PRAGMATIC APPROACHES TOWARD AUTOMATED EXTRACTION AND UNDERSTANDING OF LARGE SCALE HEALTH-RELATED TEXTS

by

Tianyi Xie

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2021

Approved by:

Dr. Yaorong Ge

Dr. Shi Chen

Dr. Albert Park

Dr. Jing Yang

ABSTRACT

TIANYI XIE. Pragmatic approaches toward automated extraction and understanding of large scale health-related texts. (Under the direction of DR. YAORONG GE)

Natural language processing has become a very popular tool in many areas and has also drawn great attention in the community of health informatics. It is a series of processes that allows informaticists to take advantage by extracting and understanding the information hidden in the unstructured text. Such a process can help clinicians making more accurate decisions, filtering more useful information, and better understand public health-related social norms. However, due to the uniqueness of the health-related content, the regular workflow of NLP has limited application because of the challenges in the annotation. The thesis primarily focuses on shortening the gap in annotation by integrating deep learning NLP approaches in the workflow to reduce the task in annotation, or to realize semi-automatic and automatic annotation in certain tasks. In this dissertation, I first present a deep learning-based phenotyping system that allows extraction of blood pressure readings from unstructured clinical notes. The workflow employs a pre-filtering approach that can reduce the workload in annotation, and can be applied in different domains. The second part presents an extractive text summarization system that utilizes the information in the abstract of scientific publications. The system uses a self-supervised approach that does not require any annotation while generating a classifier that can detect the content in the body text of the publication which should be extracted. In the third part, I proposed a workflow that performs info-surveillance on social media about COVID-19. By using a small group of annotated social media posts, the workflow will be able to monitor the trend and sentiment of the different topics being discussed on social media based on different times and locations.

ACKNOWLEDGEMENTS

This work would not be possible without the help from many people. I would first like to share my gratitude to my doctoral advisor, Dr. Yaorong Ge. I really appreciate all the guidance and help you gave over the years and all the advises you gave to my research. Your support really means a lot to me. I am really luck to have a mentor like you.

I would like to thank Dr. Shi Chen for being my co-advisor. It is a pleasure to work with you on solving real-world problems, and I really learned a lot. Thank you for the ideas and suggestions you gave during my research.

I would like to acknowledge my dissertation committee members: Dr. Albert Park, and Dr. Jing Yang. Thank you for playing such essential role in my dissertation research and your time, advice, and help.

In addition, I would like to thank Dr. Yi Zhen, Dr. Maryam Tavakoli, Dr. Tianqi Li and Ms. Chuqin Li for the collaboration. It is a pleasure to work with you to solve research questions.

Also, I would like to thank Dr. Yaorong Ge and Dr. Shi Chen for providing the financial support. I would like to thank the Graduate Assistant Support Plan (GASP) at the University of North Carolina at Charlotte.

Last but not least, I would like to thank my family. My wife and my parents who have been supporting me physically and mentally. My daughters who filled the home with joy. I am always grateful to have you in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BERT Bidirectional Encoder Representations from Transformers

CBOW Continuous Bag-of-Words

CNN Convolutional Neural Networks

CT-BERT COVID-Twitter-BERT

EHR Electronic Health Record

GloVe Global Vectors for Word Representation

LDA Latent Dirichlet Allocation

NLP Natural Language Processing

NPIs Nonpharmaceutical Interventions

PPV Positive Predictive Value

RNN Recurrent Neural Networks

TF-IDF Term frequencyâInverse document frequency

VADAR Valence Aware Dictio-nary and sEntiment Reasoner

CHAPTER 1: INTRODUCTION

In recent years, the unstructured health-related texts have grown exponentially [1], and have caught researchersâ attention in applying Natural Language Processing (NLP) methods to extract and understand useful information from the unstructured texts [2, 3]. However, these health-related texts are presented in different forms, which can be presented as short social media posts or long scientific publications [4]. In addition, these texts are often noisy and incomplete. Thus, it posts challenge to informaticists to retrieve useful information from the texts. Traditionally, NLP methods relies on frequency based statistical methods to form a series of NLP process [5, 6, 7]. More recently, deep learning-based methods [8, 9]are gaining popularity in the research community, as it shows significant performance gain compared to the conventional NLP models. The key difference between deep learning-based methods and the conventional methods is deep learning models perform feature engineering automatically [10]. As a result, end-to-end learnings can be performed without feature engineering, which often requires exceptional domain knowledge.

Although deep learning methods provide convenience to the researchers, some problems remain to be solved. Most deep learnings models require the tuning of a large set of parameters, thus often require a large set of training data.[11] This means a good amount of resources are placed into the annotating of the texts[12, 13], which is time consuming and often requires the domain knowledge. In addition, the form of health-related unstructured text varies significantly, as a result, it will be difficult to develop an effective and efficient model that fit all cases.

Researcher community was well aware of the problems and have attempted to develop solutions to address the problems. These include deep learning models that allows transfer learning[**bert**, 14], active learning with small samples[15], and self-supervised models[16]. Many of the models has shown significant performance gain compare with conventional models, and significant reduction of labor-intensive annotation, and automation in constructing the NLP processes. There are still potentials that could further refine these models to achieve the goal of extraction and understanding the useful information of the health-related unstructured texts.

In this dissertation, we propose to (1) establish a deep-learning based system for accurate extraction of blood pressure data in clinical narratives, (2) create a self-supervised extractive text summarization workflow for biomedical literatures, (3) develop a BERT-based active learning model to monitor the trend of COVID-19 related social media posts. In this dissertation, different types of health-related texts were explored, which include Electronic Health Records (EHR), scientific publications, and social media posts. Different deep learning models were utilized, which include Convolutional Neural Network (CNN), Bidirectional Transformer (BERT) and its variant. Different downstream NLP task were solved, which include phenotype detection, text summarization, topic classification, and sentiment analysis. With the combination of different types of texts, models, and tasks, novel automated workflows were developed to achieve the downstream tasks with different texts and models.

1.1 Types of health-related Texts

To this date, there have been various type of health-related texts that researchers has been working on. Traditionally, EHR was one of the most used data sources for clinical informatics research[17]. EHR has useful structured information include diagnoses, prescriptions, treatment plans, radiology images, and laboratory and test results. These types of information are useful as they are readily available and can be extracted directly. In addition to the structured information, lots of the clinical valuable information are hidden in the unstructured texts. The texts include discharge summaries, progress notes, orders, procedure reports, and many other kinds of specific documentations, while the type of information being extracted can vary significantly. More importantly, the quality of the text can vary significantly as the text comes from different sources[18]. It is extremely difficult to build one NLP model to extract information form all kinds of clinical documents.

While EHR provide important clinical information, scientific publications also offer important information. Those publications usually contain information like the most up-to-date research results and guidelines. Generally, the text quality of the scientific publications is relatively high, however, the information is still scattered in the text that requires extraction[19].

On the other hand, social media has recently become an important source for researchers to harvest public health related data. It has been observed that the public is very interested in participating the discussion in health-related topics. The social media allows researchers to extract publicâs opinions and perceptions on certain topic, thus allows policy makers to adjust the policy accordingly[20]. Generally, the data quality of the social media is still questionable, however, with the availability of the large quantity of data, it is still possible to plot a trend in a scale of time or geolocations.

Table 1.1: Comparing the three types of health-related texts in different aspects.

Type of Text	Quality	Complexity	Easy to get
EHR	Low	High	Difficult
Publications	High	High	Easy
Social Media	Low	Low	Easy

Overall, as shown in Table 1.1, the different types of the health-related texts have different inherit features, making a single model approach difficult. It also suffers from typical natural language challenge of being context and domain specific, ambiguous, erroneous, informal, and synonymatic. Researchers have been keeping develop new models and tools to address the challenge, and many new approaches has already show great performance in many tasks.

CHAPTER 2: A DEEP-LEARNING BASED SYSTEM FOR ACCURATE EXTRACTION OF BLOOD PRESSURE DATA IN CLINICAL NARRATIVES

2.1 Abstract

This study presents a novel workflow for identifying and analyzing blood pressure readings in clinical narratives using a Convolution Neural Network. The network performs three tasks: identifying blood pressure readings, determining the exactness of the readings, and then classifying the readings into three classes: general, treatment, and suggestion. The system can be easily set up and deployed by people who are not experts in clinical Natural Language Processing. The validation results on an independent test set show the first two of the three tasks achieve a precision, recall, and F-measure over or close to 95%, and the third task achieves an overall accuracy of 85.4%. The study demonstrates that the proposed workflow is effective for extracting blood pressure data in clinical notes. The workflow is general and can be easily adapted to analyze other clinical concepts for phenotyping tasks.

2.2 Background

It is estimated that in 2015, 84% of all non-federal acute care hospitals had adopted a Basic EHR (Electronic health record) with clinician notes. 96% of non-federal acute care hospitals have possession of an EHR certified by HHS (United States Department of Health Human Services), and the percentage has held though 2017[1]. The quick adoption of EHR presents a challenge as a great portion of data is often recorded in narrative clinical notes[2], where these notes are unstructured and require an extra step of extraction. Whereas the information in clinical notes is often viewed as the key to solving the problems of improving the quality of care, clinical decision support, and clinical research[3]. Multiple systems[4, 5, 6] of clinical NLP (Natural Language Processing) pipelines have been developed to address the challenges. These pipelines can be used for different tasks, which include named entity recognition, event extraction, relation extraction, etc. And can be used to code the diagnoses, diseases, medications, and lab results. However, these pipeline requires extensive knowledge in NLP, and may not perform well in very specific tasks. More recently, deep learning[7] models are getting attention among health informaticians, as they show improved performance on multiple tasks including clinical phenotyping over the existing clinical NLP pipelines[8, 9].

2.3 Case Description

In this study, our goal is to identify and extract blood pressure readings in clinical notes. Blood pressure is considered vital signs of a patient, and the readings are often stored in the clinical notes in an unstructured format. They can be precious especially considering that we can gather a patient's blood pressure readings over an extensive period, and use them to monitor patient's health status and provide assistance in making clinical decisions. Additional information is often associated with blood pressure readings, which can help us understand the purpose of the readings. Thus, in this study, we will demonstrate a system that can identify the blood pressure readings in the clinical narrative notes, determine if the reading is vague or exact, and classify the purpose of the reading.

2.4 Method

2.4.1 Data Preprocessing and Preparation

In this study, we used a de-identified dataset of electronic medical records of cardiology patients from Wake Forest Baptist Medical Center, which in total contained 101696 clinical notes. We then randomly selected 3000 clinical notes for further processing.

After manually reviewing these clinical notes, we found about 15% of the notes containing one or more blood pressure readings. It would be time-consuming to annotate all 3000 notes. Thus, we developed search patterns using certain keywords to find potential blood pressure readings. We randomly selected 1000 clinical notes from the 3000 notes, and had all blood pressure readings annotated. We then used 500 of the notes to develop the search patterns, which were designed to find all potential blood pressure readings including false-positive cases. The patterns were tested on the other half of 1000 notes, and the result showed it had covered all blood pressure readings. The developed search patterns are in (Table 2.1). The search patterns were then applied to the 3000 notes, when search patterns found a match in the clinical notes, a sub-string of 80 characters were extracted, tokenized, and any incomplete token at the beginning and the end was removed. In total, 1465 potential blood pressure readings were found.

