 games postponement issue came up in the third position with 363 articles. South Korea is an important neighboring country for Japan. Hence, during the pandemic, South Korea’s news was

widely reported in the Japanese media and came on the fifth position. In contrast, North Korea is seen as a threat by Japan; it keeps a close eye on its activity. The top 10 topics with topic size, top words, and topic labels are shown in Table 5.4. At the same time, Figure 5.3 shows the 4th biggest topic in Japan’s dataset.

Table 5.4: **Japan’s Top 10 Topics with Topic Size, Topic words, and Topic Label**

No.	Topic Size	Top 10 Words	Topic Labels
1	409	Investors, Stocks, Dow, NASDAQ, Gains, Benchmark, Composite, Markets, Wall, Treasury	Global Stock Exchanges News
2	382	Decliners, Section, Shares, Advancers, Yen, Brokers, Outnumbered, Nikkei, Fetched, Unchanged	Nikkei News
3	363	IOC, Bach, Olympics, Olympic, Organizing, Postponement, Athletes, Games, Federations, Paralympics	Tokyo 2020 Games Postponement
4	314	Eiju, Inpatients, Outpatients, Taito, Man, Hospital, Admitted, Woman, Hospitalized, Outpatient	COVID-19 Cases in different prefectures
5	293	KCDC, Seoul, Sye, Daegu, Kwon, Kyun, Eun, Gyeongsang, Kyeong, Itaewon	South Korea News
6	259	Biden, Joe, Nominee, Hillary, Kamala, Harris, Clinton, Sanders, Democrats, Delaware	US Election
7	258	Declaration, Prefectures, Pachinko, Governors, Emergency, Parlors, Hyogo, Requests, Nishimura, Outings	Declaration of Emergency in Japan and related news
8	251	Globalization, Democracies, Societies, Superpower, Technological, Authoritarian, Geopolitical, Multilateral, Nationalism, Alliances	Politics over Coronavirus Pandemic
9	241	Pyongyang, Jong, Kim, Un, North, KCNA, Korean, Kaesong, Koreas, Korea	North Korea News
10	240	Elementary, Education, School, Schools, Teachers, Classes, Hagiuda, Junior, Academic, Students	School Education amid Pandemic

5.3.1.3 Describing South Korea’s Top Topics

Our COVID-19 dataset has the lowest number of articles (10,076) from South Korea, and this produced the lowest number of topics - 127. News related to budget and economic relief package provided during the pandemic in South Korea came up at the

Table 5.5: South Korea’s Top 10 Topics with Topic Size, Topic words, and Topic Label

No.	Topic Size	Top 10 Words	Topic Label
1	284	Supplementary, Budget, Extra, Fiscal, Budgets, Households, Handouts, Earners, Bill, Relief	Various Economic Relief during Pandemic
2	251	Syndicate, Europeans, Eurozone, Economies, Globalization, Geopolitical, Continent, Crises, Policymakers Africa	Geopolitics During Pandemic
3	244	Stores, Shopping, Store, Mart, Shinsegae, Customers, Retailers, Mall, Items, Lotte	Pandemic’s impact on Retail and Shopping
4	210	Monetary, BOK, Contraction, Forecast, Economy, Slashed, Outlook, Shrink, Projection, Growth	Lower GDP forecast Due to Pandemic
5	205	Untraceable, Sporadic, Cluster, Infections, Surrounds, Cases, Traced, Area, Caseload, Digits	Coronavirus Cluster Cases
6	191	League, Matches, Season, Teams, Football, Players, Champions, Stadium, KBO, Match	Various Sports League in Korea amid Pandemic
7	178	Smartphone, Smartphones, Chip, Memory, Dram, Chips, Semiconductor, NAND, Foundry, Huawei	Semiconductor Industry News
8	166	Tribune, Editorial, Americans, Governors, Columnist, Florida, Herd, Sick, Dangerous, Republican	US related news
9	150	Daenam, Cheongdo, Hospital, Woman, Patient, Pneumonia, Daegu, Spreader, Patients, Old	Daenam Hospital Cluster Coronavirus Cases
10	148	Clinical, Trials, Drug, antibody, Celltrion, Drugs, CT, Phase, Treatments, Pipeline	Celltrion’s Vaccine Development

top position with 284 articles. Other top topics include – the pandemic’s impact on retail and markets, geopolitics, low GDP forecast, sports, US, and vaccine trials. Out of the top ten topics, four topics (topic one, three, four, and seven) are economy-related in the South Korean dataset. This result shows South Korean media’s excessive attention to economy-related issues during the pandemic. Table 5.5 shows the top ten topics in the South Korean dataset, and Figure 5.4 shows Celltrion’s Vaccine development-related topic (topic ten).

Table 5.7: Common Topics in top ten position

	US	Economy	Education	Sports
India	✓	✓	✓	✓
Japan	✓	✓	✓	✓
South Korea	✓	✓	✗	✓
UK	✓	✓	✓	✗

5.3.1.5 Comparative Analysis

From our observation of Table 5.3, 5.4, 5.5, and 5.6, we found out that some topics are present in each country's top topics. These topics are - US, Economy, Education, and Sports. For example, the US-related topics present in - UK (topic four and six), India (topic two), Japan (topic six), and South Korea (topic eight). Similarly, economy-related topics present in – the UK (topic seven and nine), India (topic seven), Japan (topic one and two), and South Korea (topic one, three, four, and seven). Table 5.7 shows these common topics. We can infer that economy, education, and sports are the most affected sectors by the COVID-19 pandemic. Simultaneously, the US's presence at the top position in every dataset has two-fold implications. One - the worst affected country by the COVID-19 pandemic and second – the US presidential election's significance for the world. We would use the comparative analysis results in our second step of this research (sentiment classification) to investigate how the news sentiments of these common issues vary in all four countries.

5.3.2 K-means Clustering with Davies Bouldin Index

For the k-means clustering, we used the topic vectors produced from the top2vec model. Before applying the k-means algorithm on the topic vectors, labeling topics (all topics for each country) was completed. These labels were given based on the keywords

(50 keywords for each topic). Next, the dimensionality reduction step is performed. The topics vectors were in 300 dimensions, so first, we used Principal Component Analysis (PCA) for dimensionality reduction [83] and reduced the topic vectors into three dimensions. Then, to evaluate the optimal number of clusters, we used the Davies Bouldin score. The lowest score represents the optimal number of clusters. After finding out the number of clusters, then we visualize the clusters in a scatter plot. For the k-means clustering, we used the scikit-learn Python library.

5.3.2.1 K-means Clustering on Indian Topics

By applying the top2vec model to the Indian dataset, we got 402 topics. These 402 topics are further classified into clusters to understand the distribution of these topics. Davies Bouldin score projected that these 402 topics can be classified into four clusters (lowest point). Figure 5.6 shows Davies Bouldin score and number of clusters.

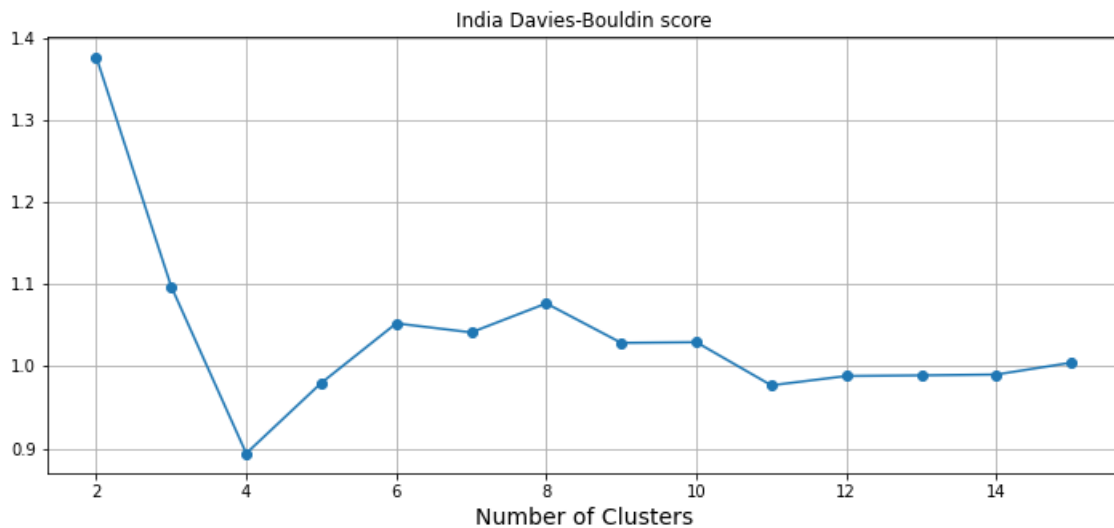


Figure 5.6: **Davies Bouldin Score and Number of Clusters (Indian Topics)**

These four clusters can be labeled as 1) Various Topics, 2) Foreign Topics, 3) Different States Topics, and 4) Showbiz Topics. Figure 5.7 shows cluster no. 4 (showbiz

news) as an example.

```

CLUSTER #4
Topics: ['Janta Curfew [IND] [125]', 'Showbiz News [IND] [230]', 'Showbiz News [IND] [373]', 'Celebrities and COVID
-19 [IND] [271]', 'Death of Famous People [IND] [168]', 'Charity and Donation [IND] [91]', 'Disinfection and Hand w
ashing [IND] [123]', 'Showbiz News [IND] [49]', 'Sports [IND] [280]', 'JK Rowling Controversy [IND] [362]', 'SNS [I
ND] [55]', 'Misinformation SNS [IND] [170]', 'Sports [IND] [284]', 'Ramzan and Hajj [IND] [114]', 'Athletes and Oly
mpics [IND] [355]', 'Art [IND] [112]', 'Janta Curfew [IND] [216]', 'Yoga and Ayurveda [IND] [269]', 'Soumitra Chat
terjee News [IND] [292]', 'Showbiz News [IND] [219]', 'Showbiz News [IND] [344]', 'Mann ki Baat [IND] [173]', 'Life
style and Fitness [IND] [174]', 'SNS [IND] [339]', 'Miscellaneous [IND] [73]', 'Hollywood Celebrities News [IND] [1
85]', 'Showbiz News [IND] [175]', 'Showbiz News [IND] [177]', 'Art and Music [IND] [78]', 'Misinformation [IND] [30
5]', 'Smartphone Apps [IND] [94]', 'Showbiz News [IND] [312]', 'Kapoor Family [IND] [375]', 'Pets and Animals [IND]
[180]', 'Showbiz News [IND] [8]', 'Sushant Singh Rajput Case [IND] [381]', 'Cricket and Bollywood Celebrities [IND]
[15]', 'Showbiz News [IND] [392]', 'Showbiz News [IND] [260]', 'Showbiz News [IND] [20]', 'Showbiz News [IND] [146]
', 'TV Programme [IND] [256]', 'Sports [IND] [144]', 'Showbiz News [IND] [152]', 'Celebrities and Wedding [IND] [15
3]', 'Frontline Warrior [IND] [154]', 'Pranab Mukherjee [IND] [391]', 'Showbiz News [IND] [400]', 'Stress and Menta
l Wellbeing [IND] [29]', 'Showbiz News [IND] [253]']

```

Figure 5.7: Indian Topics Classified in Cluster No. 4

Figure 5.8 is the representation of all topics in the scatter plot. The plot shows that Indian topics are sparsely distributed, and many topics in cluster three (green) and four (red) are outliers.

5.3.2.2 K-means Clustering on Japan's Topics

Japan's COVID-19 data consists of 255 topics, and these topics are further classified into clusters. The Davies Bouldin score shows that these 255 topics can be classified into only two topics. Figure 5.9 presents the Davies Bouldin score and number of clusters.

Few topics (8) related to COVID-19 cases in various prefectures are categorized in cluster 2, and other than that, all the remaining topics are classified into one cluster. Figure 5.10 shows topics in cluster two, and Figure 5.11 present the scatter plot of Japan's topics.

It is clear from the figure 5.10 that Japan's topics are very closely distributed. Eight topics (orange) and two blue topics on the upper left-hand corner can be seen as outliers. The results from k-means clustering of Japan data present that Japan's COVID-19 news is very focused.

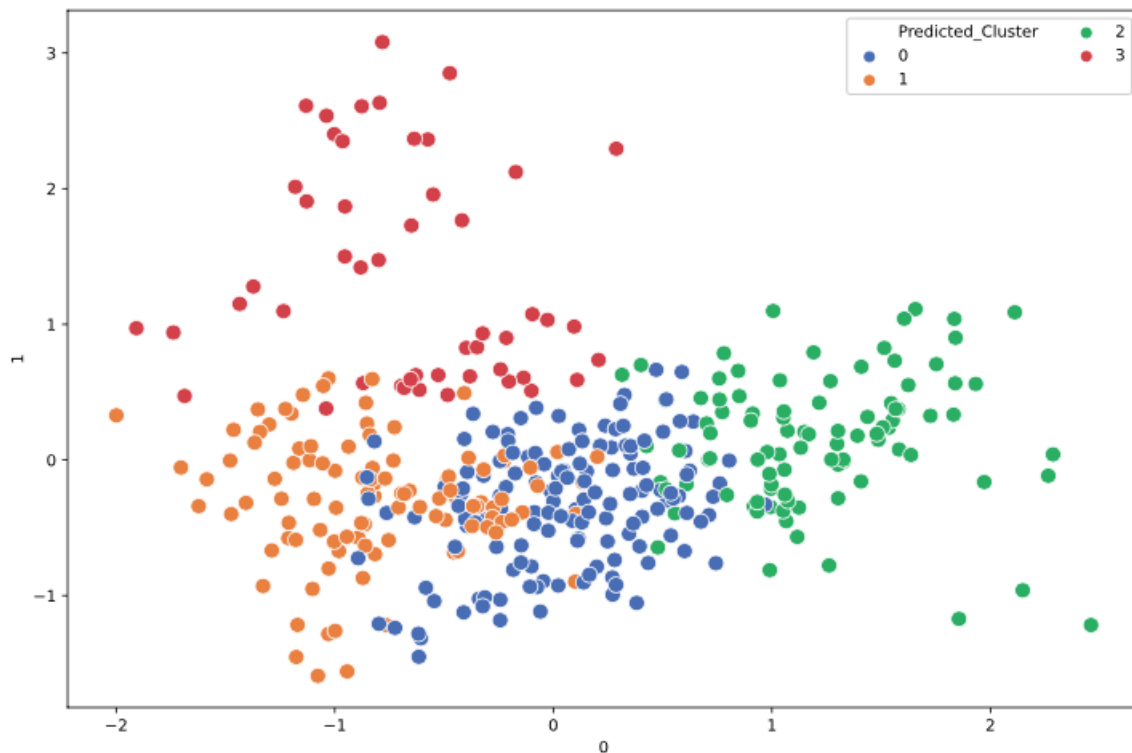


Figure 5.8: Scatter Plot of Indian Topics

5.3.2.3 K-means Clustering on South Korea's Topics

South Korea dataset was the smallest (10,076 articles), and the topics produced by the top2vec model were also the lowest (127 topics). However, the number of clusters predicted by Davies Bouldin score was the highest – 11 clusters. Figure 5.12 presents Davies Bouldin score and the number of clusters, and this figure shows that 11 clusters achieve the lowest score. This result makes the South Korean COVID-19 dataset most diversified and sparsely distributed among the four countries studied in this case study. (the UK produced four clusters, and the description is provided in section 5.3.2.4.)

