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

ガーシヤ, 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：

（別紙様式 2）

氏　　名　：ガーシヤ　ピユシュ

論 文 名　　: Research on Analysis of Social Phenomena in the Cyberspace by Utilizing Data-driven Approach
(　データ駆動型アプローチを利用したサイバースペースにおける社会現象の分析に関する研究　)

区　　分　：甲

# 論文内容の要旨

The era of big data is providing many opportunities for conducting impactful research. 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 – is gaining popularity over traditional hypothesis-driven also known as theory-driven approach. Some of the advantages or contribution 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.

Out of the generated big data, 80% 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 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. For this purpose, we have selected two social phenomena: 1) Cybersecurity and 2) the COVID-19 pandemic, as our case study.

Before introducing our case studies, first it is important to discuss two key challenges of this research. The first challenge is related to topic modeling. In topic modeling, the specification of optimal number of topics is required up front to produce 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 TC-W2V metric measure 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.

Cyberspace is a defining feature of modern life and the growing cybersecurity threats or issues are capable in not only disrupting the individual life but also pose a challenge for national security. More than 1000% increase in the coverage of cybersecurity related stories in New York Times in the past

10 years explain the criticality of cybersecurity. For that reason, investigating cybersecurity news is a timely exercise. Similarly, the COVID-19 pandemic is a health crisis which impacted more than 200 countries or territories worldwide. For the last one year, newspapers are full of COVID-19 stories. Other than the medical experts, virologist, and epidemiologist, researchers from various other fields are trying to comprehend the impact of this pandemic on economy, society, and human life. Our research would explore the COVID-19 related news to filter critical information from large data.

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.