After annotating the potential blood pressure readings, we found 881 of the 1465 notes containing true blood pressure readings. We further annotated these by:

1. Whether the blood pressure reading is exact or is a range.

2. The purpose of the reading, whether it is a general blood pressure reading, or it is a reading after certain drug treatment, or it is a suggestion on how blood pressure should be controlled.

The example of the two types has been listed in the (Table 2.2).

Table 2.1: Search Patterns and Examples

Search Pattern	Example	Reading
bp	with good effect. On day of discharge BP 126/63 resting and 133/67 with ambulation; HR 59. Patient instructed to take	Yes
	The patient did pass SBP on 09/22 but secondary to his altered mental status it was felt he was not	No
blood.{,10}pressure	on Date of Discharge: Vital Signs: Blood pressure 145/73 pulse 78 temperature 97.7 F (36.5 C) temperature source	Yes
	checks 3. Continued home diltiazem for blood pressure control 4. Continued home divalproex 250 mg BID PO 5. Continued	No
rvsp	with severe pulmonary hypertension and RVSP of 73.7 mmHg. No significant valvular vegetations or shunt. CT PE protocol	Yes
right ventricular	pulmonary hypertension. Estimated right ventricular systolic pressure is 52 mmHg. Estimated right atrial	Yes
mmhg|mm hg	(*) 7.350 - 7.450 PCO2 42.5 35 - 45 mm Hg PO2 76.0 (*) 80 - 100 mm Hg HCO3 12.1 (*) 22 - 26 mmol/L BASE	No

Table 2.2: Blood pressure notes classification

Example	Exact	Purpose
on Date of Discharge: Vital Signs: Blood pressure 121/69 pulse 83 temperature 98.9 F (37.2 C) temperature source	Yes	General
Diltiazem and metoprolol given SBP down to 130s overnight - Restarted metoprolol succinate 100 mg daily given	No	Treatment
and mag replaced. Monitor INR. Keep SBP > 120 for brain perfusion. No need for PM per EP cont to hold BB/CCB for	No	Suggestion

In the 881 actual blood pressure readings, 666 of them were exact readings, which could be a reading with both systolic and diastolic blood pressure, or either one of the two. In terms of the purpose of reading, 776 of the readings were classified as general-purpose reading, 40 of them were readings after treatment, and 65 of them were classified as readings for suggestions.

For each classification task, the notes were randomly assigned to the training, validation, and test set with a ratio of 4:3:3. For the classification tasks having imbalanced data, efforts must be made to ensure the imbalance would not impact the effectiveness of the classier. In this study, excessive data that may dominate the classification in training and validation were removed to ensure the model will not be biased, while the test case retain the ratio of different classes. Thus, leaving 430 total case for the exact readings task, and 170 total cases for the purpose of the reading task. The test set was independent of the training and validation process so that it could reflect the real-world performance of the system.

2.4.2 Classification System

To identify the blood pressure readings and classify the notes into different types, we have developed a CNN (Convolutional Neural Networks) model. The input to CNN was word embedding of the blood pressure reading candidates. Each blood pressure reading was transferred into a 2D-matrix using the word embedding algorithm described in [10, 11]. In this process, each word in a note was transformed into a vector as reference using a GloVe[10] representation. We used a GloVe model trained from Wikipedia 2014 and Gigaword 5. The vectors were then grouped into an 80 by 100 matrix. This matrix served as the input of the model.

Once we have a matrix representation of the notes, we can then use the data to train the model. In this study, a three-layer CNN model (Table 2.3) has been built for the classification tasks. We used in total 3 layers of 1-Dimensional convolution, and each one is coupled with a max pooling layer for dimension reduction. Softmax and ReLu were used as the activation function in different layers, which provided nonliterary and efficiency during the model training. As for other hyper-parameters, we set the number of epoch to 15, and a batch size of 8.

Table 2.3: CNN Layers

Layer (type)	**Output Shape**	**Activation Function**
Embedding	80*100	
conv1d_1 (Conv1D)/(MaxPooling)	79*128/39*128	ReLu
conv1d_2 (Conv1D)/(MaxPooling)	38*256/19*256	ReLu
conv1d_3 (Conv1D)/(MaxPooling)	18*256/6*256	Softmax
flatten_1 (Flatten)	1536	
dense_1 (Dense)	128	Relu
dense_2 (Dense)	# of Class	Softmax

2.4.3 Evaluation

We used 30% of the available data in each task as the test sets, and the test sets were isolated from the training and validating process. The system generated the predicted output of classification for each note, then it was compared with previously annotated labels. Since some certain tasks had imbalanced data, as a result, in those tasks some classes may have relatively small dataset.

Precision, Recall, F-measure, and Accuracy were calculated for each task and each type of class, where:

$$Positive\ Predictive\ Value(Precision) = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$True\ Positive\ Rate(Recall) = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$True\ Negative\ Rate(Specificity) = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}$$

2.5 Results

Figure 2.1 shows the relationship between the number of epoch and loss, and the number of epoch and accuracy. From the figure, we noticed that for identifying blood pressure readings and classifying if readings are exact, the model converged after 5 epochs. While for the task of identifying the purpose of blood pressure readings, due to the limited number of training and validating cases, the model converged after 12 iterations. The accuracy plot shows the first two tasks perform well with accuracy around 0.95 on the validation set, while the third task has relatively low accuracy compared with the first two. The figure also shows that the models are relatively stable, and there are no signs of over-fit.

We then performed a test using the test set (Table 2.4). The model being used is the trained model after 15 epochs. For the task of identifying the blood pressure readings, we had an overall accuracy of 0.963, with the precision, recall, and F-measure for each class maintained over 0.95. For the task of identifying the exactness of readings, we had an overall accuracy of 0.936, with the precision, recall, and F-measure around 0.95 for the exact reading, and over 0.85 for the non-exact reading. For the task of classifying the purpose of the reading, the overall accuracy is 0.826, and the recall are relative acceptable, while the precision were low for some classes because of heavily skewed data.

Figure 2.1: Epoch vs. Loss and Accuracy

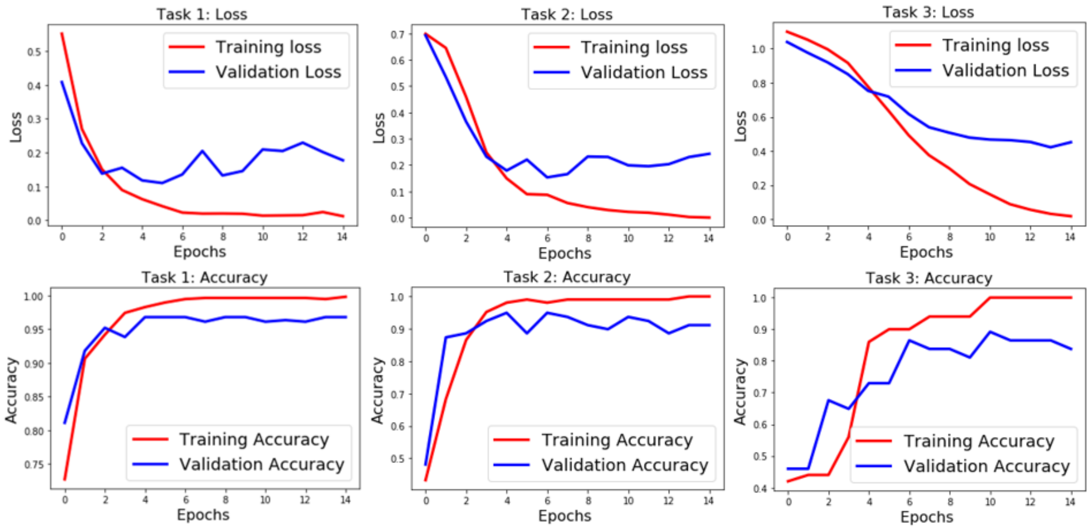


Table 2.4: Test set result

Task	**Class**	**Total #**	**Precision**	**Recall**	**F-Measure**	**Accuracy**
1.Reading	Yes	247	0.96	0.98	0.97	0.963
	No	193	0.97	0.95	0.96	
2.Exact	Yes	195	0.97	0.94	0.96	0.936
	No	70	0.85	0.91	0.88	
3.Purpose	General	239	0.99	0.83	0.90	0.826
	Suggestion	17	0.43	0.71	0.53	
	Treatment	9	0.23	0.89	0.36	

2.6 Discussion

In this paper, we present an approach to identify the blood pressure readings and classify the readings. There have been various efforts made to extract the blood pressure readings, or extract blood pressure related information from clinical notes. One approach is to use a rule-based system[12] or regular expressions[13] to extract the blood pressure information from clinical notes. The benefit of such systems is that for someone familiar with text processing and clinical context, such systems can be built in a short time and can achieve relatively good results. However, such systems can also be inflexible. If there is a new note pattern, or the way of using certain keywords has been changed, the rules need to be updated and re-evaluated. Another approach is to use certain clinical NLP systems[14, 5, 15] to map and extract medical terms in the clinical notes. These systems use a pipeline of NLP modules, usually with machine learning models and rule-based models, coupled with bio-medical vocabularies[16], to perform tasks like medical term mapping and extraction. Such systems also have drawbacks. These systems are designed for more general tasks, where they often under-performs in very specific tasks. We noticed that some systems may not be able to identify "bp" as blood pressure in their general workflow. Additional fine-tuning is required if these systems are to be used for a specific task.

Our system used a novel workflow that can solve the issue of some existing systems. It first used certain keywords to find notes that are related to blood pressure. These keywords can be very general, and false-positive cases are allowed in this process. Then the long clinical notes are fragmented into small pieces, and only the ones contain potential blood pressure readings will be recorded. These notes can be annotated at a relatively fast pace since they are selected which contains blood pressure information and are relatively short. After the annotation, by using the annotated data as the training and validating set, we can create a CNN model that identifies the blood pressure information in the clinical notes. This model does not require someone to look into the clinical notes extensively to find certain patterns and summarize them into rules. This can be a huge benefit as such systems can be deployed by someone who is not a text processing or NLP expert.

In the result section, we have demonstrated that our system had a satisfactory outcome on the first two tasks. This shows that the CNN model can extract the features and text patterns that are associated with blood pressure. We also noticed the slow convergence on the third task, and the result seems to be not as good as the previous two tasks. We reckon this is because the heavily skewed dataset cannot provide sufficient samples to train the model properly. And the precision can be heavily influenced if the dataset is not well balanced. The recall shows the model was able to extract the classes with very small proportions. We believe if we can bring in more data for this task, the model can be substantially improved. Additionally, the annotation was done by one of the author. For the first and second task, the annotation was very clear, however, for the third task, annotation can be subjective if only annotated by one annotator. Although the model did clearly reflect the choice of the annotator, a dataset with good quality would definitely benefit the training of the predictive model.