Based on the clusters suggested by the above figure, the k-means clustering algorithm is applied to South Korean topics to classify them into different clusters. Cluster 7 consists of topics about 'COVID-19 infection and reinfection,' cluster 9 consists of

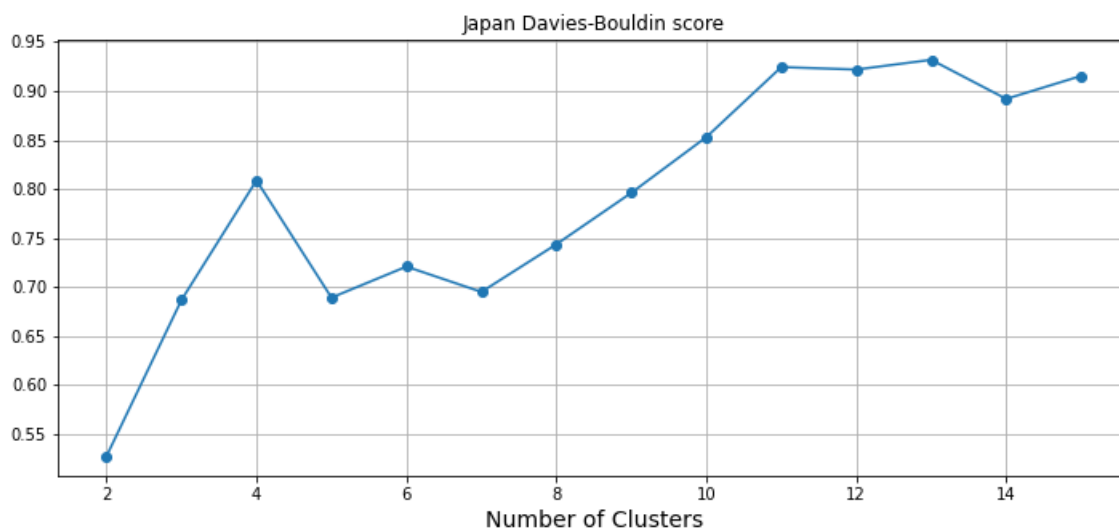


Figure 5.9: Davies Bouldin Score and Number of Clusters (Japan’s Topics)

CLUSTER #2
 Topics: ['Daily Cases Figures [JP] [255]', 'Cases in Prefectures [JP] [169]', 'COVID-19 Cases in Japan [JP] [206]', 'Cases in Prefectures [JP] [237]', 'Nightlife Restrictions [JP] [14]', 'Various Complications due to COVID-19 [JP] [109]', 'COVID-19 Cases in Japan [JP] [88]', 'COVID-19 Cases in Japan [JP] [84]']

Figure 5.10: Japan Topics Classified in Cluster No. 2

topics about 'Foreign countries,' and cluster 11 have topics about 'Art and Culture.' These are some examples of different clusters in South Korean topics. Figure 5.13 shows clusters no. 7 to 11 along with topics, and Figure 5.14 clearly illustrates the distribution of all the topics in the South Korean COVID-19 dataset in the scatter plot.

5.3.2.4 K-means Clustering on UK’s Topics

Top2Vec model produced 308 topics in UK COVID-19 dataset that has 23,821 articles. The Davies Bouldin score predicted that these 308 topics could be classified into four clusters. Figure 5.15 shows Davies Bouldin score and number of clusters for UK topics.

K-means clustering classified UK topics into four clusters: 1) Cluster with US-related topics, 2) Cluster with Australia-related topics, 3) Cluster with UK-related topics, and

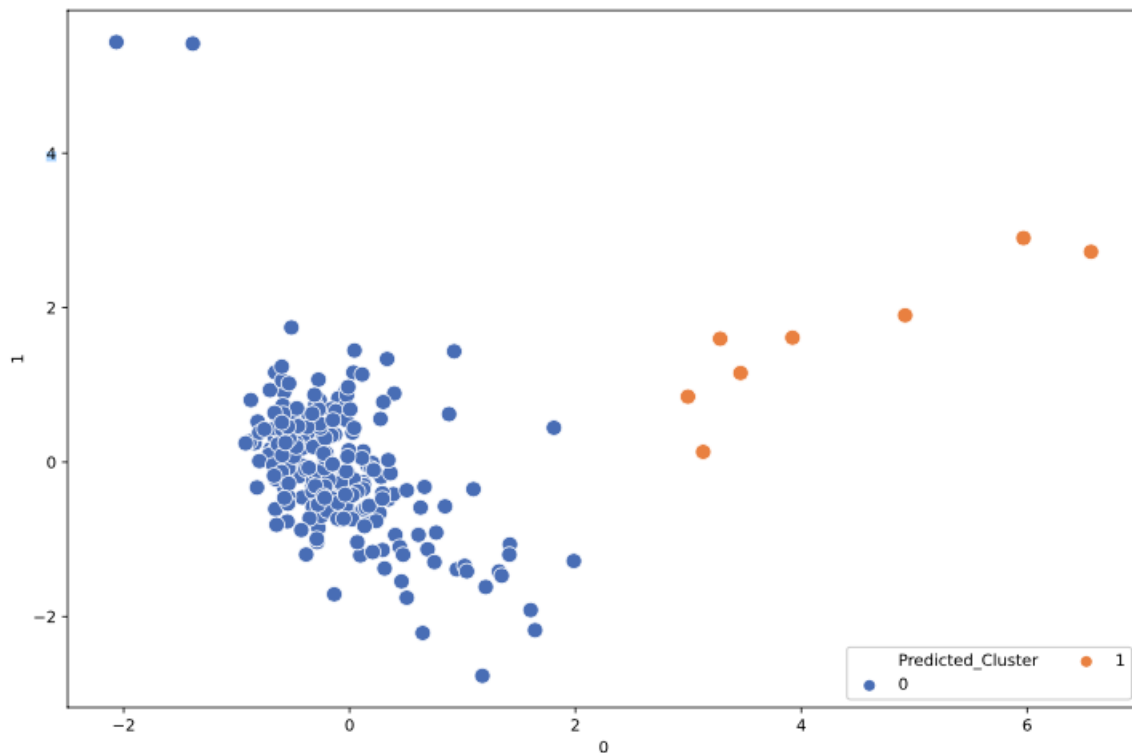


Figure 5.11: Scatter Plot of Japan's Topics

4) Cluster that includes all the Miscellaneous topics. Figure 5.16 illustrates the example of cluster 2 (Australia-related topics). Further, the scatter plot visualizes the topics. We can see that even though UK topics are classified into only four clusters, the topics are not densely distributed compared to Japan or India.

Sections 5.3.1, and 5.3.2 presented the results of the top2vec model and then further classified the results (topics) into clusters to understand the nature of the COVID-19 dataset. The next section would apply NMF topic modeling on the COVID-19 dataset and then compare the results of NMF with Top2Vec topics.

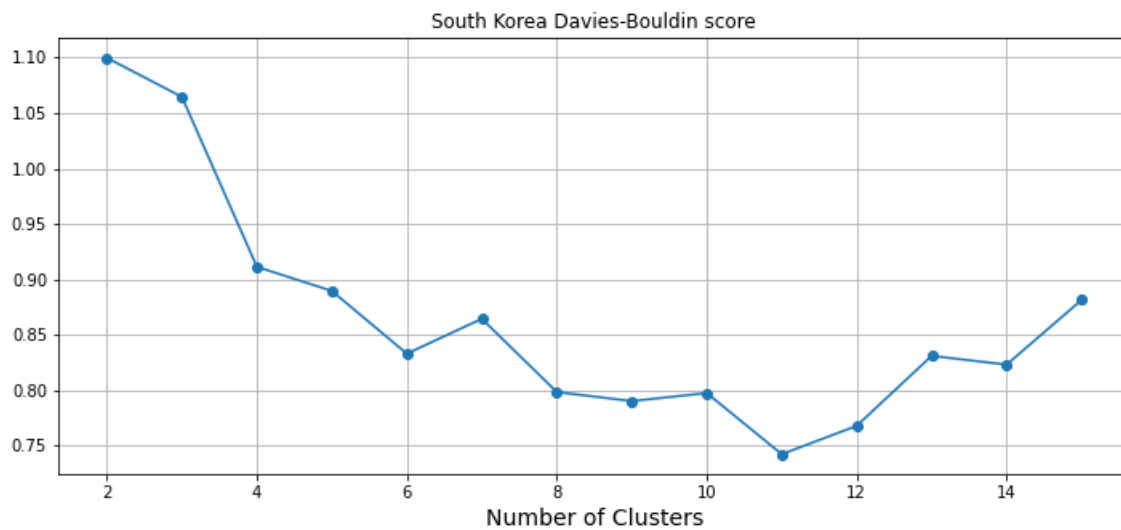


Figure 5.12: Davies Bouldin Score and Number of Clusters (South Korea's Topics)

CLUSTER #7
 Topics: ['COVID-19 Reinfection [SK] [38]', 'US and Travel Advisory [SK] [88]', 'Diamond Princess [SK] [111]', 'Diamond Princess [SK] [79]', 'USFK [SK] [22]', 'Quarantine and Fine [SK] [69]', 'Coronavirus Origin [SK] [96]', 'Traveling Restrictions [SK] [51]', 'USFK [SK] [41]']

CLUSTER #8
 Topics: ['Miscellaneous [SK] [63]', 'Shincheonji [SK] [53]', 'Traveling During Holidays [SK] [86]', 'Contact Tracing and Privacy [SK] [65]', 'Flu Vaccination [SK] [93]', 'Election and Voting [SK] [94]', 'Shincheonji [SK] [58]', 'Stress and Mental Wellbeing [SK] [76]', 'Religious Activity [SK] [78]', 'Infections in Pets [SK] [80]', 'Education [SK] [60]', 'Fine for Violation [SK] [90]', 'Domestic Politics [SK] [99]', 'Elementary School [SK] [11]', 'Rallies and Protest [SK] [98]']

CLUSTER #9
 Topics: ['North Korea [SK] [37]', 'Typhoon and Flood [SK] [81]', 'ASEAN Teleconference [SK] [61]', 'North Korea [SK] [35]', 'WTO and Reforms [SK] [124]', 'Japan Politics [SK] [91]', 'USFK [SK] [36]', 'Kim Jong Um and North Korea [SK] [122]', 'North Korea [SK] [109]', 'QUAD [SK] [72]', 'Summit Teleconferencing [SK] [12]', 'RCEP [SK] [115]', 'North Korea [SK] [26]', 'Middle East and Teleconference [SK] [100]', 'China [SK] [71]']

CLUSTER #10
 Topics: ['Art and Culture [SK] [103]', 'Mobile Games and Apps [SK] [123]', 'Film and Film festival [SK] [56]', 'Music and Theatre [SK] [50]', 'Music and Theatre [SK] [87]', 'Cherry Blossom in SK [SK] [126]', 'Music Concert [SK] [85]', 'Art, Opera and Theatre [SK] [119]']

CLUSTER #11
 Topics: ['Electric and Hydrogen Cars [SK] [46]', 'Easter Jet Airline Bankruptcy [SK] [110]', 'Banking and Stocks [SK] [49]', 'Asiana and other Airlines [SK] [48]', 'Banking and Stocks [SK] [39]', 'Telecom, AI and ICT [SK] [33]', 'Ssangyong and Mahindra [SK] [121]', 'Carmakers [SK] [30]', 'Banking [SK] [29]', 'Banking and Insurance [SK] [102]', 'KOSPI [SK] [73]', 'Carmakers [SK] [120]', 'Economy and Manufacturing [SK] [107]', 'Jobs and Recruitment [SK] [55]', 'Mobile Games and Apps [SK] [83]', 'Supermarket [SK] [3]', 'Housing and Property Tax [SK] [92]', 'Smartphone Manufacturing [SK] [7]']

Figure 5.13: South Korea Topics Classified in Cluster No. 7 to 11

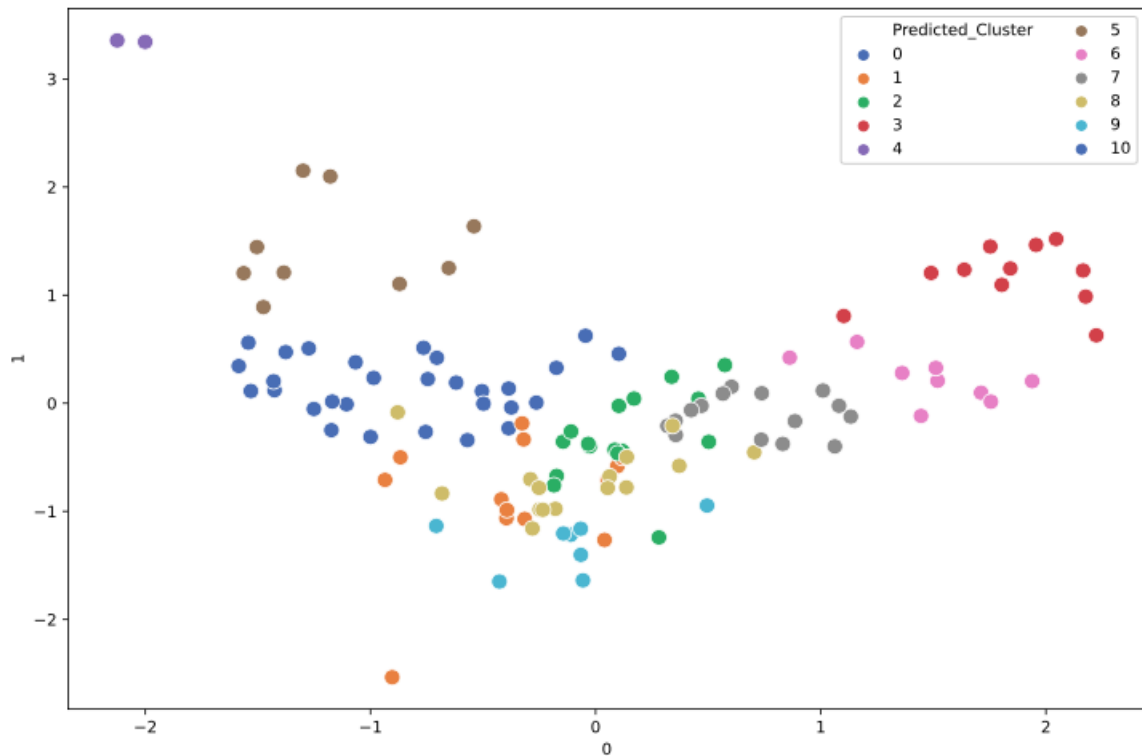


Figure 5.14: Scatter Plot of South Korea's Topics

5.3.3 NMF Topic Modeling of COVID-19 Dataset and Comparison with Top2Vec

The COVID-19 dataset has four countries, and for each country, there would be two steps as mentioned in the research methodology. In the first part (A), TC-W2V is used to find the optimal number of topics and producing topics (NMF algorithm), and the second part (B) is the comparative analysis of NMF topics and Top2Vec topics. Before going into the details, it would be beneficial to compare the two approaches' major differences. Table 5.8 compares some of the critical parameters of both topic modeling approaches. The major differences between the two models are the algorithms and representation. While the NMF algorithm is based on linear algebra and multivariate