Currently, in our system, the notes were embedded by GloVe, which is a general-purpose word to vector representation. We believe it is more appropriate to use a Word2Vec that is specifically trained from the biomedical domain. This approach ensures the word embeddings are more representative of the clinical notes. Another thing we want to investigate in the future is to other up-to-date deep learning models like BERT[17] or Bio-BERT[18] to perform the tasks.

In conclusion, our study demonstrated that the workflow we designed can efficiently extract focused information from clinical narrative documents. Our system can be built and deployed at a fast pace and does not require extensive feature hunting or rule generation.

2.7 Contribution

The major contribution of this part is to demonstrate an approach of using NLP and deep learning models with minimal annotation and human intervention to perform information extraction for clinical narratives. The result shows such approach can outperform state of the art models in extracting blood pressures, and such approach can also be applied for different information extraction tasks.

Charpter 2 Reference

[1] *Non-federal Acute Care Hospital Electronic Health Record Adoption*. 2017 (accessed July 10, 2019). URL: https://dashboard.healthit.gov/quickstats/pages/FIG-Hospital-EHR-Adoption.php.

[2] George Hripcsak et al. "Unlocking Clinical Data from Narrative Reports: A Study of Natural Language Processing". In: *Annals of Internal Medicine* 122.9 (May 1995), pp. 681–688. ISSN: 0003-4819. DOI: 10.7326/0003-4819-122-9-199505010-00007. eprint: https://annals.org/acp/content_public/journal/aim/19829/0000605-199505010-00007.pdf. URL: https://doi.org/10.7326/0003-4819-122-9-199505010-00007.

[3] John Hornberger. "Electronic Health Records: A Guide for Clinicians and Administrators". In: *JAMA* 301 (Jan. 2009), pp. 110–110. DOI: 10.1001/jama.2008.910.

[4] Alan R Aronson and FranÃ§ois-Michel Lang. "An overview of MetaMap: historical perspective and recent advances". In: *Journal of the American Medical Informatics Association* 17.3 (May 2010), pp. 229–236. ISSN: 1527-974X. DOI: 10.1136/jamia.2009.002733. eprint: http://oup.prod.sis.lan/jamia/article-pdf/17/3/229/6021555/17-3-229.pdf. URL: https://doi.org/10.1136/jamia.2009.002733.

[5] Guergana K Savova et al. "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications". In: *Journal of the American Medical Informatics Association* 17.5 (Sept. 2010), pp. 507–513. ISSN: 1527-974X. DOI: 10.1136/jamia.2009.001560. eprint:

http://oup.prod.sis.lan/jamia/article-pdf/17/5/507/5940551/17-5-507.pdf. URL: https://doi.org/10.1136/jamia.2009.001560.

[6] Ergin Soysal et al. “CLAMP â a toolkit for efficiently building customized clinical natural language processing pipelines”. In: *Journal of the American Medical Informatics Association* (Nov. 2017), ocx132. DOI: 10.1093/jamia/ocx132.

[7] Jung Hoon Son et al. “Deep Phenotyping on Electronic Health Records Facilitates Genetic Diagnosis by Clinical Exomes”. In: *The American Journal of Human Genetics* 103 (June 2018). DOI: 10.1016/j.ajhg.2018.05.010.

[8] Sebastian Gehrmann et al. “Comparing Rule-Based and Deep Learning Models for Patient Phenotyping”. In: (Mar. 2017).

[9] Sebastian Gehrmann et al. “Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives”. In: *PLOS ONE* 13 (Feb. 2018), e0192360. DOI: 10.1371/journal.pone.0192360.

[10] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. “Glove: Global vectors for word representation”. In: *In EMNLP*. 2014.

[11] Tomas Mikolov et al. “Efficient Estimation of Word Representations in Vector Space”. In: *CoRR* abs/1301.3781 (2013).

[12] Jitendra Jonnagaddala et al. “HTNSystem: Hypertension Information Extraction System for Unstructured Clinical Notes”. In: *Technologies and Applications of Artificial Intelligence*. Ed. by Shin-Ming Cheng and Min-Yuh Day. Cham: Springer International Publishing, 2014, pp. 219–227. ISBN: 978-3-319-13987-6.

[13] Alexander Turchin et al. “Using Regular Expressions to Abstract Blood Pressure and Treatment Intensification Information from the Text of Physician Notes”. In: *Journal of the American Medical Informatics Association* 13.6 (Nov. 2006), pp. 691–695. ISSN: 1527-974X. DOI: 10.1197/jamia.M2078. eprint: http:

//oup.prod.sis.lan/jamia/article-pdf/13/6/691/2201107/13-6-691.pdf. URL: https://doi.org/10.1197/jamia.M2078.

[14] Hua Xu et al. “MedEx: a medication information extraction system for clinical narratives”. In: *Journal of the American Medical Informatics Association* 17.1 (Jan. 2010), pp. 19–24. ISSN: 1527-974X. DOI: 10.1197/jamia.M3378. eprint: http://oup.prod.sis.lan/jamia/article-pdf/17/1/19/6025347/17-1-19.pdf. URL: https://doi.org/10.1197/jamia.M3378.

[15] Li Zhou et al. “Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to Process Medication Information in Outpatient Clinical Notes”. In: *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2011 (Jan. 2011), pp. 1639–48.

[16] Olivier Bodenreider. “The Unified Medical Language System (UMLS): Integrating Biomedical Terminology”. In: *Nucleic acids research* 32 (Feb. 2004), pp. D267–70. DOI: 10.1093/nar/gkh061.

[17] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *arXiv e-prints*, arXiv:1810.04805 (Oct. 2018), arXiv:1810.04805. arXiv: 1810.04805 [cs.CL].

[18] Jinhyuk Lee et al. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. In: *Bioinformatics* (Sept. 2019). btz682. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz682. eprint: http://oup.prod.sis.lan/bioinformatics/advance-article-pdf/doi/10.1093/bioinformatics/btz682/30132027/btz682.pdf. URL: https://doi.org/10.1093/bioinformatics/btz682.

CHAPTER 3: SELF-SUPERVISED EXTRACTIVE TEXT SUMMARIZATION FOR BIOMEDICAL LITERATURES

3.1 Abstract

In this study, we propose a self-supervised approach to extractive text summarization for biomedical literature. The approach uses abstracts to find the most informative content in the article, then generate a summary for training a classification model. The Sentences in the abstract and literature were first embedded using BERT. A similarity-based model was then applied to label the informative sentences for training the classifier. We used logistic regression as our classification model and used the features of sentence embedding for the classification. The results showed the feasibility of employing the abstract to perform self-supervised training of a classification model to generate extractive summarization. This approach can enable automatic generation of one or two page executive summaries of biomedical literature to keep clinicians and biomedical researchers up to date with the latest development.

3.2 Introduction

An enormous amount of information exists in biomedical literature and is still growing exponentially [1]. Clinicians need to retrieve and integrate available clinical information from biomedical literature to support evidence-based medicine. Biomedical researchers also need to understand the state-of-the-art of their domain from the scientific literature. Due to increasing research studies and the high demand for facilitating evidence-based medicine, it is becoming difficult for both clinicians and researchers to maintain and adopt the most up-to-date information effectively and efficiently. As text mining and Natural Language Processing (NLP) being widely applied in the biomedical domain, text summarization has become a desirable approach to retrieving, conveying and managing valuable information contained in biomedical literature. However, training a biomedical text summarizer is a high-cost task that typically requires domain experts to manually perform a large amount of annotation and curation of the rapidly increasing biomedical literature. We were thus motivated to propose an automatic text summarization method based on self-supervised learning to eliminate the need for high-cost manual annotation.

We designed a self-supervised learning based text summarizer for biomedical literature through embedding, sentence classification, and NLP methods. We investigated the feasibility of generating labels with self-supervised learning and applying pre-trained language representations for biomedical text summarization. In our study, we explored two hypotheses to perform text summarization tasks in biomedical literature. The first hypothesis is that we could use abstracts to navigate the most informative contents of each biomedical article, which means the generated summary should be composed of the most similar contents to the abstract. The other hypothesis is that sentences being selected for extractive summarization contain common linguistic features that are conclusive and definitive, and the common features can be represented by the pre-trained language representation. We trained the text summarizer as a sentence classifier to include key sentences into the generated summary. To further determine the pre-trained embedding model in the summarizer, we compared the performance of using BERT [2] with using BioBERT [3] as an intermediate representation. We measured the summarizerâs performance with the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [4] metric.

3.3 Related Work

Text summarization is the task of producing a concise and fluent summary while preserving key information content and conveying the main idea of the text. In the biomedical field, text summarization has been proposed as an efficient solution to advance information extraction, knowledge discovery, decision making, and evidence-based medicine[5]. There are two categories of text summarization methods: abstractive summarization and extractive summarization [6]. Abstractive summarization introduces novel words and phrases and generates a new shorter version that conveys the main idea of the input text. In contrast, extractive summarization selects original sentences that are highly relevant to the main idea of the input text to generate a coherent summary. In most cases, extractive summarization methods give better results compared to abstractive summarization methods [7]. Concerning the complexity of interpreting domain knowledge and intolerance of information distortion, more efforts have been put on extractive summarization in the biomedical field. Significant advances in text summarization techniques have been achieved in solving biomedical problems. Abstractive summarization are focusing on summarizing Medline citations to obtain decision support data [8], treatment on disease [9, 10],and drug adverse events-drug interactions [11]. Abstractive summarization methods applied in the biomedical field mostly produce graphical summaries [12, 13]. On the other hand, the majority of extractive summarization systems focus on producing textual summaries. Extractive summarization methods have been widely studied in the biomedical domain for different tasks [14], such as summarizing clinical notes [15, 16], patient-specific clinical evidence for decision making support [17, 18], electronic medical records [19, 20], radiology report [21] and clinical trial descriptions [22]. Text summarization can also be divided into two categories by the number of inputs: single-document summarization and multi-document summarization. Multi-document summarization is more challenging to generate a cohesive summary for heterogeneous documents [23]. In our study, we focus on a single-document extractive summarization method to obtain a summary for biomedical articles efficiently and effectively.

In general, an extractive summarization method consists of the following three steps: 1) represent input document into an intermediate representation; 2) score sentences based on the representation; 3) construct summary by selecting a set of key sentences. Various representation models, scoring algorithms, and selection strategies were employed to perform different biomedical tasks. The graph-based summarizers [24, 7, 25] represent input document into an undirected weighted graph, where vertices are the sentences and the links are the similarity relations between these sentences. Latent Semantic Analysis (LSA) is used to represent the importance of each sentence based on the presence of word combination patterns in text summarization [26]. LSA is a statistical method for extracting the contextual meaning of words and the similarity of sentences. However, both graphical representation and LSA-based representation lack semantic information, sentence structure, and contexture feature. Deep contextualized language models are imported into text summarization [27]. In this way, a text summarizer can effectively represent input documents with contextual features while without utilizing biomedical knowledge bases. Most text summarizers utilized clustering [28] to group sentences by sub-themes. Clustering-based summarizer has less training cost but lower performance than supervised learning-based summarizer [29]. We proposed a text summarization method with the contextualized language model and self-supervised learning.

3.4 Proposed Summarization workflow

The workflow of our summarization pipeline is illustrated in Figure 1. In the following sections, we will discuss how we built a classifier that enables self-trainning automatically to perform extractive summarization.

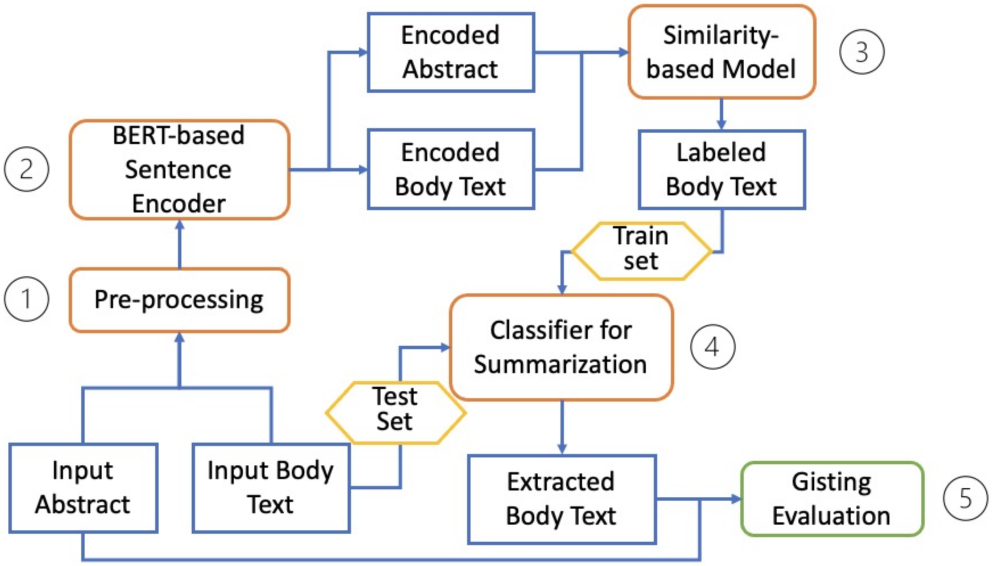


Figure 3.1: The workflow of proposed summarization pipeline.

3.4.1 Pre-processing of documents

The first step is to pre-process the document. Our study focuses on summarizing documents or published scientific articles in the biomedical domain. For each document, we extracted its abstract and body text separately, then tokenized them into sentences. We also applied pre-processing to clean the document by removing content in the parentheses, special characters, white spaces, references, and URLs. The pre-processing is done by using Natural Language ToolKit (NLTK) Library [30].

3.4.2 BERT-based Sentence Embedding

Once pre-processing has been applied to the documents, we apply a sentence level encoder to encode each sentence into vectors, which allows the extraction of certain linguistic features. Then the vector can be used to compare the similarity between the sentences in the abstracts and the sentence in the body text, and build classification models to extract the common linguistic features. In this study, we use the uncased BERT base model to encode the sentences. BERT has been proved effective in many downstream tasks in NLP, and pre-training can be applied to make the model domain-specific. Although many models including BioBERT and BlueBERT[31] have been pre-trained on using the cohort similar to our study, we still plan to use BERT base to show that our pipeline is generalizable, and the linguistic features being extracted in this study are not domain-specific.

For sentence-level encoding, we use the hidden state corresponding to the first token as the embedding of the sentence, which is a vector with a length of 768. This token is a special classification token and is used as the aggregate sequence representation of classification tasks.

3.4.3 Similarity-based model

To train a model that can classify whether the sentences need to be extracted from the document based on linguistic features, we need to first generate a dataset with such labels. To achieve this, we propose a novel approach that utilizes the abstract to find the sentences in the documents that need to be extracted. Since the abstract can be seen as a concise summarization of the document, there are strong similarities between the abstract and the sentences to be extracted. Thus, the workflow we propose are based on cosine similarity:

$$\cos(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = \frac{\sum_{i=1}^{n}\mathbf{x}_i\mathbf{y}_i}{\sqrt{\sum_{i=1}^{n}(\mathbf{x}_i)^2}\sqrt{\sum_{i=1}^{n}(\mathbf{y}_i)^2}} \tag{3.1}$$

The cosine similarity measures the similarity between vectors, and since all sentences are embedded by BERT in the form of vectors, thus the similarity between sentences can be measured. To find sentences with the highest similarity with the sentences in the abstract, the following algorithm is proposed to find sentences to be extracted:

Algorithm 1 Algorithm for Similarity-based Model

```
Create an empty set for $SentencetoExtract$
$NumofSentence$ = max(len($abstract$) * 2, len($document$) * 15%)
LOOP Process
for $x$ in $abstract$ do
    Find two sentences $y$ in document with highest similarity compared to $x$.
    Add the two sentences $y$ to the $SentencetoExtract$.
end for
while len($SentencetoExtract$) $<$ $NumofSentence$ do
    Find the sentence $y$ with highest similarity between any $x$ in $abstract$ and any
    $y$ in $document$ that are not in $SentencetoExtract$ .
    Add this sentence $y$ to $SentencetoExtract$.
end while
return $SentencetoExtract$
```

The algorithm will form a group of sentences that have high similarities compared with the sentences in the abstract, and the length of the group is the maximum of two times the length of the abstract or 15% length of the document. For a typical publication, it will generate 1 to 2 pages of extensive summary, which include more details directly quoted from the publication. We also tested for different length of the document, and we used elbow method to determine that 15% is the optimal, as it showed good coverage of the abstract and presented enough detail, while not being unnecessarily long.

3.5 Classification of sentences for summarization

In this section, we discuss the training of the classification model that can classify the sentences for summarization. By using the Similarity-based model, we were able to generate class labels for the sentences, where the classification model is trained based on the class labels. The purpose of the classification model is to identify if any sentences should be extracted to form the summarization. The classification model is purely based on the linguistic features of the sentence, which has been previously extracted and embedded by the BERT model.

The model we propose in this study is logistic regression. Although there are machine learning models that perform better in such tasks, the primary purpose of this study is to validate the hypotheses, to see if common linguistic patterns can be found among the embedding of the sentence. The performance of the classification will be measured the matrices including Precision, Recall, F-measure, and AUC-ROC [32].

3.6 Experiment

3.6.1 Data

In this study, we searched 300 Radiation Therapy-related scientific articles from PubMed published in 2020. 10 of the 300 publications were removed for missing abstracts or being incomplete. We follow the pre-processing procedure mentioned in the previous section and have all the sentences in the publication encoded using BERT. The ratio of training data and testing data is 7:3, where 203 randomly selected publications were used to train the classifier and 87 were used for testing. In total, there are 71943 sentences in the dataset, which including both abstract and body text. The detailed number of sentences is listed in Table 1.

Table 3.1: Number of total sentences

	Abstract	**Body text**
Training	2042	48628
Testing	858	20415
Total	2900	69043

3.6.2 Labeling of sentences

Once all the sentences were pre-processed and embedded, we followed the proposed similarity-based model to label the sentence based on the similarity of each sentence in the body text. We also performed a test based on the ROUGE score to make sure the combination of the labeled sentence is a good representation of the abstract. It is based on the 290 publications, and the detailed ROUGE scores are listed in Table 2.

Table 3.2: ROUGE scores on similarity-based model

	Recall	**F-score**
ROUGE-1	0.71	0.32
ROUGE-2	0.31	0.14
ROUGE-L	0.59	0.33

The result shows our approach of labeling the sentence for training had relatively good performance on the recall of ROUGE-1 and ROUGE-L, showed the summarization can be a good representation of the abstract. The relatively low F-1 score is within expectation, as our proposed plan is to generate a summary that is on average 3 to 4 times the length of the abstract. Overall, the test shows the proposed similarity-based model is capable of generating sentence labels for training the classification model.

3.6.3 Classification of sentence labels

Once we have the labeled sentences for training the classification, we follow the proposed method of using logistic regression for the classification. The feature being used is purely based on the linguistic features that BERT generated for each sentence. In total, 48,628 sentences were used for training the classifier and 20,415 are used for testing. Since the two classes are heavily imbalanced with a 0:1 ratio of approximately 6:1, we applied weight to the model to ensure the number of sentences being classified as 1 has the same ratio as the training data. In the end, the testing data shows the metrics for the classification are listed in Table 3.

Table 3.3: ROUGE scores on similarity-based model

	Precision	**Recall**	**F-score**	**Support**
0	0.89	0.88	0.88	17293
1	0.35	0.37	0.37	3122
accuracy	0.8			20415
weighted avg	0.8	0.8	0.8	20415

The classification result shows it has decent performance on the sentences that are not labeled to be extracted. However, since we train the classifier to classify the data based on its imbalance ratio, we were only able to capture some of the sentences to be extracted based on its linguistic features. Given the AUC score of 0.73, it shows the classifier can capture certain common linguistic features from the sentence to perform the classification.

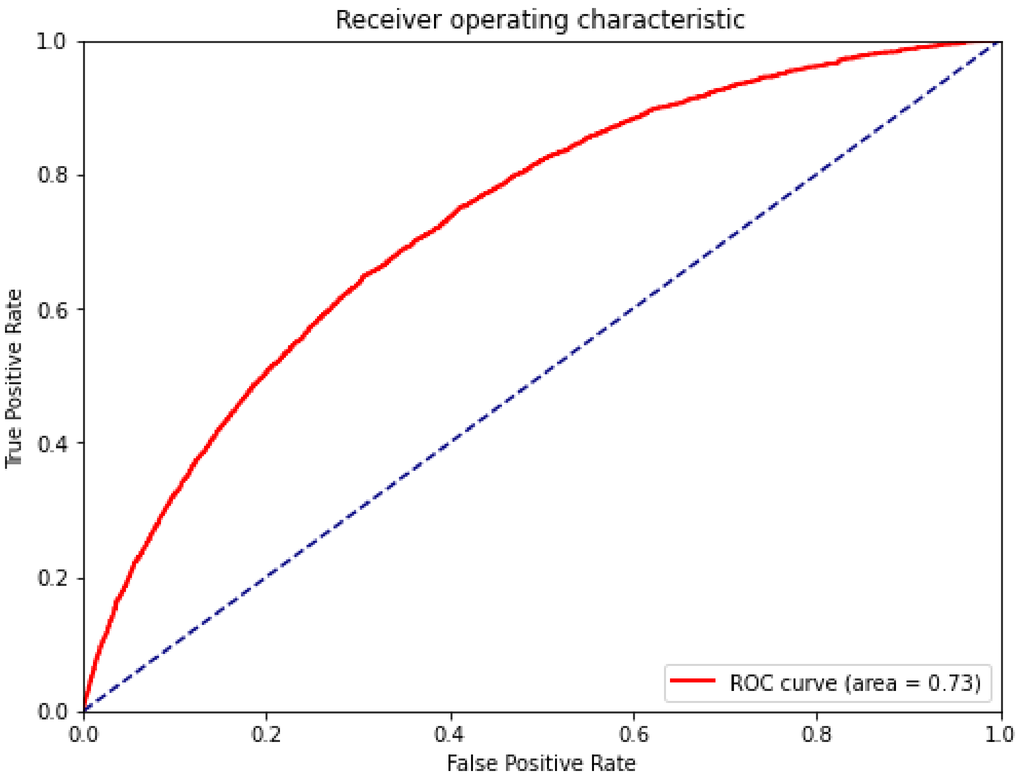


Figure 3.2: Receiver Operating Characteristic curve

3.6.4 Gisting Evaluation

The classifier was able to generate class labels for the testing publications. By combining the selected sentence, we were able to generate a summarization for the publication. We then conducted another ROUGE test on the classifier-generated summarizations.