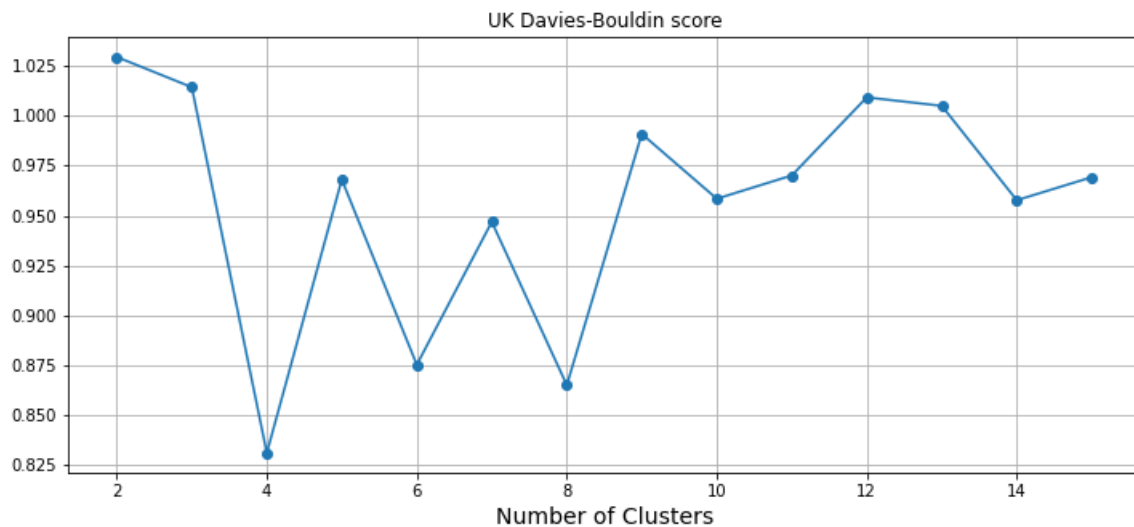


Figure 5.15: Davies Bouldin Score and Number of Clusters (UK's Topics)

CLUSTER #2
 Topics: ['Australia News [UK] [218]', 'Australia News [UK] [34]', 'Cruise Ship [UK] [232]', 'Australia News [UK] [26]', 'Traveling [UK] [129]', 'Australia News [UK] [3]', 'Restrictions in Australia [UK] [237]', 'Australia News [UK] [240]', 'Australia News [UK] [214]', 'Australia News [UK] [136]', 'Various Countries News [UK] [102]', 'Australia News [UK] [255]', 'Australia News [UK] [300]', 'New Zealand and Australia [UK] [8]', 'Celebrities News [UK] [261]', 'Economy and Jobs [UK] [9]', 'Prison Australia [UK] [294]', 'Australia News [UK] [264]', 'Cruise Ship [UK] [40]', 'Airlines [UK] [21]', 'Australia News [UK] [13]', 'Australia News [UK] [275]', 'Australia News [UK] [101]', 'Australia News [UK] [159]', 'Australia News [UK] [234]', 'Australia News [UK] [85]', 'Protest by Aboriginal [UK] [140]', 'Outdoor Activities [UK] [186]', 'Australia News [UK] [209]', 'Real Estate [UK] [50]', 'Australia News [UK] [153]', 'Australia News [UK] [59]', 'Fitness [UK] [173]', 'Australia News [UK] [172]', 'Australia News [UK] [48]', 'Australia News [UK] [167]', 'Australia News [UK] [201]', 'Australia News [UK] [190]', 'Australia News [UK] [208]', 'Australia News [UK] [58]', 'Australia Education [UK] [158]', 'COVID-19 App [UK] [163]', 'Melbourne News [UK] [165]']

Figure 5.16: UK Topics Classified in Cluster No. 2

analysis, the top2vec algorithm uses neural networks. Similarly, NMF is based on the traditional bag-of-words representation, while top2vec uses semantic embedding. Because of this reason, the top2vec results take into consideration the context of the text. Hence, produce comparatively better results than other bag-of-words-based topic modeling approaches. NMF (and even LDA) has the drawback of not getting the best results without knowing the optimal number of topics. For NMF, this drawback can be overcome by applying the TC-W2V algorithm in the NMF topic model. Top2Vec does not have this drawback as it produces topics automatically.

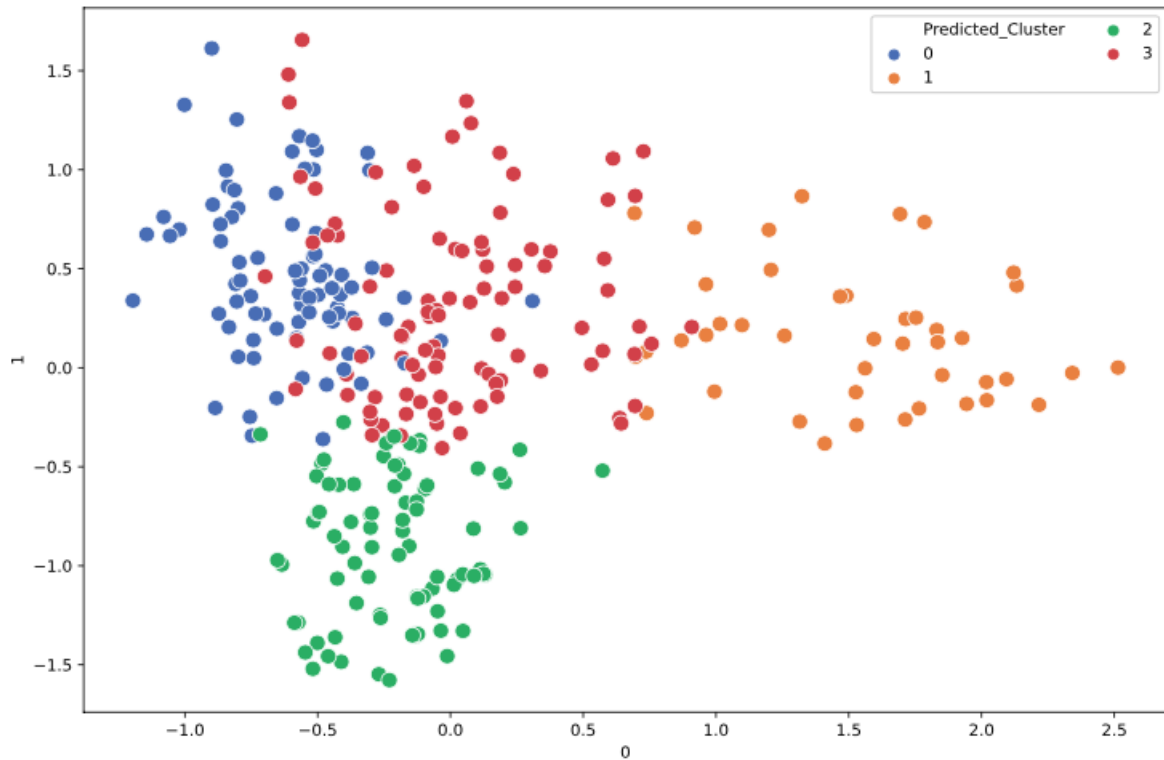


Figure 5.17: Scatter Plot of UK's Topics

Table 5.8: Comparing NMF and Top2Vec Approaches

No.	Features	NMF	Top2Vec
1.	Approach	Matrix Factorization	Distributed Representation
2.	Algorithm	Linear Algebra and Multivariate Analysis	Neural Networks
3.	Proposed/Popularized By	Lee & Seung, 1999	Angelov, 2020
4.	Preprocessing	Required	Not Required
5.	Representation	Bag-of-Words	Semantic Embedding
6.	Type	Unsupervised (Semi-supervised with TC-W2C)	Unsupervised
7.	For Better Results	Previous Knowledge of Number of Topics Necessary	No previous knowledge required

5.3.3.1 NMF Topic Modeling and Comparative Analysis with Top2Vec for Indian Dataset

- (A) **TC-W2V and NMF Topics** – Indian dataset of COVID-19 articles consists of 47,342 articles. The NMF model is trained from $k=5$ to $k=69$ and found that 68 topics are the optimal number for Indian data. Figure 5.18 presents the mean coherence and number of topics for the Indian dataset. The top ten of the resultant topics is shown in table 5.9. Some of these topics are: 1) work from home (4.6%), 2) COVID-19 testing (4.2%), 3) COVID-19 news from various states (3.9%), 4) Government orders and notifications (3.5%), and 5) Essential goods (3.3%).

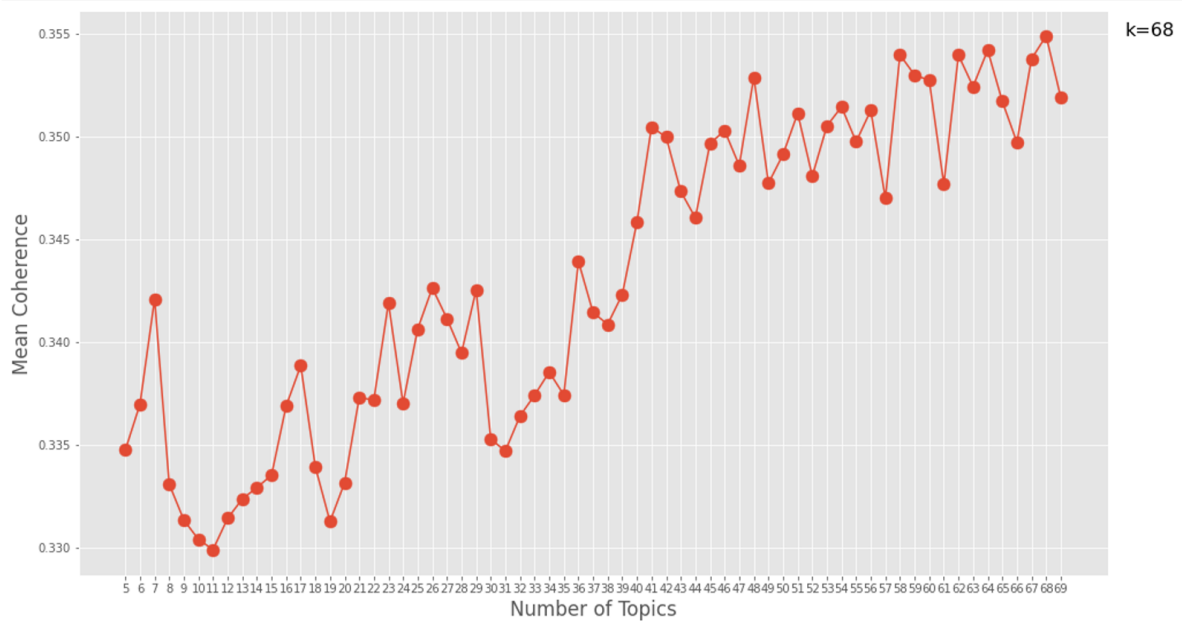


Figure 5.18: Mean Coherence and Number of Topics (India)

- (B) **Comparing NMF and Top2Vec Topics (Top ten)** – While Top2Vec produced 402 topics, NMF suggested only 68 topics. Out of the top ten topics, except two topics, all are different. COVID-19 testing and Showbiz News/Celebrities on SNS (topic 6 in NMF and topic 8 in Top2Vec) are common in both. Table 5.10

Table 5.9: Top Ten Topics from NMF from Indian Dataset

No.	Topic Weightage (%)	Top 10 Words	Topic Label
1.	4.6%	Time, Get, Go, Work, People, Think, Home, Feel, Things, Help	Work from Home
2.	4.2%	Test, Positive, Negative, Antigen, Conduct, Result, Rapid, Symptoms, Virus, Asymptomatic	COVID-19 Testing
3.	3.9%	State, Kerala, Chief, Department, Pradesh, Minister, Maharashtra, Health, Karnataka, Rajasthan	Various States News
4.	3.5%	Government, Order, Private, Issue, Decision, Hospitals, Central, Official, Take, Secretary	Various Government Orders and Notifications
5.	3.3%	Food, Ration, Supply, People, Residents, Essential, Provide, Help, Distribute, Ensure	Essential Goods
6.	3%	Film, Actor, Share, Post, Read, Video, Write, Pdt, Shoot, Khan	Celebrities on SNS
7.	2.5%	Employees, Work, Office, Staff, Offices, Pay, Company, Employee, Noida, Salaries	Jobs and Employment
8.	2.5%	Coronavirus, Virus, People, Health, Countries, World, Country, Infections, Italy, Outbreak	Global Coronavirus News
9.	2.4%	Lockdown, Extend, April, March, Till, Nationwide, Restrictions, Announce, Impose, Lift	Lockdown
10.	2.4%	India, Countries, Ministry, Country, Indian, World, Modi, Indians, Global, Foreign	Bringing Indian's from Abroad

presents the top ten topics in both approaches for the Indian dataset.

5.3.3.2 NMF Topic Modeling and Comparative Analysis with Top2Vec for Japan's Dataset

- (A) **TC-W2V and NMF Topics** – Japanese dataset of COVID-19 articles is consists of 21,039 articles. The NMF model is trained from $k=5$ to $k=64$ and found out that 41 topics are the optimal number for Japanese data. Figure 5.19 present the mean coherence and number of topics for the Japanese dataset. The top ten of the resultant topics is shown in table 5.11. Some of these topics are: 1) work from home (8.2%), 2) COVID-19 testing (5.1%), 3) COVID-19 Outbreak in Wuhan (4.4%), 4) COVID-19 Cases (4.3%), and 5) US News (3.9%).