Previously we mentioned that we used the similarity-based model to find the sentences to be extracted, we conducted a ROUGE test on the testing data using this method as the benchmark. We also randomly picked sentences from the testing data that match the number generated by the similarity-based model, which gives us the baseline performance when sentences are picked randomly. The result of the ROUGE score on each approach is listed in Table IV.

The result shows our classification model significantly outperforms randomly se-

Table 3.4: ROUGE scores on test set

	Similarity-based Model	Classifier Generated	Randomly selected
ROUGE-1 Recall	0.71	0.65	0.59
ROUGE-2 Recall	0.31	0.24	0.17
ROUGE-L Recall	0.58	0.52	0.45

lected sentences for summarization, while was not able to achieve the performance of the similarity-based model. We primarily focus on the recall of the ROUGE Score, to see if the summarization generated by the classifier has a good representation of the abstract. Precision and F-score are not the primary focus, as the length of the abstract and the length of the generated summary are disproportional to each other. Surprisingly, we noticed even with randomly selected sentences, we were able to get a ROUGE-1 score on recall of 0.59. We believe this is because of the disproportion between the abstract and generated summary, where the generated summary is much longer than the summary, resulting in it having very high recall while does not fully reflect the contextual information of the abstract. Still, the classifier-generated summary has significant improvement compared with randomly selected sentences for the summarization. Thus, the result demonstrated the feasibility of using this self-supervised workflow to training a classifier based on linguistics features to perform extractive text summarization.

3.7 Discussion

In this paper, we proposed a pipeline for extractive text summarization of biomedical literature. We partitioned the pipeline into three parts, namely document pre-processing, sentence embedding, and model training. In this section, we conclude our approach and discuss some potential future works for the latter two steps in this pipeline.

After document pre-processing, the sentences are translated into numerical vectors, or embedding, representing their linguistic features. Recent years have seen increasing research in this area. Traditional embeddings like FastText [33] or Word2Vec [34] serve for this purpose well. Later, context-dependent pre-trained models stand out for their ability to provide richer and more dynamic information. Among them, in this paper, we rely on the so-called BERT, a pre-trained Transformer model [35] which has been utilized for various NLP tasks. For each token, BERT learns information from its left and right side when pre-training. This bidirectionality ensures the context-dependence which is essential to better understand the accurate semantic meaning of the language. A significantly important property of BERT is the ability to inflect a word's distinct meanings in different context settings by providing unequal vector representations, whereas embeddings like Word2Vec fail to do so. Another advantage of BERT is that it is indeed not only an embedding – it is tunable. However, in this paper, we only benefited from BERT's expert comprehension of text, treating it as an embedding. One future direction is fine-tuning BERT which has been applied in many NLP tasks including text summarization [36, 37, 38].

Additionally, BERT can be extended by training further with domain-specific supplementary. When restricting on the biomedical area, BioBERT and BlueBERT are introduced, for example. We also have experiments utilizing BioBERT as the embedding tool. Unfortunately, we observed subtle changes in the result. Though one guess could be for text summarization task, the influence of domain-specific words is not significant enough, more exploration is needed. In addition to the domain-specific pre-training, we will also explore the possibility of using additional features. The locational feature can be useful, as the sentences at the beginning or the end of a paragraph can contain useful terms for summarization [39]. The length of the sentence and its located section can also be useful. Additionally, the presence of certain cue words can also be the signal of disclosure, as important information is usually followed by 'conclusion' and 'in particular'. Sentences with a high overlay with the title or heading may also be an indicator of importance, as it often related to concepts directly related to the article. These features will be part of our future work, and we will combine these features with text embedding to train a classifier. We will experiment to test if these features can potentially improve the performance of the classifier.

The encoded sentences are then passed to the decoder which aims to extract sentences from the document. One problem that we suffered from is lacking the labeled documents. To handle it, in this paper, we proposed a labeling technique by computing the similarity between the abstract and the document. This labeling strategy is meaningful as the ROUGE test shows it is capable of generating summarizations that have a good representation of the abstract. In this study, we used a proprietary parameter to determine the length of the summary. It reflects the proportion of sentences we highlighted while reading an article. In future work, we will optimize this parameter to generate a summary with a more suitable length. The strategy we applied to determine the similarity between the sentences is using the cosine similarity. We will experiment on using a different approach to determine the similarity between the sentences, including using cosine similarity and ROUGE score.

We applied logistic regression for sentence extraction and compared the recall of ROUGE with the similarity-based and random model. The comparison indicates that even such a simple rather efficient model leads to a valid result. Accordingly, it is reasonable to extend to other models taking advantage of deep learning, listed as follows. Recall that BERT is a pre-trained Transformer encoder. Naturally, the Transformer decoder is treated as our first potential model to decode and classify the sentences. An alternate can be recurrent neural networks or RNN. At every phase, the sentences extracted hitherto, together with encoded representation, produce possibilities for all sentences induced by cross-entropy loss. Then the one with the highest score will be picked as the next prediction. Both models show desirable performance in numerous cases. Nevertheless, challenges may be introduced simultaneously. Indeed, there exists a discrepancy between the training objective and the evaluation, or ROUGE, in the testing stage [40]. In addition, as pointed out in [36], different from the pre-trained BERT, the decoders need to be trained from scratch. As a result, it is possible that BERT overfits the data whereas the decoder underfits, or vice versa. This conflict might be resolved by setting different learning rates in encoder and decoder [36]. To overcome the first mismatch, reinforcement learning, introduced by [40], is the third method we are investigating. This model, rather than utilizing the cross-entropy loss, globally optimizes ROUGE score directly and ranks sentences. Our last attempt inspired by [41], takes both Transformers and reinforcement learning into consideration, anticipating benefits from both models.

In addition to the pipeline, we also consider ideas to improve our experiments. For example, the dataset we collected for this study is composed of 300 Radiation Therapy-related scientific articles from PubMed published in 2020, containing 71943 sentences. One observation is that the AUC of the logistic regression model on the training data is 0.8, whereas the AUC on the testing data is 0.73, revealing an overfitting circumstance. One common solution to get rid of this trouble is to aggregate more data. Henceforth, we are extending our dataset by adding a larger number of publications covering more general topics in the biomedical area.

We also consider presenting the pipeline and its generated summaries to the clinicians and researchers, who may potentially benefited from such a system, to collect feedback on whether the generated summaries have summarized the publications and extracted important information. Such user test allows us to get feedback from the targeted audience, and validate the proposed pipeline for extractive text summarization of biomedical literature.

3.8 Contribution

The major contribution of this part is that we proposed a system that utilize self supervised approach that can automatically generate extractive text summarizations for biomedical literature. The system take a unique path of using abstract to find conclusive sentences, and built a classification model to find such kind of sentences. The result shows the classification model is able to generate exclusive summary that cover the topics in abstracts, and proof the approach is feasible of automatically generate extractive text summery. Such kind of approach can possibly be applied on different domain, and used for transfer learning.

Charpter 3 Reference

[1] Rashmi Mishra et al. "Text summarization in the biomedical domain: a systematic review of recent research". In: *Journal of biomedical informatics* 52 (2014), pp. 457–467.

[2] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

[3] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* (Sept. 2019). Ed. by JonathanEditor Wren. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btz682. URL: http://dx.doi.org/10.1093/bioinformatics/btz682.

[4] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL: https://www.aclweb.org/anthology/W04-1013.

[5] Mehdi Allahyari et al. "Text summarization techniques: a brief survey". In: *arXiv preprint arXiv:1707.02268* (2017).

[6] Milad Moradi and Nasser Ghadiri. "Text Summarization in the Biomedical Domain". In: *arXiv preprint arXiv:1908.02285* (2019).

[7] Günes Erkan and Dragomir R Radev. "Lexrank: Graph-based lexical centrality as salience in text summarization". In: *Journal of artificial intelligence research* 22 (2004), pp. 457–479.

[8] T Elizabeth Workman, Marcelo Fiszman, and John F Hurdle. "Text summarization as a decision support aid". In: *BMC medical informatics and decision making* 12.1 (2012), pp. 1–12.

[9] Marcelo Fiszman et al. "Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation". In: *Journal of biomedical informatics* 42.5 (2009), pp. 801–813.

[10] George Simon et al. "Applying artificial intelligence to address the knowledge gaps in cancer care". In: *The oncologist* 24.6 (2019), p. 772.

[11] Marcelo Fiszman, Thomas C Rindflesch, and Halil Kilicoglu. "Summarizing drug information in Medline citations". In: *AMIA Annual Symposium Proceedings*. Vol. 2006. American Medical Informatics Association. 2006, p. 254.

[12] Han Zhang et al. "Clustering cliques for graph-based summarization of the biomedical research literature". In: *BMC bioinformatics* 14.1 (2013), pp. 1–15.

[13] Nicole Sultanum et al. "Doccurate: A curation-based approach for clinical text visualization". In: *IEEE transactions on visualization and computer graphics* 25.1 (2018), pp. 142–151.

[14] Duy Duc An Bui et al. "Extractive text summarization system to aid data extraction from full text in systematic review development". In: *Journal of biomedical informatics* 64 (2016), pp. 265–272.

[15] Hans Moen et al. "Comparison of automatic summarisation methods for clinical free text notes". In: *Artificial intelligence in medicine* 67 (2016), pp. 25–37.

[16] Ayelet Goldstein et al. “Evaluation of an automated knowledge-based textual summarization system for longitudinal clinical data, in the intensive care domain”. In: *Artificial intelligence in medicine* 82 (2017), pp. 20–33.

[17] Guilherme Del Fiol et al. “Formative evaluation of a patient-specific clinical knowledge summarization tool”. In: *International journal of medical informatics* 86 (2016), pp. 126–134.

[18] Mohammad Amin Morid et al. “Classification of clinically useful sentences in clinical evidence resources”. In: *Journal of biomedical informatics* 60 (2016), pp. 14–22.

[19] Rimma Pivovarov and Noémie Elhadad. “Automated methods for the summarization of electronic health records”. In: *Journal of the American Medical Informatics Association* 22.5 (2015), pp. 938–947.

[20] Ayelet Goldstein and Yuval Shahar. “An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data”. In: *Journal of biomedical informatics* 61 (2016), pp. 159–175.

[21] Daniel J Goff and Thomas W Loehfelm. “Automated radiology report summarization using an open-source natural language processing pipeline”. In: *Journal of digital imaging* 31.2 (2018), pp. 185–192.

[22] Christian Gulden et al. “Extractive summarization of clinical trial descriptions”. In: *International journal of medical informatics* 129 (2019), pp. 114–121.

[23] Elena Baralis et al. “Multi-document summarization exploiting frequent itemsets”. In: *Proceedings of the 27th Annual ACM Symposium on Applied Computing*. 2012, pp. 782–786.

[24] Mozhgan Nasr Azadani, Nasser Ghadiri, and Ensieh Davoodijam. “Graph-based biomedical text summarization: An itemset mining and sentence clustering approach”. In: *Journal of biomedical informatics* 84 (2018), pp. 42–58.

[25] Federico Barrios et al. “Variations of the similarity function of textrank for automated summarization”. In: *arXiv preprint arXiv:1602.03606* (2016).

[26] Josef Steinberger, Karel Jezek, et al. “Using latent semantic analysis in text summarization and summary evaluation”. In: *Proc. ISIM* 4 (2004), pp. 93–100.

[27] Naveen Saini et al. “Extractive single document summarization using binary differential evolution: Optimization of different sentence quality measures”. In: *PloS one* 14.11 (2019), e0223477.

[28] Milad Moradi. “CIBS: A biomedical text summarizer using topic-based sentence clustering”. In: *Journal of biomedical informatics* 88 (2018), pp. 53–61.

[29] Milad Moradi and Nasser Ghadiri. “Different approaches for identifying important concepts in probabilistic biomedical text summarization”. In: *Artificial intelligence in medicine* 84 (2018), pp. 101–116.

[30] Edward Loper and Steven Bird. “NLTK: The Natural Language Toolkit”. In: *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*. 2002.

[31] Yifan Peng, Shankai Yan, and Zhiyong Lu. “Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets”. In: *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*. 2019, pp. 58–65.

[32] Mohammad Hossin and Sulaiman M.N. “A Review on Evaluation Metrics for Data Classification Evaluations”. In: vol. 5. Mar. 2015, pp. 01–11. DOI: 10.5121/ijdkp.2015.5201.

[33] Armand Joulin et al. “Bag of Tricks for Efficient Text Classification”. In: *arXiv preprint arXiv:1607.01759* (2016).

[34] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[35] Ashish Vaswani et al. "Attention is all you need". In: *arXiv preprint arXiv:1706.03762* (2017).

[36] Yang Liu and Mirella Lapata. "Text summarization with pretrained encoders". In: *arXiv preprint arXiv:1908.08345* (2019).

[37] Chi Sun et al. "How to fine-tune BERT for text classification?" In: *China National Conference on Chinese Computational Linguistics*. Springer. 2019, pp. 194–206.

[38] Yunqiu Shao et al. "BERT-PLI: Modeling Paragraph-Level Interactions for Legal Case Retrieval". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. 2020, pp. 3501–3507.

[39] Asad Abdi et al. "An Automated Summarization Assessment Algorithm for Identifying Summarizing Strategies". In: *PLOS ONE* 11 (Jan. 2016), pp. 1–34. DOI: `10.1371/journal.pone.0145809`. URL: `https://doi.org/10.1371/journal.pone.0145809`.

[40] Shashi Narayan, Shay B Cohen, and Mirella Lapata. "Ranking sentences for extractive summarization with reinforcement learning". In: *arXiv preprint arXiv:1802.08636* (2018).

[41] Haoyu Zhang, Jianjun Xu, and Ji Wang. "Pretraining-based natural language generation for text summarization". In: *arXiv preprint arXiv:1902.09243* (2019).

CHAPTER 4: Understanding of Public's Attitude about COVID-19: a social media analysis with BERT

4.1 Introduction

Social media has become an important platform for the public to not only receive public health-related information from health agencies and news outlets but also share opinions and engage in discussions regarding public health-related issues. It has also become an important source for various health agencies and researchers to understand the public's opinion and promote certain health campaigns. It has seen significant use during the global infectious disease pandemic, by both health agencies and individuals. During the pandemic of 2014 Ebola, researchers noticed the significant upward trend of Twitter posts and Google search in the USA[1, 2]. And in the 2016 Zika pandemic, multiple health agencies start to use social media as a communication channel and have adopted effective communication strategies to improve the dissemination of public health-related issues[3]. To this date, COVID-19 has become one of the most discussed topics on social media platforms across the globe.

Any global pandemic will not solely be a health or medical issue. In fact, it is often associated with cultural, social, economic, and political issues[4, 5]. These issues can be seen in the social media discussions, whereas a lot of discussions are around social and political topics, rather than the health or medical topic alone. In the early stage of COVID-19, the majority of the discussion was on quarantine and social distancing. As the pandemic progresses, the discussion has shifted towards mask-wearing, the governmentâs handling of the crisis, and the development of vaccines. To this date, COVID-19 is still one of the most popular topics on social media[6], and a lot of internet users retrieve COVID-19 related information and share their opinions on social media.

Research about monitoring and surveillance on social media discussion about health issues has started decades ago. The idea is that monitoring the trend of certain social media discussion, and the change in trend it captures can be used to predict the outbreak of an epidemic of transmissive diseases, such as influenza[7, 8, 9], Zika virus[10], and the recent COVID-19[11, 12]. These researches show the monitoring and surveillance of social media discussion on health-related topics can improve the prediction of the severity of the epidemic, and can possibly detect the uprise of an epidemic. In addition, non-pharmaceutical interventions (NPI), including social distancing, board restrictions, quarantine, and mask-wearing has been proven to be effective in the reduction of airborne transmission diseases [13, 14, 15]. Researchers have already begun to pay attention to the discussions on certain topics of NPI[16, 17].

The monitoring and surveillance are typically done by the combination of natural language processing (NLP), time series analysis, and geospatial analysis. Various NLP applications including topic modeling, topic classification, sentiment analysis, and semantic analysis, are applied to give a comprehensive understanding of the topic, sentiment, and semantic of each individual social media post. By aggregating the results from NLP in a time scale and geospatial scale, the trend of social media discussion can be formed to understand the publicâs attention and attitude to health-related topics.

The approach of understanding the topic of social media can be seen as two-fold. First, researchers use a combination of certain keywords to determine the topic discussed in social media. Then use the number of social media posts found in a certain time period to determine the outbreak of an epidemic[18, 19, 20]. Such approaches have been proved to show a high correlation in detecting the outbreak of influenza with the data published by the influenza surveillance system[21]. Although such a method is successful in detecting the outbreak of an epidemic, it is not capable of understanding the sentiment and semantic feature in the social media discussion. As mentioned in the previous section, such epidemics are not health issues alone, and are often associated with multiple aspects. Such an approach cannot deliver a comprehensive analysis. The other approach that has been widely adopted is using NLP. The downstream tasks in NLP can often provide a comprehensive insight into the text with the rich linguistic, sentiment, and semantic features. There are word frequency-based approaches such as term frequency-inverse document frequency (TF-IDF) and Latent Dirichlet allocation (LDA)[22], with the combination of certain machine learning models, they can achieve the goal of downstream tasks. Another approach is to the encoding of text with pre-trained embeddings include Word2Vec[23], GloVe[24], and BERT[25], where the models are based on Continuous Bag-of-Words (CBOW) or Transformer. A lot of attention has been made on pre-training the domain-specific embedding, where transfer learning is applied on base models that are often trained for general purposes. The embeddings are then fed into certain machine learning or deep learning methods such as Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) for the downstream tasks. Overall, NLP methods have proven to perform well in monitoring and surveillance of social media.

In this study, we will primarily focus on using word or sentence level embedding to understand the contextual information of the tweets. Word embedding is the process of translating words into numerical vectors, and it has long been a hot research topic in NLP. There are traditional static word embeddings including Word2Vec, FastText[26], or GloVe, where the embedding is trained based on a large cohort of text. A typical example of these embeddings Figure 4.1 is: king - queen = man - women, where the cosine similarity between synonyms or words having the same part of speech is usually very high. However, the drawback of this kind of embedding is that it cannot reflect the true meaning of the word in different contexts, as a word may have different meanings in different contexts. Another potential problem is that these text embeddings are usually trained in a more general corpus, as embedding news to be versatile in different contexts. However, such embedding will often perform not as well in certain specific contexts. In this study, the language used in social media can be very different than the corpus where these text embeddings are trained upon, thus can resulting in low performance in the topic modeling tasks.

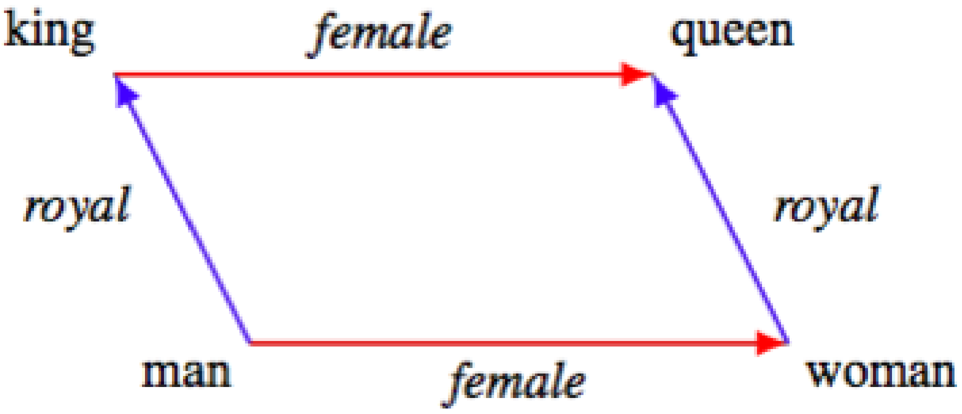


Figure 4.1: Word representations in 2D space

To address the problems, context-dependent pre-trained models have been developed to provide richer and more dynamic information, these models include BERT, ELMO[27], XLNet[28], GPT-2[29], and etc. The major difference between these models and the static word embeddings is that these models can learn from the contexts. BERT is one of the most popular models and has been widely adopted in many tasks. For each token, BERT learns from its context within the input series. Although the name suggests it is bidirectional, the transformer model which BERT is based on will encode the token with its initial embedding and positional information. Then it will pass through 12 multi-head attention layers, where the model will give each token its contextual information within the input series. Such an approach will ensure the context-dependence which is essential to better understand the accurate semantic meaning of the language. In specific, BERT is able to infect a wordâs distinct meanings in different context settings by providing unequal vector representations, whereas static embeddings fail to do so. Another feature of BERT is it allows pre-training the model to the specific domain. It employes the technique of transfer learning, where a model is trained on top of a base model. Such an approach allows the model to have a better representation of the specific domain that the model is fine-tuned upon and usually leads to better performance in downstream tasks. Additionally, it does not require training a model from scratch, which eliminates the issue of heavy computational load and lacking data for training. There have been many examples of pre-training BERT. In the bio-medical area, examples include BioBERT[30], BlueBERT[31], and Med-BERT[32], where these models are trained on biomedical publications or electronic health records. For social media, examples include BERTweet, and there are even models pre-trained on COVID-related tweets like COVID-Twitter-BERT[6]. Such pre-trained models all have substantial performance improvements on certain downstream tasks compare to the original BERT model. In addition to models focuses on the token level, there have been models that focuses on the semantic embedding of sentences. SentenceBERT[33] is an example and has proven to perform well in many sentence-level NLP tasks, especially in comparing the semantic similarity between the sentences.