Table 5.10: Comparing Top Ten Topics from NMF and Top2Vec (India)

No.	NMF Topics	Top2Vec Topics
1.	Work from Home	Cricket
2.	COVID-19 Testing	US Election
3.	Various States News	Punjab News
4.	Various Government Orders and Notifications	Crime
5.	Essential Goods	Vaccine
6.	Celebrities on SNS	School Education
7.	Jobs and Employment	Economy
8.	Global Coronavirus News	Showbiz News
9.	Lockdown	University Education
10.	Bringing Indian's from Abroad	COVID-19 testing

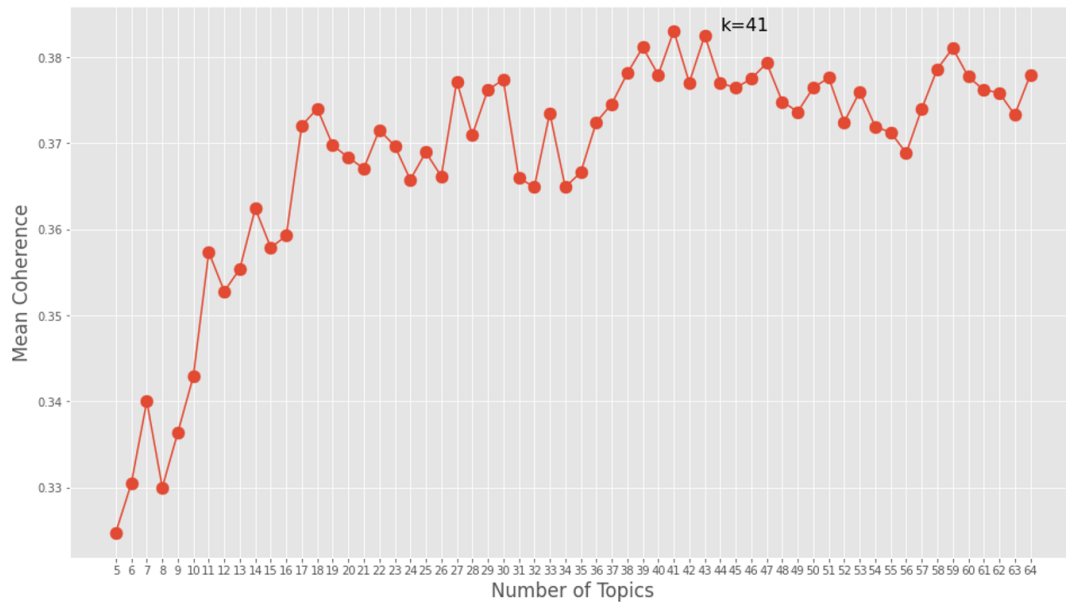


Figure 5.19: Mean Coherence and Number of Topics (Japan)

Table 5.11: Top Ten Topics from NMF from Japan's Dataset

No.	Topic Weightage (%)	Top 10 Words	Topic Label
1.	8.2%	People, Go, Get, Home, Time, Think, Live, Work, Feel, Children	Work from Home
2.	5.1%	Test, Positive, PCR, Virus, Result, Health, Kit, Negative, People, Infect	COVID-19 Testing
3.	4.4%	Wuhan, Virus, China, Hubei, Chinese, Outbreak, Province, City, Health, People	Outbreak in Wuhan
4.	4.3%	Case, Report, Confirm, Number, Total, Infections, Deaths, Daily, Day, Tally	COVID-19 Cases in Japan
5.	3.9%	Trump, President, White, House, Administration, Donald, Washington, Americans, State, Pence	US News
6.	3.9%	Reopen, Lockdown, Restrictions, Deaths, State, People, City, Health, Distance, Restaurants	Lockdown and Reopening
7.	3.2%	News, Japan, Subscribe, Crucial, Information, Times, Practical, Cope, Please, Stag	News and Information
8.	2.9%	Mask, Wear, Face, Cloth, Supply, Sell, Distribute, Surgical, Medical, Cover	Mask
9.	2.8%	Emergency, State, Prefectures, Government, Lift, Declaration, Request, Declare, Tokyo, Panel	State of Emergency
10.	2.8%	Tokyo, Infections, Capital, Metropolitan, Average, Record, Day, Mainichi, Total, Koike	Tokyo COVID-19 Cases

- (B) **Comparing NMF and Top2Vec Topics (Top ten)** – The number of topics produced for Japan's COVID-19 data by Top2Vec and NMF is 255 and 41, respectively. As visible in table 5.12, the topics (at least top ten) produced by both approaches are very different. Topic about the State of Emergency in Japan (topic 9 in NMF and topic 7 in Top2Vec) and COVID-19 Cases in Japan (topic 4 in both) are only two common topics.

Table 5.12: Comparing Top Ten Topics from NMF and Top2Vec (Japan)

No.	NMF Topics	Top2Vec Topics
1.	Work from Home	Global Stock Market
2.	COVID-19 Testing	Nikkei
3.	Outbreak in Wuhan	Tokyo 2020 Games
4.	COVID-19 Cases in Japan	COVID-19 Cases in Japan
5.	US News	COVID-19 Cases in South Korea
6.	Lockdown and Reopening	US Election
7.	News and Information	State of Emergency in Japan
8.	Mask	Multilateralism
9.	State of Emergency in Japan	North Korea News
10.	Tokyo COVID-19 Cases	School Education

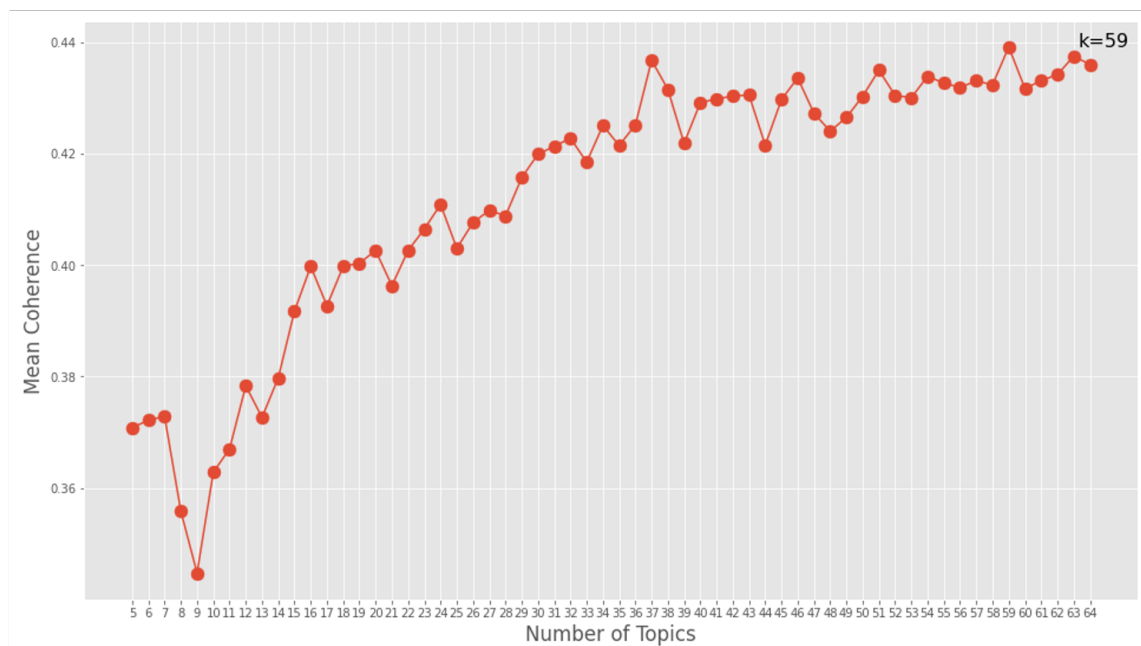


Figure 5.20: Mean Coherence and Number of Topics (South Korea)

5.3.3.3 NMF Topic Modeling and Comparative Analysis with Top2Vec for South Korean Dataset

- (A) **TC-W2V and NMF Topics** – There were 10,076 COVID-19 related articles collected from South Korean newspapers. The NMF model is trained from $k=5$ to $k=64$ and found that 59 topics are the optimal number for the South Korea dataset. Figure 5.20 present the mean coherence and number of topics for South Korea. The top ten of the resultant topics is shown in table 5.13. Some of these topics are: 1) Miscellaneous (5.5%), 2) Business (4.8%), 3) COVID-19 Cases Global (3.7%), 4) Rally and Protest in SK (3.4%), and 5) North Korea News (3.3%).

Table 5.13: Top Ten Topics from NMF from South Korea’s Dataset

No.	Topic Weightage (%)	Top 10 Words	Topic Label
1.	5.5%	People, World, Live, Time, Get, Think, Go, Need, Change, Work	Miscellaneous
2.	4.8%	Company, Firm, SK, Business, Telecom, KT, Employees, CEO, Service, Industry	Business
3.	3.7%	Virus, Health, Deaths, Coronavirus, Case, Italy, Spread, People, Outbreak, World	COVID-19 Cases Global
4.	3.4%	Rally, Seoul, City, Police, Government, Gwanghwamun, Court, Park, Jun, Metropolitan	Rally and Protest in SK
5.	3.3%	North, Korea, Pyongyang, Korean, Border, Sanction, Nuclear, South, Leader, Unification	North Korea News
6.	2.9%	Minister, Foreign, Kang, Cooperation, Ministry, South, Talk, Korea, Seoul, Meet	Government Meeting
7.	2.8%	Market, Stock, Investors, Price, KOSPI, Securities, Investment, Sell, Analyst, Index	Stock Market
8.	2.5%	Trade, WTO, Yoo, Eu, Countries, Global, World, International, European, Multilateral	Global Trade
9.	2.5%	Win, Trillion, Bond, Worth, Debt, Corporate, Net, Amount, Data, Increase	Corporate News
10.	2.5%	Case, Infections, Report, Total, Country, Number, Import, Seoul, Korea, Virus	COVID-19 Cases in SK

- (B) **Comparing NMF and Top2Vec Topics (Top ten)** – Top2Vec produced 127, and the NMF model predicted 59 topics from South Korea data. As it is visible from table 5.14, there are no common topics among the topics (top ten) produced by both approaches.

Table 5.14: Comparing Top Ten Topics from NMF and Top2Vec (South Korea)

No.	NMF Topics	Top2Vec Topics
1.	Miscellaneous	Relief Package
2.	Business	Multilateralism
3.	COVID-19 Cases Global	Supermarket
4.	Rally and Protest in SK	Economic Outlook
5.	North Korea	Cluster Cases
6.	Government Meeting	Sports
7.	Stock Market	Smartphone Manufacturing
8.	Global Trade	US COVID-19 Cases
9.	Corporate News	Shincheonji
10.	COVID-19 Cases in SK	Vaccine Trials

5.3.3.4 NMF Topic Modeling and Comparative Analysis with Top2Vec for the UK Dataset

- (A) **TC-W2V and NMF Topics** – There were 23,821 COVID-19 related articles collected from the UK newspapers. The NMF model is trained from $k=5$ to $k=64$ and found out that 31 topics are the optimal number for the UK dataset. Figure 5.21 present the mean coherence and number of topics for the UK. The top ten of the resultant topics is shown in table 5.15. Some of these topics are: 1) Family

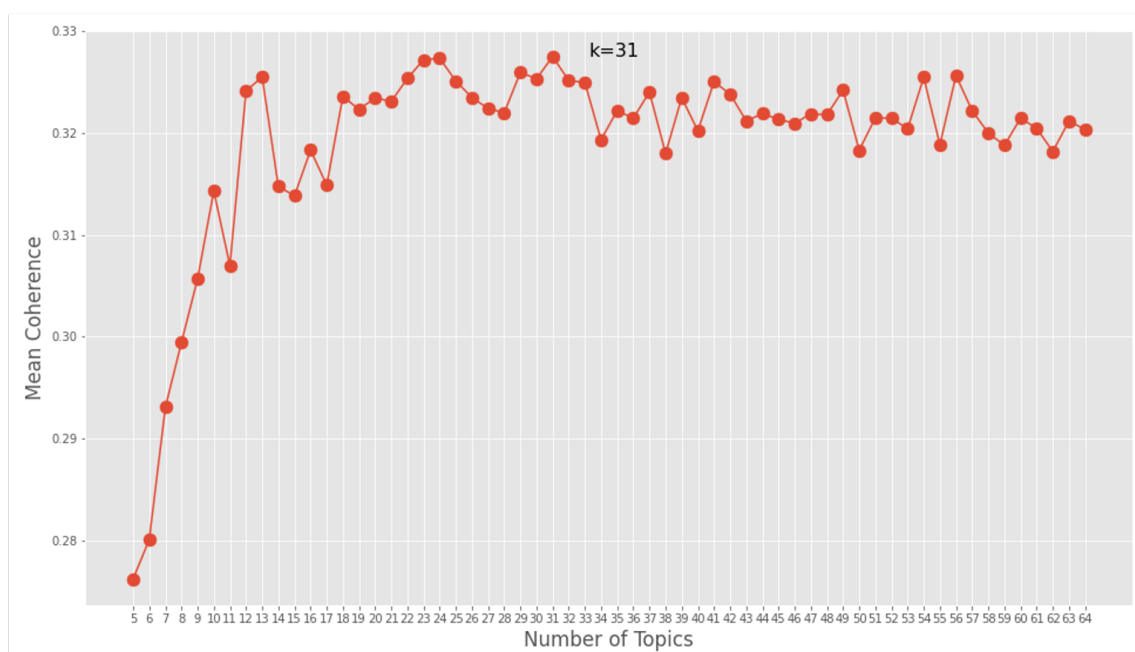


Figure 5.21: Mean Coherence and Number of Topics (UK)

Stories (7.5%), 2) Miscellaneous (6.9%), 3) UK Government (6.5%), 4) COVID-19 (5.3%), and 5) Jobs and Unemployment (4.7%).

- (B) **Comparing NMF and Top2Vec Topics (Top ten)** – Top2Vec produced 308, and the NMF model predicted 31 topics from UK data. As it is visible from table 5.16, there are no common topics among the topics (top ten) produced by both approaches.

From the comparison presented in various subsections of section 5.3.3, it is clear that the results (topics) produced by both approaches are completely different (at least for the top ten topics). Further, for top2vec, the number of topics is directly proportionate to the number of articles. The largest dataset has the highest number of topics. However, this was not true for NMF and TC-W2V approaches. In this approach, India (largest dataset) produced the highest number of topics (68). However, this was not true for others as South Korea, with the smallest number of articles (10,076), produced

Table 5.15: Top Ten Topics from NMF from the UK's Dataset

No.	Topic Weightage (%)	Top 10 Words	Topic Label
1.	7.5%	Family, Mother, Go, Die, Father, Love, Daughter, Children, Get, Wife	Family Stories
2.	6.9%	Watch, Video, Updatestyleonhover, Onpos, Activeclass, Wocc, Pagesize, Pagecount, Moment, Videos	Miscellaneous
3.	6.5%	Johnson, Minister, UK, Government, Boris, Prime, Lockdown, Today, PM, Secretary	UK government
4.	5.3%	Test, Positive, Antibody, Swab, Result, Symptoms, Kit, Health, Contact, Trace	COVID-19 Testing
5.	4.7%	Job, Tax, Pay, Businesses, Unemployment, Scheme, Workers, Furlough, Economy, Company	Jobs and Unemployment
6.	4.4%	Care, Home, Residents, Deaths, Staff, Nurse, Die, PPE, Elderly, Dementia	Elderly Care
7.	4.4%	Reopen, Restaurants, Bar, Allow, Pubs, Restrictions, Open, Limit, Distance, Businesses	Lockdown and Reopening
8.	4.2%	Police, Officer, Protest, Arrest, Protesters, Man, Fin, Black, Floyd, Charge	BLM Protest in the US
9.	3.9%	Trump, President, White, House, Donald, Fauci, Americans, Rally, Administration, Pence	US Coronavirus
10.	3.8%	Virus, Viruses, Infect, Sars, Case, Spread, Coronaviruses, Wuhan, Outbreak, Animals	About Coronavirus

the second-highest number of topics (59).

5.3.4 Pros and Cons of NMF and Top2Vec

NMF is a well-established and popular topic modeling approach. However, just like other popular and old approaches such as LDA, NMF also have same problems. Top2Vec tried to overcome some of these issues. Below are some pros and cons of both NMF and Top2Vec approaches.

- (A) **Pros and Cons of NMF** – The main pros of NMF in comparison to Top2Vec is that it produces topic model really fast (less than one hour). The problem of finding optimal number of topics can be solved by using method such as TC-

Table 5.16: Comparing Top Ten Topics from NMF and Top2Vec (UK)

No.	NMF Topics	Top2Vec Topics
1.	Miscellaneous	Relief Package
2.	Business	Multilateralism
3.	COVID-19 Cases Global	Supermarket
4.	Rally and Protest in SK	Economic Outlook
5.	North Korea	Cluster Cases
6.	Government Meeting	Sports
7.	Stock Market	Smartphone Manufacturing
8.	Global Trade	US COVID-19 Cases
9.	Corporate News	Shincheonji
10.	COVID-19 Cases in SK	Vaccine Trials

W2V. The overall topic model can be visualize by utilizing MDS. However, as NMF is based on bag-of-words approach, the context of text is not considered in the topics detection. This is the biggest drawback of NMF. Though NMF requires preprocessing, however it is not a significant drawback.

- (B) **Pros and Cons of Top2Vec** – The biggest advantage of Top2Vec is that since it is based on neural networks, it takes into consideration the context. Other pros of using Top2Vec is that it is automatic. It means there is no need to find the optimal number of topics. The only parameter a research need to select is the training (‘fast-learn,’ ‘learn,’ and ‘deep-learn’). Out of these three, ‘deep-learn’ produces best results but takes long time (around six-seven hours). This is the biggest drawback of Top2Vec.

Topic Information Gain is one method to calculate the quality of topics. However, till now, this feature is not included in Top2Vec or the implementation has not been shared

by the researcher who proposed Top2Vec approach. Hence, at present, the comparison of the quality of topics produced by both NMF and Top2Vec is not possible.

5.4 Limitations

In all topic modeling approaches, topic labeling has to be done manually, and that is one of the biggest limitations of topic modeling. In this research, also for the topic generated by both NMF and Top2Vec, the topic labeling is done manually based on topics keywords. If someone else labels them, they might label them differently than this researcher. The other limitation is analyzing only the top ten topics for each country and each topic modeling approach. Though this research also compares NMF and Top2Vec results (top ten topics) but did not venture into these topics' quality.

5.5 Consideration

Top2Vec is new method and can be seen as a work in progress. At present evaluation feature is not available. This is the reason that in this chapter we manually analyzed only top ten topics. It might be possible that in future, the researcher who proposed this method would include evaluation feature. Our future research would focus of evaluation method for top2vec. This would also help us in comparison of quality of topics generated by top2vec vis-a-vis other topic modeling methods such as NMF and LDA.

Chapter 6

Sentiment Analysis of COVID-19 News

6.1 Motivation

The COVID-19 pandemic is not the first and will not be the last pandemic that humankind suffered/will suffer from. Just in the past 100 years, the world has seen several pandemics such as Spanish Flu (1918-20), Asian Flu (1957-58), AIDS (1981-present day), H1N1 Swine Flu (2009-10), and Zika Virus (2015-present day). In the past outbreaks, information exchange was relatively slow. However, the COVID-19 pandemic occurs at a time when 3.7 billion people worldwide (around 49.7% of the world's population) used web-based information¹. For information, people rely on digital news media or/and social media. People keep themselves up to date about preventions, treatments, and cases of the COVID-19 pandemic by using these digital media. Further, computer technologies provide profound opportunities to fight infectious disease outbreaks. G. Eysenbach discussed SARS and population health technology [84], and K. Goldschmidt researched the usage of technology to support children's wellbeing during the COVID-19 pan-

¹<https://data.worldbank.org/indicator/it.net.user.zs?end=2018start=2010>

demic [85]. R. Singh et al. also tried to predict an epidemic using a sentiment analysis approach in their 2018 research [86]. The applications of sentiment analysis in fighting the COVID-19 pandemic are discussed in detail by A. H. Alamoodi et al. in their paper titled "Sentiment analysis and its applications in fighting COVID-19, and other infectious diseases: A systematic review" [87]. Therefore, the importance of sentiment analysis in studying pandemics, such as the COVID-19, cannot be underestimated.

As it is common with sentiment analysis generally, most studies are performed on social media data as people express their opinion on these platforms. Because of this reason, social media sentiment analysis is useful for understanding the popular public sentiment during health emergencies or other tragic events such as natural disasters. However, misinformation on social media is becoming a huge problem, and it is very much visible during the COVID-19 pandemic also [88] [89]. On the other hand, news media is relatively less prone to misinformation. Sentiment analysis of news headlines can inform us about what kind of news (negative or positive) is coming out from news media during the COVID-19 pandemic. Further, it can also help us understand whether there is any correlation between the impact of COVID-19 in a particular country and the sentiment of news headlines. For example, whether the highly impacted societies such as the US, Brazil, and the UK have a high percentage of negative news. In this chapter, we would try to find out news headlines' sentiment by applying state-of-the-art sentiment classification models. Next, we provide descriptive sentiment analysis.

6.2 Research Methodology

6.2.1 Data Acquisition and Text Preprocessing

As this chapter is the second part of the COVID-19 case study, the data is related to COVID-19. During the initial data collection (which we described in detail in section 5.2.1. of chapter 5), we also collected headlines of the COVID-19 articles. The period

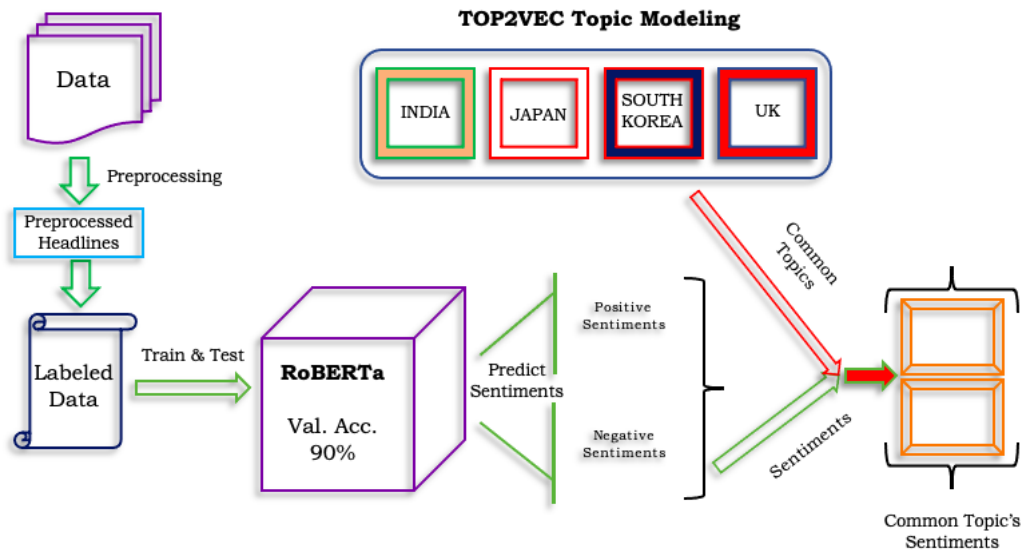


Figure 6.1: **Research Methodology for Sentiment Analysis of COVID-19 News Headlines**

was the same, from January 1, 2020, to December 1, 2020. However, as some headlines were missing and some were in a different language (Korean), the number of headlines is less (154) than the articles. Sentiment analysis is performed on these headlines. Table 6.1 present the collected COVID-19 headlines. The research methodology for this chapter is illustrated in Figure 6.1.

Table 6.1: **Collected COVID-19 News Headlines**

Country	Newspapers	No. of Headlines
India	Hindustan Times, The Indian Express	47,210
Japan	The Japan Times, Asahi Shimbun, Mainichi Shimbun	21,038
South Korea	Korea Herald, Korea Times	10,055
UK	The Daily Mail	23,821
Total		102,124

After collecting the headlines, next is the text preprocessing step. The steps for text preprocessing are similar to the previously described (section 4.2.2 of chapter 4) steps.

An illustration of those steps is provided in Figure 6.2.



Figure 6.2: **Preprocessing Steps Used in Cleaning the Headlines**

After data collection and text preprocessing, the next step is to create a labeled dataset for our sentiment classification model training and testing.

6.2.2 Labeling Headlines for Sentiment Classification

The majority of sentiment analysis research is being conducted on Twitter or other social media posts because these posts tend to be highly subjective, making them a good resource for sentiment analysis. In contrast, news and news articles present facts that make them more objective (except editorials and opinion pieces). In any supervised machine learning approach, labeled data is extremely critical, and it is not easy to manually verify the labels on large news articles. Owing to this reason, we chose headlines. The first step to label the headlines for this research is: using the unsupervised sentiment analysis method. We used the three most popular python-based libraries 1) VADER, 2) Textblob, and 3) SentiWordNet on 102,124 headlines. The second step is: to keep headlines that all three libraries categorize as positive and negative. We discarded all the headlines, which are categorized differently by these three libraries. For example: if a headline is categorized as positive by VADER but negative by Textblob and SentiWordNet, we discarded it from the labeling process. For this process, we wrote a simple python code.

After the second step, we were left with only around 15% of the total headlines. In the third step, we manually confirmed all the headlines from the second step. Lastly, we used oversampling to balance (almost) the labeled data. Oversampling can lead to

overfitting. One way to find whether the model is overfitting is to check the difference between training accuracy and validation accuracy. The training accuracy of the 3rd (last) epoch of our classification model (which gave 90% validation accuracy) has 95.08%. As the difference between training and validation accuracy is not very high, it can be said that our model is not overfitting. Figure 6.3 shows our sentiment classification model summary. In this way, we collected 10,727 headlines (5369 positives and 5358 negatives). This labeled dataset is used to fine-tune the RoBERTa model. Table 6.2 shows few samples of labeled headlines.

```
Epoch 1/3
WARNING:tensorflow:Gradients do not exist for variables ['tf_roberta_for_sequence_classification_1/roberta/pooler/dense/kernel:0', 'tf_roberta_for_sequence_classification_1/roberta/pooler/dense/bias:0'] when minimizing the loss.
WARNING:tensorflow:Gradients do not exist for variables ['tf_roberta_for_sequence_classification_1/roberta/pooler/dense/kernel:0', 'tf_roberta_for_sequence_classification_1/roberta/pooler/dense/bias:0'] when minimizing the loss.
WARNING:tensorflow:Gradients do not exist for variables ['tf_roberta_for_sequence_classification_1/roberta/pooler/dense/kernel:0', 'tf_roberta_for_sequence_classification_1/roberta/pooler/dense/bias:0'] when minimizing the loss.
WARNING:tensorflow:Gradients do not exist for variables ['tf_roberta_for_sequence_classification_1/roberta/pooler/dense/kernel:0', 'tf_roberta_for_sequence_classification_1/roberta/pooler/dense/bias:0'] when minimizing the loss.
668/668 [=====] - ETA: 0s - loss: 0.3157 - accuracy: 0.8545
Epoch 00001: val_accuracy improved from -inf to 0.82000, saving model to /Users/piyushghasiya/PycharmProjects/News/checkpoint-01-0.8200.h5
668/668 [=====] - 37970s 57s/step - loss: 0.3157 - accuracy: 0.8545 - val_loss: 0.2290 - val_accuracy: 0.8200
Epoch 2/3
668/668 [=====] - ETA: 0s - loss: 0.1734 - accuracy: 0.9292
Epoch 00002: val_accuracy improved from 0.82000 to 0.86000, saving model to /Users/piyushghasiya/PycharmProjects/News/checkpoint-02-0.8600.h5
668/668 [=====] - 40163s 60s/step - loss: 0.1734 - accuracy: 0.9292 - val_loss: 0.2991 - val_accuracy: 0.8600
Epoch 3/3
668/668 [=====] - ETA: 0s - loss: 0.1223 - accuracy: 0.9508
Epoch 00003: val_accuracy improved from 0.86000 to 0.90000, saving model to /Users/piyushghasiya/PycharmProjects/News/checkpoint-03-0.9000.h5
668/668 [=====] - 37870s 57s/step - loss: 0.1223 - accuracy: 0.9508 - val_loss: 0.3374 - val_accuracy: 0.9000
```

Figure 6.3: RoBERTa Sentiment Classification Model Summary

Table 6.2: Sample Labeled Headlines with Sentiment

Headlines	Labeled Sentiment
Coronavirus crisis: 100,000 dead in 101 days, half of them in just a week	Negative
Coronavirus leading to heavy job losses among vulnerable workers in Japan	Negative
Covid-19 vaccine tracker, August 27: Moderna shot promises to be equally effective in older patients	Positive
New Zealand is now free of COVID-19	Positive

6.2.3 Sentiment Classification with RoBERTa

The traditional bag-of-words-based NLP classifiers such as LogisticRegressionCV, LinearSVC, and Naïve Bayes discard word order (hence context) and, in turn, the meaning of words in the document (semantics). However, context and meaning are very crucial to the problems of NLP. On the other hand, Bidirectional Encoder Representations from Transformers (BERT) [90], Robustly Optimized BERT Pretraining Approach (RoBERTa) [91] XLNet, GPT-2, and GPT-3 are some of the examples of the pre-trained model based on Transformers – a deep learning model introduced in 2017 [92]. RoBERTa is an extension of the original BERT Model. All these transformers-based models take into account the context for each occurrence of a given word. This is one reason behind the better performance of these models compared to the classical bag-of-words-based models. The original BERT model produced state-of-the-art results in many NLP tasks. However, there was still room for improvements in training objectives, duration of the training, and the data on which it is trained. Hence the Facebook AI Research (FAIR) proposed an "optimized" and "robust" version of BERT. Compared to BERT, RoBERTa is trained on 1) much larger datasets, 2) much longer, 3) longer sequences, 4) with dynamic mask generation, 5) without Next Sentence Prediction (NSP) objective, and 6) larger batches.