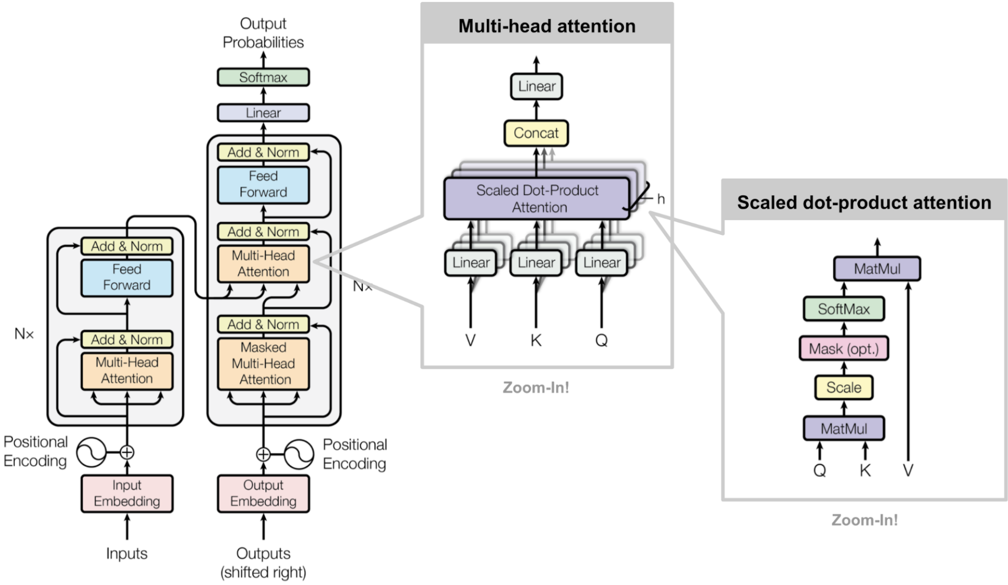


Figure 4.2: The architecture of transformer and multi-head attention mechanism in BERT

4.2 Method

4.2.1 Data

In this study, we focus on using Twitter to understand what is being mostly discussed throughout the pandemic. Twitter is one of the most heavily used and popular social media platforms for people to discuss various topics and express their opinions. In addition, there are plenty of opinion leaders on the platform who have a strong influence on the followers and general twitter use. Twitter is generating a huge volume of data daily, and there has been heated discussion about COVID-19 on the platform. In addition, Twitter has provided a convenient API feature, which allows researchers to extract COVID-related tweets for academic research. In this study, we will use a relatively small dataset for training the topic model and apply the model on a larger set for analyzing the trend of COVID-related topics. Thus, the data collection and annotation were conducted two-fold.

Firstly, we focus on the dataset that enables us to build the topic classification model. We randomly extracted 2000 tweets from a set of tweets we collected during May 2020 and June 2020 using keywords listed in Table 4.1. We further applied a filter to ensure the tweets we collected have a geolocation tag within the United States, and the tweet is written in English. Although for the annotating purpose, it is not necessary to have specific geolocation, we recon matching the data for training the model and the data that the model will apply on, it will give steadier performance and it also gives us the opportunity to validate our proposed codebook for annotating the tweets. In addition, we have excluded tweets that have only 10 or fewer tokens. This ensures the tweets will have real semantic meanings, and such meaning can be extracted by a classifier. To ensure the variety of the dataset, and the dataset that will not be overshadowed by a few super active Twitter users, the selected tweets will all belongs to different users, i.e., no single user will have two tweets in the selected sample. The tweet will be sampled in a time block of every 5 days. For each 5-day time block, 200 tweets will be randomly selected following the filtering criteria.

Table 4.1: Keyword for COVID-19 related tweet extraction

Keyword: COVID-19, COVID19, nCoV-2019, nCoV, SARS, SARS-CoV-2, COVID, coronavirus, corona virus, pandemic, PHEIC, Wuhan virus, China virus, Wuhan pneumonia, Wuhan flu, Kungflu

The set of tweets were sent to a group of annotators who have received the codebook in Table 4.2. The codebook was developed by a domain expert in the area of public health. It has covered all the topics about infectious disease and the related social and political aspects. The codebook covers 6 major topics, which reflects the different perspectives of understanding the topic of tweets, and a single tweet may belong to multiple topics and multiple sub-topics. Proper training was given to the individuals who conduct the annotation of this study. Each tweet will be annotated by two annotators, and if discrepancies were found, the tweet will be sent to the domain expert for verification.

For analyzing the trend of COVID-related tweets, we used a dataset across a larger time scale. We target to collected 12,000 random tweets daily in English from 03/01/2020 to 05/31/2021 that are COVID-19 related using Twitter's Academic API V2. 6000 of the 12,0000 tweets are geo-tagged, with their geolocation in US. There are certain days in the duration that we cannot collect enough tweets with geo-tag, due to the cool down of COVID-19 related discussion. The rest 6000 are tweets without geo-tags. Similar filtering criteria will be applied to these tweets to ensure the quality and rich semantic features. Retweets in this study will be considered as agreeing to the original tweet, thus for any retweets, we will only be focusing on the original tweet. With the application of topic classification models, we can understand the topics of the COVID-19 related tweets. This set of data allows us to monitor the trend of the most discussed topic for each day and observe if certain key events have impacted the topic being discussed on social media. In addition to analyzing

Table 4.2: Topics and Sub-topics of COVID annotating codebook

Topics	Sub-topics
Clinical and Epidemiology	Symptom, Transmission, Testing, Treatment, Prevention,Vaccine, Cases, History, Recovery, Consequence, Risk Factor, Comorbidity, Pharmacy, eHealth, Health System, Health Personnel
Countermeasures	Masks, Other PPE, Disinfection, Food, Exposure, Contact tracing, Technology, Research, Online resource
Policies Politics	Social distancing, Stay-at-home, Shelter-in-place, Constitution, Judicial system, 2020 election, GOP, Democratic Party, Trump, Political figure, Legislation, Economic policy, Curfew, Public sector, Federal
Responses and Impact	Preparedness, Shortage, Financial, Interpersonal, Riot/unrest, Protest, Domestic travel, Intl. travel, college ed, non-college ed, remote working, Business, Sports, Mental health, Suicide, Public response, Unrelated, Main religion, Folk religion, Celebrity, Product promotion, Ecosystem
Spatial scale	Local, State, National, International
Social problem	Disc. Country, Disc. Region, Disc. Ethnicity, Disc. Profession, Disc. Gender, Disc. Age, Disc. Religion, Disc. Food, Violence, Profanity, Misinformation

the topic trend in the time scale, we also seeked the possibility of differentiating the trend on a geospatial scale. We investigated if there are significant differences in topic trends among different places in the United States.

4.2.2 Preprocessing

Prior to training the topic classification model, each tweet went through a series of processes for preprocessing. It is an important process, as irrelevant information inside the text will cause disruption towards the later processes, and texts need to be cleaned and normalized before they can be sent for text encoding. Any usernames or URLs that appeared in the tweet text will be replaced by a common text token. We also replaced all Emoticons with textual representation using the Python emoji library. The tweet text are also normalized to remove certain unusual expressions.

There will be no decapitalization, stemming, or lemmatization, as BERT has already employed these processes in its workflow. The title of the URLs and the hashtags are preserved as additional features to the tweet text. Each tweet will be treated as a series of inputs towards the BERT model. It is similar to the sentence input of BERT, while the difference is that a tweet may have multiple sentences. Since Twitter has a character limit of 240, it is well within the longest sequence input limitation of BERT.

4.2.3 Text Embbeding

Text embedding is an essential part of this aim, as all the contextual, sentiment, and semantic features are reflected in the embedding of the text. Accurate embedding of the text results in a better representation of the text, which will result in better performance in the downstream tasks, specifically in this study, the topic modeling. Considering the computational complexity and the existence of a few already well-performing embeddings in the related domain, we decide not to pre-train the embedding, instead, using and comparing the existing models directly. COVID-Twitter-BERT is pre-trained on COVID-19 related tweets, which is exactly the datasets we are considering using. Additionally, it proved to have decent performance improvement over BERT-Large, which has been considered as a benchmark for most general-purpose NLP models. In this study, we used Cased-BERT-Base as the text embedding and text classification tool, due the the computation complexity. We also compared it with COVID-Twitter-BERT, and noted COVID-Twitter-BERT does have performance improvement over the BERT-Base models. However, due to the computation complexity over 4 million tweets, using a large model like COVID-Twitter-BERT in production will be computational resource heavy and time consuming. Thus, we decide to balance the performance and cost of the model, to use the BERT-Base as out production model.

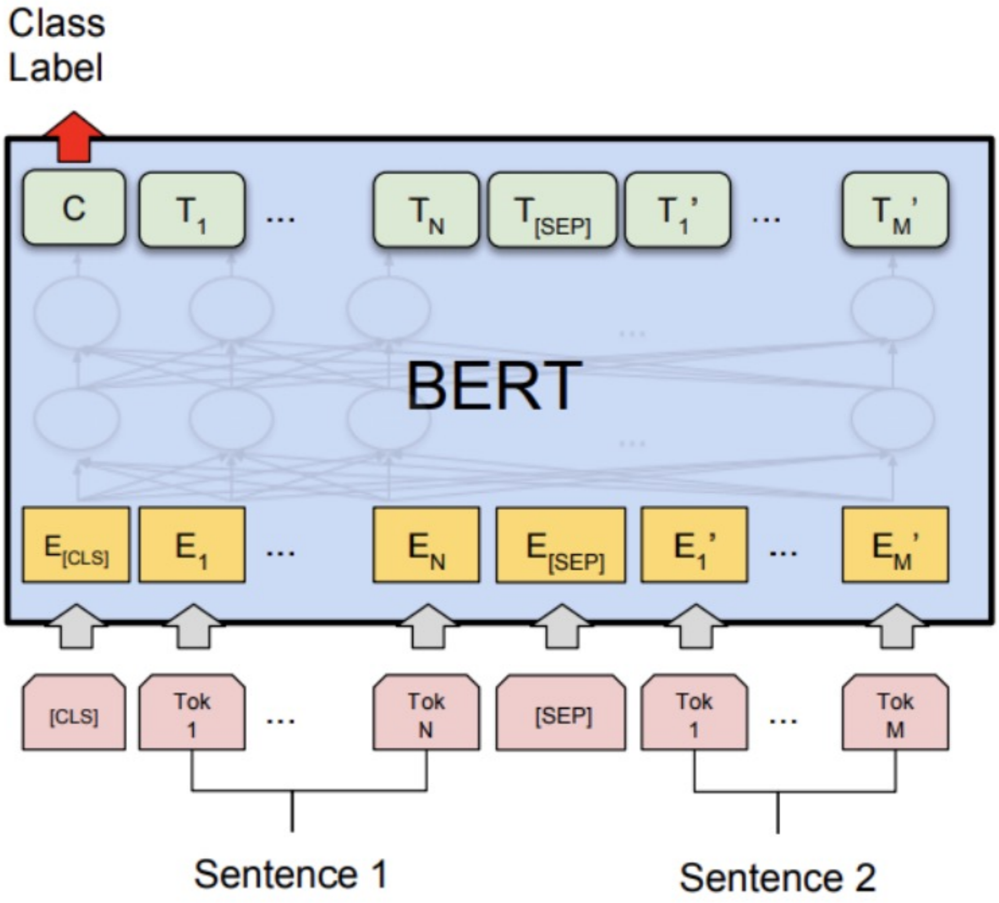


Figure 4.3: [CLS] token of BERT

4.2.4 Topic Classification

Once the tweet has been embedded, we can then use the embedding to build classification models that can properly identify the topic label the tweets belong to. Since each tweet can be assigned with multiple topic labels, we decide to turn this multi-label classification into 6 binary classification problems, that is using one-vs.-rest method for each topic and combine the classification output for each topic model. Thus, 5 binary classifiers will be trained to identify the topic of the tweet. For training these classifiers, we may encounter an imbalanced issue in the training dataset, as we are using one class against the rest of the classes. We will fine-tune the weight of each classifier to ensure the classifier can generate tweet labels that reflect the true proportion of tweets in the dataset.