6.3 Experiment, Results, and Discussion

6.3.1 Sentiment Classification

To perform sentiment classification, we used our labeled dataset of 10,727 headlines. We fine-tuned the RoBERTa BASE model with 12 layers and 768 hidden dimensions. We setup hyperparameters of 3 epochs, the learning rate of 1e-5, and a batch size of 16. The last third epoch achieved a validation accuracy of 90%. We also performed

comparative tests using traditional bag-of-words approach-based classifiers such as LinearSVC, MultinomialNB, BernoulliNB, LogisticRegressionCV, and others. Table 6.3 shows the accuracy of RoBERTa and different classifiers.

Table 6.3: Comparison of RoBERTa with Other Classifiers

System	Accuracy
RoBERTa BASE	90
LinearSVC	92
MultinomialNB	88
BernoulliNB	88
LogisticRegressionCV	92
PassiveAggressiveClassifier	94
NearestCentroid	72
AdaBoostClassifier	70
Perceptron	94

We can see that there are few classifiers such as LinearSVC (92%), LogisticRegressionCV (92%), PassiveAggressiveClassifier (94%), and Perceptron (94%) that achieve high accuracy than the RoBERTa model. However, even with little less accuracy, RoBERTa was able to classify better than these classifiers. Table 6.4 shows few headlines (taken from the Indian dataset) where RoBERTa classified correctly and other classifiers failed.

It is clear from the above Table 6.4 that RoBERTa performed better. It might be possible that for some headlines, the traditional classifier may classify correctly compare to RoBERTa. The study of this possibility and its possible explanation can be a separate research topic in itself. Studies that specially focus on the comparative analysis of traditional and transformer-based sentiment classification can investigate it further. Presently, transformer-based models have produced superior results in various NLP

Table 6.4: Example of RoBERTa and Other Classifier’s Predictions

Headlines	RoBERTa Predicted Sentiments	Other Classifiers Predicted Sentiments
Asian shares slip on faltering hopes for COVID vaccines	Negative	Positive
COVID-19 infection offers protection from reinfection for at least 6 months: Study	Positive	Negative
Maharashtra: Counterfeit masks using leading brand’s name major concern for manufacturer; cases, raids in many cities	Negative	Positive
Travel restrictions challenge vaccine rollout, airlines warn	Negative	Positive
Crude oil prices extend gains on COVID vaccine hopes, OPEC+	Positive	Negative

tasks and are considered state-of-art. Because of this reason, we opted for RoBERTa to predict the sentiment of our whole dataset. Following is the detailed description of sentiment classification in all four countries individually.

6.3.2 Sentiment Analysis

RQ 4, RQ 5 and RQ 6 of COVID-19 pandemic is answered in this section.

6.3.2.1 Sentiment Analysis of India’s COVID-19 Headlines

Indian dataset came out to be almost balanced. Overall, the Indian dataset has 24,017 (50.87%) negative and 23,193 (49.12%) positive headlines. When it comes to the sentiments of various topics, all of the topics followed a different trend than overall sentiments. The US-related headlines have the highest percentage of negative with 58.22%. In contrast, the Economy category has 57.02% positive headlines. In Figure 6.4, we can see various topics’ sentiments in the Indian dataset.

During the first year of pandemic (2020), the extent of the pandemic in India was comparatively less than the other developed countries. Per day infections in India reached at peak in mid-September 2020 (registered little less than 100,000 infections),

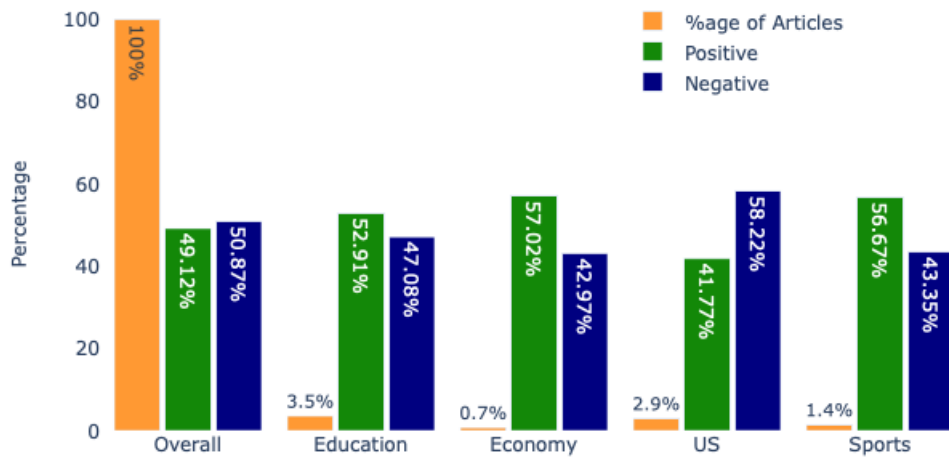


Figure 6.4: Overall and Various Topic's Sentiment in India Dataset

and since then it continued a downward trend till march 2021². This can explain the reason for balance (almost equal percentage of negative and positive headlines) nature of overall sentiments in Indian dataset. Meanwhile, other developed nations were devastated by second and third wave of pandemics. US was the worst affected country as in early January 2021, the third wave reached its peak (registered more than 300,000 infections)³. Hence, we see that high number of US related news headlines (58%) are negative. Researchers globally tried (are still trying) to understand Indian case but till now could not completely explained the reasons⁴ [93]. With these observation we can say that our sentiment analysis results from Indian dataset correspond with the actual situation.

²<https://www.worldometers.info/coronavirus/country/india/>

³<https://www.worldometers.info/coronavirus/country/us/>

⁴<https://frontline.thehindu.com/cover-story/lessons-from-the-first-covid19-wave-india-paradox-of-high-coronavirus-infection-rate-but-low-case-rate-low-case-fatality-rate/article34377410.ece>

6.3.2.2 Sentiment Analysis of Japan's COVID-19 Headlines

Japan's dataset turns out to be more negative tilted. As out of 21,038 Japan's headlines, 12,073 (57.38%) were negative, and the remaining 8965 (42.61%) were positive. Sports turns out to be following a different pattern than the rest of the topics. All topics, excluding sports, were $42\% \pm 2\%$ (for positive) and $57\% \pm 2\%$ (for negative). While the sports category has 54.91% positive and 44.08% negative headlines. On the other hand topic of education received the most negative (59.48%) than any other topic. Figure 6.5 shows sentiments of various topics in Japan's dataset.

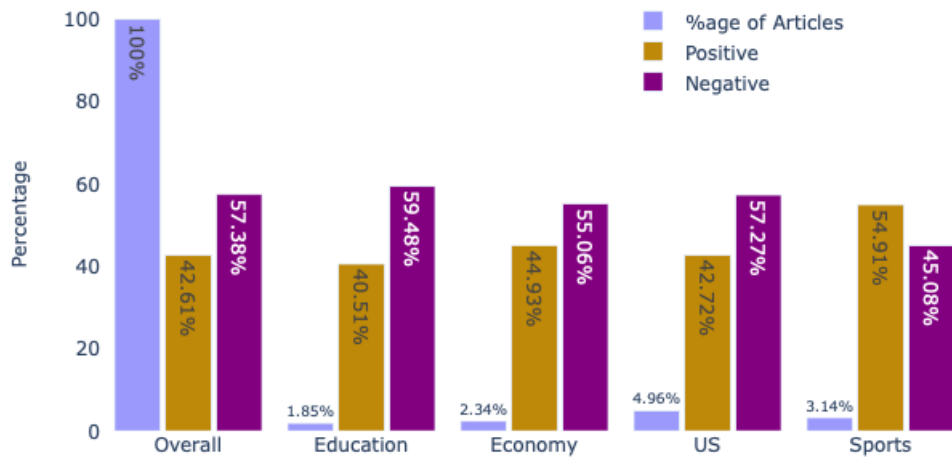


Figure 6.5: Overall and Various Topic's Sentiment in Japan's Dataset

When performance of developed western countries was compared with the developed economies of East Asia (Hong Kong, Japan, Singapore, South Korea, and Taiwan), latter outperformed former. However, within these East Asian countries, Japan's performance was lower than other four countries. Japan appeared to be less prepared compared to other countries [94]. Our results from Japan's dataset shows that it is tilted towards negative (57%), but not highly negative. We can say that our result correlate with the real situation of Japan.

6.3.2.3 Sentiment Analysis of South Korea's COVID-19 Headlines

Out of 10,055 headlines in our South Korean dataset, 54.47% (5477) were positive, and the remaining 4578 headlines (45.52%) were negative. These results make the South Korean dataset most positive among the four nations that we investigated. When we inspected various topics' sentiments, we found out that the economy-related topic was the most positive, with 60.95% of the headlines being positive. In comparison, the US and sports topics turned out to be the most negative, with 53.16% and 53.57% negative headlines, respectively. Figure 6.6 presents the sentiments of various topics in the South Korean dataset.

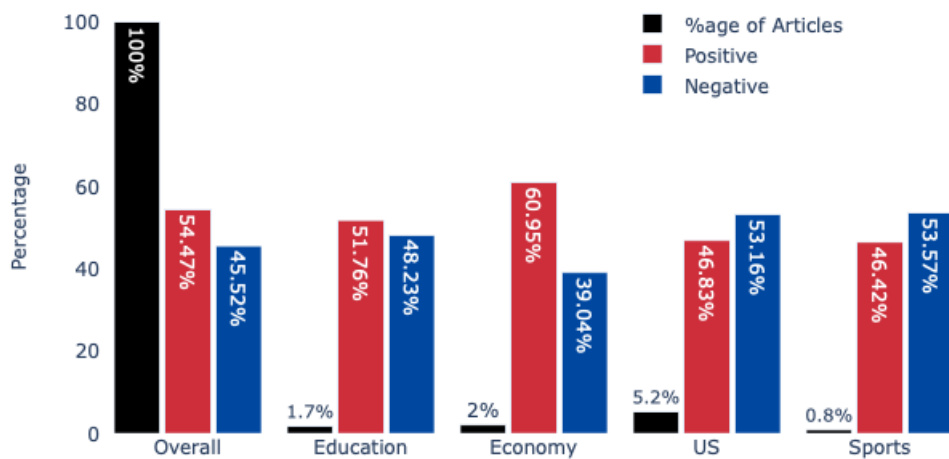


Figure 6.6: Overall and Various Topic's Sentiment in South Korea's Dataset

South Korea is one of the most successful country in world in containing the spread of Coronavirus⁵. The number of deaths per million in South Korea is one of the lowest⁶ among develop countries. Our results also confirm the success of South Korea as it is the only country in our dataset that have more positive headlines than negative. Even among the five East Asian countries (Hong Kong, Japan, Singapore, South Korea, and

⁵<https://ourworldindata.org/covid-exemplar-south-korea>

⁶<https://www.worldometers.info/coronavirus/>

Taiwan), South Korea is the most successful in containing the COVID-19 pandemic out of four countries [94]. This verify that our results correspond to the real situation. Further, our results shows that around 61% of the economy related headlines are positive. This indicate that South Korea's economy must have performed better (comparatively) than other similar economies. According to various global organizations/think tanks such as World Economic Forum (WEF) and International Monetary Fund (IMF), South Korea have the smallest downturn among the G-20 countries⁷ and it is doing better than other Organisation for Economic Co-operation and Development (OECD) countries^{8 9}. These reports confirm our results which shows high percentage of positive economy related headlines.

6.3.2.4 Sentiment Analysis of UK's COVID-19 Headlines

Overall, out of 23,821 headlines, 17,445 (73.23%) were negative, and the remaining 6376 (26.76%) were positive. This result shows high negative sentiments in UK media about the COVID-19 pandemic and other related issues. When we investigated the sentiments of the topics that were common in all four countries based on Table ?? in chapter 5, we found out that the US has more negative (74.5%) than the overall percentage. In comparison, sports were the least negative (71.09%). Figure 6.7 shows the overall and various topic sentiment of the UK's dataset. It is clear from Figure 6.7 that the various other topics' sentiments almost (slight fluctuation) follow the trend of the overall sentiments in the UK's dataset.

Sentiment analysis results from UK dataset shows that overall and all the four topics related headlines are highly negative (around two-third). During 2020, UK had

⁷<https://www.weforum.org/agenda/2020/08/south-korea-covid19-government-pandemic-response/>

⁸<https://www.imf.org/en/News/Articles/2021/04/29/na042921-mountains-after-mountains-korea-is-containing-covid-19-and-looking-ahead>

⁹<https://foreignpolicy.com/2020/09/16/coronavirus-covid-economic-impact-recession-south-korea-success/>

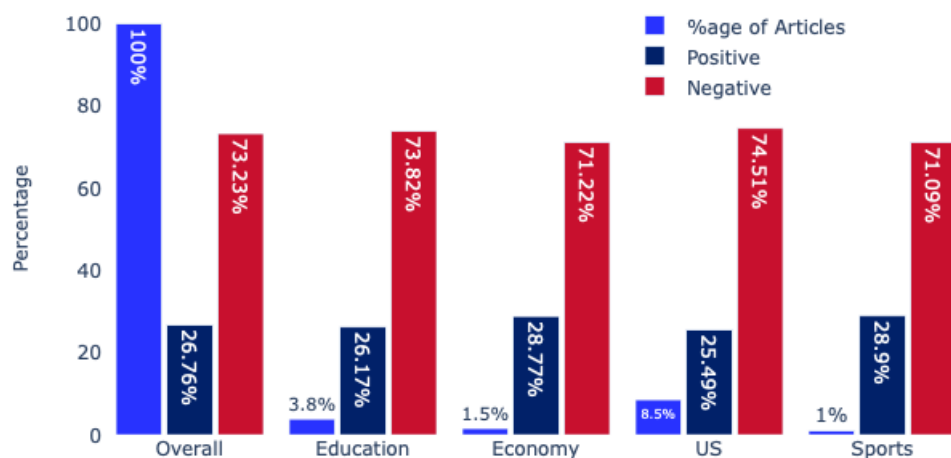


Figure 6.7: Overall and Various Topic's Sentiment in UK's Dataset

two wave of COVID-19 pandemic. The first and second wave peaked in late April and November respectively. UK was one of the worst affected country in terms of per millions death worldwide¹⁰ and in our dataset of four countries, it is the worst affected country. Further, OECD projected that UK would be hardest hit economy among the developed countries¹¹. Hence, it is not surprising that high percentage of news headlines are negative. With this, once again our results correspond to the real world situation.

6.3.3 Comparative Analysis of All Four Nation's Sentiments

When we evaluate our complete COVID-19 dataset of 102,124 headlines, we discovered that 56.9% (58,113 headlines) were negative, and the rest of 44,011 headlines (43.1%) were positive. These results show (not surprisingly) that overall there was more negative news about the COVID-19 pandemic. The UK and South Korea are the most negative and positive countries, respectively. The UK has almost three-quarters of news (73.23%) that were negative. In comparison, South Korea has 54.47% of positive news, which is

¹⁰<https://www.worldometers.info/coronavirus/>

¹¹<https://www.bbc.com/news/business-52991913>

more than 10% higher than the overall percentage.

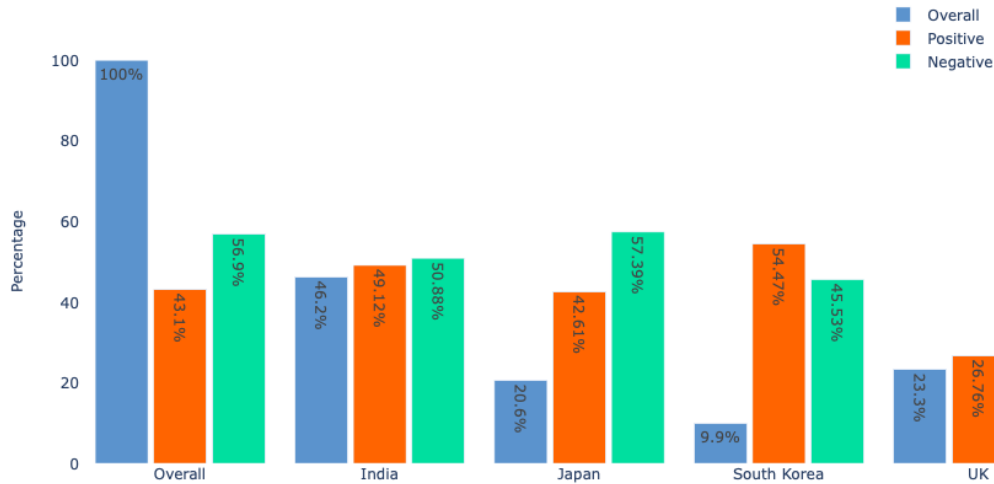


Figure 6.8: Comparison of Countries COVID-19 News Headlines Sentiments

Interestingly, out of the four countries that we studied, the UK is the worst affected country (deaths per million - 1320). At the same time, South Korea is the most successful country to tackle the COVID-19 pandemic, with only 25 deaths per million¹². India is the only developing country in our research. While in terms of infected people, India is the 2nd worst affected country after the US, but when taking deaths/million into consideration, it is better than most developed countries. Our results show that the Indian dataset is the most balanced dataset with 50.87% negative and 49.13% positive headlines. Figure 6.8 shows the comparison of all four countries' sentiments.

6.4 Limitations

Though we analyzed more than 100,000 news headlines related to COVID-19 from four countries, it is impractical to cover all the facets of such a vast domain. Sentiment classification has two variants. Binary, where sentiment is classified as positive and

¹²<https://www.worldometers.info/coronavirus/>

negative, while the other is multi-class, where text sentiment can be divided into more than two classes. For example, three classes where sentiment is classified into positive, negative, and neutral. This research has focused on classifying news headlines into either positive or negative classes (binary). However, certainly, there is a possibility that multi-class sentiment classification (positive, negative, and neutral) may produce better results.

Further, in this research, we have collected COVID-19 related news articles either from digital English language newspapers (for UK and India) or from the newspaper's English website (for Japan and South Korea). Lastly, data collection also faced limitations such as paywall restrictions, frequent changes in the website's format, and other web-scraping-related issues. Also, we have only used two keywords: Coronavirus and COVID-19, to collect news headlines.

Chapter 7

Conclusion

This research has three key components- NLP, Cybersecurity, and COVID-19 pandemic. The first component (NLP) is utilized to understand the other two components (Cybersecurity and COVID-19 pandemic). As the amount of digital data (both structured and unstructured text data) increases, the decluttering of this data can bring out critical patterns, themes, issues, and topics. These patterns/themes would be either impossible or too laborious to find by humans. NLP turns out to be a boon for this process, making it an important technique for the 21st century.

This research started with two objectives in mind: 1) Showing the application of NLP methods, and 2) Understanding and analyzing Cybersecurity and the COVID-19 pandemic. In chapters 4, 5, and 6, we utilized various NLP techniques or methods to comprehend and analyze a large number of Cybersecurity and COVID-19 related news headlines and articles. This application of NLP tools provided several themes, patterns, important issues, and sentiment (negative or positive) from the collected articles. This would be next to impossible without the usage of NLP tools. Further, our analysis of Cybersecurity and COVID-19 news also highlighted critical and interesting information.

7.1 Cybersecurity

News headlines are changing every minute, which makes the life of news remarkably short. At the same time, the amount of data is also increasing at an exponential pace. Finding critical information from such large data is like looking for a needle in a haystack. However, NLP techniques made the impossible possible. 'Russian interference in 2016 US election' and 'Huawei 5G issue' dominated the discourse for one year from April 2018 to March 2019. Also, an interesting pattern was discovered. While media from all nations except the US extensively reported the US's issues, the US media only focused on the domestically critical issue. This shows the US as the biggest influencer around the world generally and specifically in the Cybersecurity area. We also found that other than these dominant issues, all countries have different priorities based on their national interest. Japan is keeping a close eye on Vietnam's cybersecurity-related activities; voting machines' health keeps the Americans on their toes; cryptocurrencies give Indian authorities headache and other issues. These issues are critical but hidden due to an excess of information. The sentiment analysis section tried to decipher the Cybersecurity news sentiment. News articles' sentiment analysis is different from social media text. News sentiment can enlighten analysts about the media's biases and national mood towards a certain event. Our research showed that Russian interference in the US election garnered negative sentiment everywhere. On the other hand, news sentiment towards Huawei is positive, which is different from the popular sentiment (negative). This trend is different from Donald Trump's agenda to portray Huawei and its 5G technology negatively. Though this research only analyzed one-year articles but still successfully manifested the application of NLP in analyzing news articles. The methodology of this research and, to some extent, results can also help analysts and researchers by showing NLP applications and the interdisciplinary nature of Cybersecurity. Policymakers can also benefit from this research as it clearly shows each nation's

sentiments on critical cybersecurity issues. Using a similar methodology, researchers from other fields can also analyze different subjects.

Further, our Cybersecurity case study finds out that the US is the biggest influencer in Cybersecurity as all nation's newspapers extensively reported the US-related events and topics. Russian interference in the 2016 US election, Huawei 5G issue, US-China trade dispute, and US election-related news were extensively reported by almost all the countries we analyzed. Nations' geopolitical and strategic interests were also understood from the analysis of a nation's news. Presence of North Korea-related news in Japan and South Korea. Another interesting pattern is seen in the US Cybersecurity dataset. Unlike the other five countries, US media extensively focused on domestic issues as there was no foreign issue present in the top ten. Sentiment analysis of the Cybersecurity dataset finds out that overall, there was more positive news than negative. The UK with 32.1% and the US with 31.5% negative articles are the most negative dataset, and the overall average of negative news is 26%. In sentiment analysis of various critical topics, Russian interference in the 2016 US election turns out to be the most negative topic with an average of 48.6% negative articles. This is very high than the overall negative percentage. This result shows that the Russian interference issue garnered high negative sentiments.

7.2 COVID-19 Topic Modeling

The COVID-19 pandemic is a new healthcare crisis, and its impact is felt by people worldwide. Because of this pandemic's huge scale and its impact on the social, political, economic, and geopolitical spheres, it becomes critical to understand this pandemic's nature and sentiment. The SARS-CoV-2 virus is highly infectious and spread through the respiratory route when an infected person coughs, sneezes, sings, talks or breath. To prevent the spread of this virus, governments worldwide shut down their countries, and

even all the activities were stopped except the essential services. By April 2020, about half of the world's population was under lockdown, with more than 3.9 billion people in more than 90 countries or territories ordered to stay at home. In this situation, news media, especially digital news (newspapers or news channels), was the main source to know about the COVID-19 situation inside the country and around the world. Our research tried to understand what kind of COVID-19 news was prevalent, whether a nation's strategic interest also influences the COVID-19 news, or whether there is any correlation between the negative news and the impact of the COVID-19 pandemic on a country.

Topic modeling is a popular method to generate themes, issues, and patterns in a large textual dataset. This is either impossible to do manually or very time-consuming, laborious, and requires huge human resources. This COVID-19 case study, where more than 100,000 news articles were analyzed with the help of topic modeling, is a case point that proves the vitality of this approach. Our topic modeling experiment and analysis show that the US, Economy, Education, and Sports are among the most widely reported issues in all four countries.

In the first part of this research - topic modeling - we used the top2vec model and produced topics for each country. The number of topics turns out to be directly proportionate to the number of articles, as India, with the highest number of articles (47,342), produced the highest number of topics (402). Further, our descriptive analysis of the top ten topics across all four nations showed that the US, Economy, Education, and Sports are the most common issues. The presence of the US in the top ten in all countries shows the significance of the US. While education, economy, and sports can be seen as the worst affected sectors during this pandemic. Other than the results mentioned above, this analytical study helps us understand the nature of newsmaking. By looking at the topics in different countries, we can see that even COVID-19 news reporting was based on the overall national interest and strategic significance. For

example, the US's presence in all four countries, Australia-related news in the UK media, North and South Korea-related news in Japan's media.

Further, we also applied k-means clustering on the topic vectors generated by top2vec model. The objective was to classify the topics into clusters to observe the sparsity and distribution of topics. These results showed that Japan's 255 topics could be classified into only two clusters, making Japan's COVID-19 dataset most concentrated. On the other hand, South Korea, with the smallest dataset (10,076 articles), is classified into 11 clusters, which makes it the most sparsely distributed dataset in this case study.

We also compared NMF and Top2Vec topic modeling approaches. Top2Vec is a new model (proposed in 2020) but based on neural networks and utilizes word embedding representation. Utilization of these advanced concepts makes the results (generated topics) coherent and easy to comprehend. However, the NMF algorithm though fairly old (popularized in 1999), uses traditional bag-of-words representation, which does not consider the context. Top2Vec does not require preprocessing and advance knowledge of the number of topics, while NMF requires both. To overcome the problem of finding the optimal number of topics, this research utilized the TC-W2C metric measure. When the results from both topic modeling approaches are finally compared, we find no similarity (only one or two common topics in two countries), at least among the top ten topics.

The future work would try to work on the methods to determine the quality of topics produced by both topics modeling approach. Lastly, we would also try to overcome the limitations of this research, as mentioned in section 5.4

7.3 COVID-19 Sentiment Analysis

This research presents the sentiment classification and analysis of COVID-19 news headlines. For this, we utilized the state-of-the-art RoBERTa model for sentiment clas-

sification of headlines. Our model achieved 90% validation accuracy and was able to classify headlines better than other traditional classifiers. Our implementation of the RoBERTa model for sentiment classification on our complete dataset showed that the UK has the most negative news (73.23%). In comparison, South Korea was the most positive country with 54.47% positive news. These results correspond with the real situation of the COVID-19 pandemic in these countries, as the UK is one of the worst affected countries in terms of deaths/million. South Korea is one of the best-performing countries with only 15 deaths/million. With these results, this study can be used as a template to study the COVID-19 news worldwide, helping us discover and understand the critical issues and their representative sentiments in COVID-19 news in those countries. Our sentiment analysis result points towards a possible correlation between negative COVID-19 news and a countries' affectedness. For example, in our dataset, the UK is the worst affected country and has the highest negative headlines. This correlation can be further tested by deploying the same or improved methodology by other researchers. The results and subsequent analysis can also help researchers understand the newsmaking process or the cross-cultural socio-political-economic COVID-19 impact. While most COVID-19 sentiment analysis researches focus on social media posts, our study presents an alternative perspective to this by introducing COVID-19 news headlines sentiment analysis.

The future work would first try to overcome the limitations of this research. New and interesting methods have been proposed to detect emotion [20] and sentiment [21] by using models such as LSTM and lexicon-based convolutional neural network. We would also try to explore these models in our future studies. Lastly, we would add more countries to our dataset to enlarge this research scope to make it more wide-ranging and global.

Bibliography

- [1] B. D. Prasad, “Content analysis,” *Research methods for social work*, vol. 5, pp. 1–20, 2008.
- [2] O. Pearman, M. Boykoff, J. Osborne-Gowey, M. Aoyagi, A. G. Ballantyne, P. Chandler, M. Daly, K. Doi, R. Fernández-Reyes, I. Jiménez-Gómez *et al.*, “Covid-19 media coverage decreasing despite deepening crisis,” *The Lancet Planetary Health*, vol. 5, no. 1, pp. e6–e7, 2021.
- [3] K. Krishnan and S. P. Rogers, *Social data analytics: Collaboration for the enterprise*. Newnes, 2014.
- [4] A. M. Turing, “Computing machinery and intelligence,” in *Parsing the turing test*. Springer, 2009, pp. 23–65.
- [5] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” *arXiv preprint arXiv:1609.08144*, 2016.
- [6] I. Annamoradnejad, “Colbert: Using bert sentence embedding for humor detection,” *arXiv preprint arXiv:2004.12765*, 2020.
- [7] S. D. Reese, “Prologue–framing public life,” *Framing public life. Perspectives on media and our understanding of the social world*. Oxon: Routledge, 2001.

- [8] L. Kappelman, V. Johnson, R. Torres, C. Maurer, and E. McLean, “A study of information systems issues, practices, and leadership in europe,” *European Journal of Information Systems*, vol. 28, no. 1, pp. 26–42, 2019.
- [9] C. Wang, P. W. Horby, F. G. Hayden, and G. F. Gao, “A novel coronavirus outbreak of global health concern,” *The lancet*, vol. 395, no. 10223, pp. 470–473, 2020.
- [10] F. Krammer, “Sars-cov-2 vaccines in development,” *Nature*, vol. 586, no. 7830, pp. 516–527, 2020.
- [11] G. Forni and A. Mantovani, “Covid-19 vaccines: where we stand and challenges ahead,” *Cell Death & Differentiation*, vol. 28, no. 2, pp. 626–639, 2021.
- [12] V. Chandrashekhar, “1.3 billion people. a 21-day lockdown. can india curb the coronavirus,” *Science*, vol. 10, 2020.
- [13] P. Ghasiya and K. Okamura, “COVID-19 News Articles,” 2021. [Online]. Available: <https://dx.doi.org/10.21227/gdq8-ej60>
- [14] T. M. Georgescu, “Natural language processing model for automatic analysis of cybersecurity-related documents,” *Symmetry*, vol. 12, no. 3, p. 354, 2020.
- [15] C. L. Jones, R. A. Bridges, K. M. Huffer, and J. R. Goodall, “Towards a relation extraction framework for cyber-security concepts,” in *Proceedings of the 10th Annual Cyber and Information Security Research Conference*, 2015, pp. 1–4.
- [16] H. Gasmi, J. Laval, and A. Bouras, “Information extraction of cybersecurity concepts: An lstm approach,” *Applied Sciences*, vol. 9, no. 19, p. 3945, 2019.
- [17] O. Mendsaikhan, H. Hasegawa, Y. Yamaguchi, and H. Shimada, “Identification of cybersecurity specific content using the doc2vec language model,” in *2019*

- IEEE 43rd annual computer software and applications conference (COMPSAC)*, vol. 1. IEEE, 2019, pp. 396–401.
- [18] M. R. Alagheband, A. Mashatan, and M. Zihayat, “Time-based gap analysis of cybersecurity trends in academic and digital media,” *ACM Transactions on Management Information Systems (TMIS)*, vol. 11, no. 4, pp. 1–20, 2020.
- [19] F. Kolini and L. Janczewski, “Clustering and topic modelling: A new approach for analysis of national cyber security strategies,” in *Pacific Asia Conference on Information Systems (PACIS)*. Association For Information Systems, 2017.
- [20] B. Gupta, S. Sharma, and A. Chennamaneni, “Twitter sentiment analysis: An examination of cybersecurity attitudes and behavior,” *Proceedings of the 2016 Pre-ICIS SIGDSA/IFIP WG8*, vol. 3, 2016.
- [21] K. Al-Rowaily, M. Abulaish, N. A.-H. Haldar, and M. Al-Rubaian, “Bisal—a bilingual sentiment analysis lexicon to analyze dark web forums for cyber security,” *Digital Investigation*, vol. 14, pp. 53–62, 2015.
- [22] T.-J. Kim, “Covid-19 news analysis using news big data: Focusing on topic modeling analysis,” *The Journal of the Korea Contents Association*, vol. 20, no. 5, pp. 457–466, 2020.
- [23] M. Costola, M. Nofer, O. Hinz, and L. Pelizzon, “Machine learning sentiment analysis, covid-19 news and stock market reactions,” 2020.
- [24] K. Krawczyk, T. Chelkowski, S. Mishra, D. Xifara, B. Gibert, D. J. Laydon, S. Flaxman, T. Mellan, V. Schwämmle, R. Röttger *et al.*, “Quantifying the online news media coverage of the covid-19 pandemic,” *medRxiv*, 2020.
- [25] W. Poirier, C. Ouellet, M.-A. Rancourt, J. Béchar, and Y. Dufresne, “(un) covering the covid-19 pandemic: Framing analysis of the crisis in canada,”

- Canadian Journal of Political Science/Revue canadienne de science politique*, vol. 53, no. 2, pp. 365–371, 2020.
- [26] W. Jo and D. Chang, “Political consequences of covid-19 and media framing in south korea,” *Frontiers in public health*, vol. 8, 2020.
- [27] Q. Liu, Z. Zheng, J. Zheng, Q. Chen, G. Liu, S. Chen, B. Chu, H. Zhu, B. Akinwunmi, J. Huang *et al.*, “Health communication through news media during the early stage of the covid-19 outbreak in china: digital topic modeling approach,” *Journal of medical Internet research*, vol. 22, no. 4, p. e19118, 2020.
- [28] T. de Melo and C. M. Figueiredo, “Comparing news articles and tweets about covid-19 in brazil: sentiment analysis and topic modeling approach,” *JMIR Public Health and Surveillance*, vol. 7, no. 2, p. e24585, 2021.
- [29] Y. Bai, S. Jia, and L. Chen, “Topic evolution analysis of covid-19 news articles,” in *Journal of Physics: Conference Series*, vol. 1601, no. 5. IOP Publishing, 2020, p. 052009.
- [30] E. De Santis, A. Martino, and A. Rizzi, “An infoveillance system for detecting and tracking relevant topics from italian tweets during the covid-19 event,” *IEEE Access*, vol. 8, pp. 132 527–132 538, 2020.
- [31] S. Noor, Y. Guo, S. H. H. Shah, P. Fournier-Viger, and M. S. Nawaz, “Analysis of public reactions to the novel coronavirus (covid-19) outbreak on twitter,” *Kybernetes*, 2020.
- [32] J. Samuel, G. Ali, M. Rahman, E. Esawi, Y. Samuel *et al.*, “Covid-19 public sentiment insights and machine learning for tweets classification,” *Information*, vol. 11, no. 6, p. 314, 2020.

- [33] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, "Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets," *IEEE Access*, vol. 8, pp. 181 074–181 090, 2020.
- [34] M. Huang, H. Xie, Y. Rao, Y. Liu, L. K. Poon, and F. L. Wang, "Lexicon-based sentiment convolutional neural networks for online review analysis," *IEEE Transactions on Affective Computing*, 2020.
- [35] S. Boon-Itt and Y. Skunkan, "Public perception of the covid-19 pandemic on twitter: sentiment analysis and topic modeling study," *JMIR Public Health and Surveillance*, vol. 6, no. 4, p. e21978, 2020.
- [36] G. Barkur and G. B. K. Vibha, "Sentiment analysis of nationwide lockdown due to covid 19 outbreak: Evidence from india," *Asian journal of psychiatry*, vol. 51, p. 102089, 2020.
- [37] S. Das and A. Dutta, "Characterizing public emotions and sentiments in covid-19 environment: A case study of india," *Journal of Human Behavior in the Social Environment*, pp. 1–14, 2020.
- [38] R. Chandrasekaran, V. Mehta, T. Valkunde, and E. Moustakas, "Topics, trends, and sentiments of tweets about the covid-19 pandemic: Temporal infoveillance study," *Journal of medical Internet research*, vol. 22, no. 10, p. e22624, 2020.
- [39] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, and T. Zhu, "Public discourse and sentiment during the covid 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter," *PloS one*, vol. 15, no. 9, p. e0239441, 2020.

- [40] R. Xie, S. K. W. Chu, D. K. W. Chiu, and Y. Wang, “Exploring public response to covid-19 on weibo with lda topic modeling and sentiment analysis,” *Data and Information Management*, vol. 5, no. 1, pp. 86–99, 2021.
- [41] D. Sarkar, *Text analytics with Python: a practitioner’s guide to natural language processing*. Apress, 2019.
- [42] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”, 2009.
- [43] G. A. Miller, “Wordnet: a lexical database for english,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
- [44] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala, “Latent semantic indexing: A probabilistic analysis,” *Journal of Computer and System Sciences*, vol. 61, no. 2, pp. 217–235, 2000.
- [45] T. Hofmann, “Probabilistic latent semantic indexing,” in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 50–57.
- [46] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.
- [47] J. D. Lafferty and D. M. Blei, “Correlated topic models,” *Advances in neural information processing systems*, vol. 18, pp. 147–154, 2006.
- [48] P. Paatero, U. Tapper, P. Aalto, and M. Kulmala, “Matrix factorization methods for analysing diffusion battery data,” *Journal of Aerosol Science*, vol. 22, pp. S273–S276, 1991.
- [49] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

- [50] S. Arora, R. Ge, and A. Moitra, “Learning topic models—going beyond svd,” in *2012 IEEE 53rd annual symposium on foundations of computer science*. IEEE, 2012, pp. 1–10.
- [51] D. Angelov, “Top2vec: Distributed representations of topics,” *arXiv preprint arXiv:2008.09470*, 2020.
- [52] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, “Singular value decomposition and principal component analysis,” in *A practical approach to microarray data analysis*. Springer, 2003, pp. 91–109.
- [53] G. E. Hinton, “Distributed representations,” 1984.
- [54] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.
- [55] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *International conference on machine learning*. PMLR, 2014, pp. 1188–1196.
- [56] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” *arXiv preprint arXiv:1802.03426*, 2018.
- [57] L. McInnes and J. Healy, “Accelerated hierarchical density based clustering,” in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2017, pp. 33–42.
- [58] O. Maimon and L. Rokach, “Data mining and knowledge discovery handbook,” 2005.
- [59] B. Liu, *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press, 2020.

- [60] C. S.-G. Khoo, A. Nourbakhsh, and J.-C. Na, "Sentiment analysis of online news text: a case study of appraisal theory," *Online Information Review*, 2012.
- [61] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 4, no. 1, 2010.
- [62] X. Zhang, H. Fuehres, and P. A. Gloor, "Predicting stock market indicators through twitter "i hope it is not as bad as i fear"," *Procedia-Social and Behavioral Sciences*, vol. 26, pp. 55–62, 2011.
- [63] Y. Liu, X. Huang, A. An, and X. Yu, "Arsa: a sentiment-aware model for predicting sales performance using blogs," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 607–614.
- [64] S. Asur and B. A. Huberman, "Predicting the future with social media," in *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, vol. 1. IEEE, 2010, pp. 492–499.
- [65] B. Liu *et al.*, "Sentiment analysis and subjectivity." *Handbook of natural language processing*, vol. 2, no. 2010, pp. 627–666, 2010.
- [66] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams engineering journal*, vol. 5, no. 4, pp. 1093–1113, 2014.
- [67] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." in *Lrec*, vol. 10, no. 2010, 2010, pp. 2200–2204.

- [68] C. Hutto and E. Gilbert, “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*, vol. 8, no. 1, 2014.
- [69] S. Loria, “textblob documentation,” *Release 0.15*, vol. 2, 2018.
- [70] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, “Recursive deep models for semantic compositionality over a sentiment treebank,” in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.
- [71] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, “Learning word vectors for sentiment analysis,” in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011, pp. 142–150.
- [72] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” *CS224N project report, Stanford*, vol. 1, no. 12, p. 2009, 2009.
- [73] D. O’callaghan, D. Greene, J. Carthy, and P. Cunningham, “An analysis of the coherence of descriptors in topic modeling,” *Expert Systems with Applications*, vol. 42, no. 13, pp. 5645–5657, 2015.
- [74] C. Sievert and K. Shirley, “Ldavis: A method for visualizing and interpreting topics,” in *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, 2014, pp. 63–70.
- [75] M. A. Cox and T. F. Cox, “Multidimensional scaling,” in *Handbook of data visualization*. Springer, 2008, pp. 315–347.

- [76] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.
- [77] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in python,” *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.
- [78] S. Bradshaw and P. N. Howard, *The global disinformation order: 2019 global inventory of organised social media manipulation*. Project on Computational Propaganda, 2019.
- [79] D. Nadeau and S. Sekine, “A survey of named entity recognition and classification,” *Linguisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.
- [80] J.-M. Torres-Moreno, *Automatic text summarization*. John Wiley & Sons, 2014.
- [81] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.
- [82] G. Vaidyanathan, “India will supply coronavirus vaccines to the world-will its people benefit?” *Nature*, pp. 167–168, 2020.
- [83] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.
- [84] G. Eysenbach, “Sars and population health technology,” *Journal of Medical Internet Research*, vol. 5, no. 2, p. e14, 2003.
- [85] K. Goldschmidt, “The covid-19 pandemic: Technology use to support the well-being of children,” *Journal of Pediatric Nursing*, 2020.

- [86] R. Singh, R. Singh, and A. Bhatia, "Sentiment analysis using machine learning technique to predict outbreaks and epidemics," *Int. J. Adv. Sci. Res*, vol. 3, no. 2, pp. 19–24, 2018.
- [87] A. Alamoodi, B. Zaidan, A. Zaidan, O. Albahri, K. Mohammed, R. Malik, E. Almahdi, M. Chyad, Z. Tareq, A. Albahri *et al.*, "Sentiment analysis and its applications in fighting covid-19 and infectious diseases: A systematic review," *Expert systems with applications*, p. 114155, 2020.
- [88] G. K. Shahi, A. Dirkson, and T. A. Majchrzak, "An exploratory study of covid-19 misinformation on twitter," *Online Social Networks and Media*, vol. 22, p. 100104, 2021.
- [89] K. Sharma, S. Seo, C. Meng, S. Rambhatla, A. Dua, and Y. Liu, "Coronavirus on social media: Analyzing misinformation in twitter conversations," *arXiv preprint arXiv:2003.12309*, 2020.
- [90] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [91] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pre-training approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [92] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.
- [93] J. Cohen, "Is india's coronavirus death 'paradox' vanishing?" *Science*, vol. 372, no. 6542, pp. 552–553, 2021.

- [94] B. Y. An and S.-Y. Tang, “Lessons from covid-19 responses in east asia: Institutional infrastructure and enduring policy instruments,” *The American Review of Public Administration*, vol. 50, no. 6-7, pp. 790–800, 2020.

Publications

Journals (Peer Reviewed)

- [1] P. Ghasiya, and K. Okamura, “Understanding the Middle East through the eyes of Japan’s Newspapers: A topic modelling and sentiment analysis approach,” *Digital Scholarship in the Humanities*, March. 2021. doi:[10.1093/llc/fqab019](https://doi.org/10.1093/llc/fqab019).
- [2] 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, March. 2021. doi:[10.1109/ACCESS.2021.3062875](https://doi.org/10.1109/ACCESS.2021.3062875).

International Conferences (Peer Reviewed)

- [3] P. Ghasiya, and K. Okamura, “A Comparative Analysis of Japan and India COVID-19 News Using Topic Modeling Approach,” In: Sharma, H., Gupta, M.K., Tomar, G.S., Wang, L., (eds) *Communication and Intelligent Systems: Proceedings of ICCIS 2020*, Springer, Nature. doi:[10.1007/978-981-16-1089-9_36](https://doi.org/10.1007/978-981-16-1089-9_36).
- [4] P. Ghasiya, and K. Okamura, “Investigating Cybersecurity News Articles by Applying Topic Modeling Method,” *2021 International Conference on Informa-*

tion Networking (ICOIN), Jeju Island, Korea (South), pp. 432-438, Jan. 2021.
doi:[10.1109/ICOIN50884.2021.9333952](https://doi.org/10.1109/ICOIN50884.2021.9333952).

- [5] P. Ghasiya, and K. Okamura, “Comparative Analysis of Japan and the US Cybersecurity Related Newspaper Articles: A Content and Sentiment Analysis Approach,” In: Barolli L., Amato F., Moscato F., Enokido T., Takizawa M. (eds) *Advanced Information Networking and Applications. AINA 2020. Advances in Intelligent Systems and Computing*, vol 1151. Springer, Cham. doi:[10.1007/978-3-030-44041-1_39](https://doi.org/10.1007/978-3-030-44041-1_39).

Forthcoming Publication

- [6] P. Ghasiya, and K. Okamura, (Forthcoming). “A Hybrid Approach to Analyze Cybersecurity News Articles by Utilizing Information Extraction Sentiment Analysis Methods.” *International Journal of Semantic Computing*. (Accepted for Publication)