We compared the performance on topic classification using conventional machine learning model, which is based on the text embedding of the BERT-Base model and Logistic regression, with the fine-tuning the classifier in BERT models. We also compared the result using BERT-Base model and COVID-Twitter-BERT, to see the performance gain using the different models.

4.2.5 Sentiment Analysis

An important part of this study is to analyze the sentiment in the tweets. Sentiment has long been a challenge in NLP, and the models that can perform sentiment analysis are usually trained with a large volume of data with high annotation cost. Since we are lacking the resource to perform sentiment annotation to the COVID-related tweets, thus we decided to adopt the existing sentiment model to perform this task. The nature of the social media posts is the language being very informal, and there is a good amount of sarcasm which will disturb the sentiment models. Although there are challenges like this, we believe that as long as we are applying the model to a large number of tweets, by the law of large number, we will be able to get the true sentiment of the social media discussions, and it can reflect the sentiment in a time and geo-special scale.

The models we use for the sentiment analysis are VADAR (Valence Aware Dictionary and sEntiment Reasoner) and BERT-Base. VADAR is a lexion and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. VADER uses a combination of A sentiment lexicon is a list of lexical features (e.g., words) which are generally labeled according to their semantic orientation as either positive or negative. VADER not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is. Similar to topic classification task, BERT can be used to train a sentiment classifier. In this study, we tained a BERT-Base model in a 3-class fashion. Given a tweet, the model will classify the tweet into negative, neutral or positive.

4.2.6 Evaluation

The evaluation of the classifier will largely depend on the confusion matrix-based method. In this case, we will evaluate the model using the matrics include sensitivity (TPR), specificity (FPR), precision (PPV), accuracy (ACC), F1 score, and AUR-ROC.

The evaluation of the classifier is essential in our study, as we need to ensure the classification of the tweets has relevant good performance so that it can be applied to the larger set of tweets to study the trend and sentiment of the COVID-related tweets. It also allows us to compare the performances of different text embedding and classification models so that we can choose the most appropriate model with high accuracy and reliability for our later tasks.

4.3 Result

4.3.1 Topic Classification

We compared the classification between the different types of models. For the BERT model, we compared the performance using different epochs. As the Figure 4.4 shows, the optimal number of epoch should be 5. With the increase of epochs, the training loss is steadily decreasing. However, the validation loss is decreasing initially, then after 5 epochs, it start to increase, which is an indication of overfit. This happened in all BERT classifiers, and we recon 5 epochs should be optimal for training a BERT classifier.

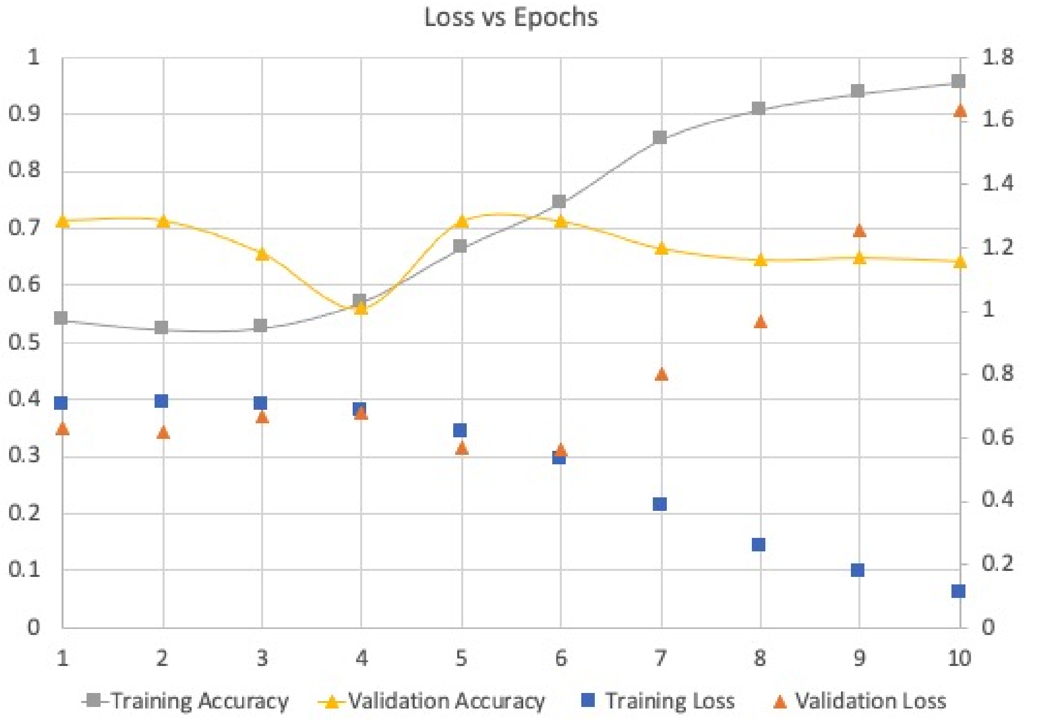


Figure 4.4: Training and Validation Loss VS number of Epochs

The comparison between different model shows that the performance of BERTâs own classifier significantly outperforms the Logistic Regression models. While COVID-Twitter-BERT shows improvement over the BERT-Base, the difference is not as significant. The performance of these three models matches our perception. COVID-

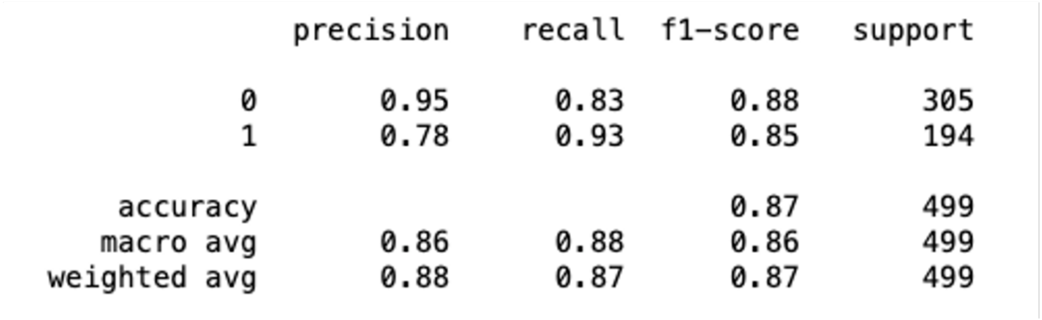
Twitter does show the advantage of large scale neural network, and the training of domain specific data.

Table 4.3: Topic Classification Accuracy comparison

Class	**Logistic Regression**	**BERT-Base**	**CT-BERT**
Clinical/Epidemiological	0.64	0.71	0.77
Countermeasure	0.63	0.80	0.82
Policies	0.67	0.77	0.81
Public Response	0.58	0.67	0.71
Social Issues	0.77	0.88	0.88

For this study, we focused on two of the topics that we are interested, the social issue that are related to COVID-19, and the Nonpharmaceutical Interventions (NPIs). The NPIs topics are combination of certain sub-topics in the class of Countermeasures and Policies, the related topics include Masks, Other PPE, Disinfection, Social distancing, Stay-at-home, and Shelter-in-place. The performance are shown in Table 4.4 and Table 4.5. Overall, the two models shows the accuracy over 87%, and the precision and recall shows the two topic classifier are relatively sensitive for the positive classes. Thus, we recon these two models can be used in the production.

Table 4.4: Social Issue BERT Topic Classifier performance



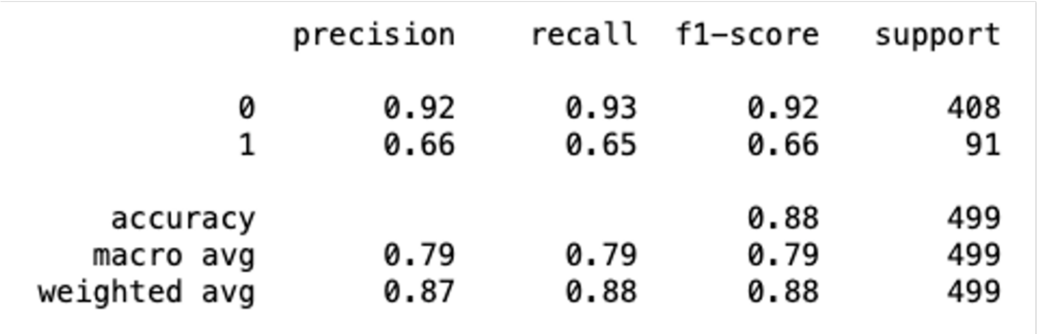
```
              precision    recall  f1-score   support

           0       0.95      0.83      0.88       305
           1       0.78      0.93      0.85       194

    accuracy                           0.87       499
   macro avg       0.86      0.88      0.86       499
weighted avg       0.88      0.87      0.87       499
```

For the two topics, we performed validation and active learning. We randomly selected 100 tweets that are positive for the two topics in the data. A domain expert performed two rounds of annotation, and any difference found in the two round of annotation were further reviewed. The result shows the validation of NPIs has similar

Table 4.5: NPIs BERT Topic Classifier performance



```
              precision    recall  f1-score   support

           0       0.92      0.93      0.92       408
           1       0.66      0.65      0.66        91

    accuracy                           0.88       499
   macro avg       0.79      0.79      0.79       499
weighted avg       0.87      0.88      0.88       499
```

performance in PPV as the test sets, while the PPV of social issue performs slightly worse than the test set. The difference in performance is within tolerance, and is likely caused by randomness in sample selection.

Table 4.6: PPV validation

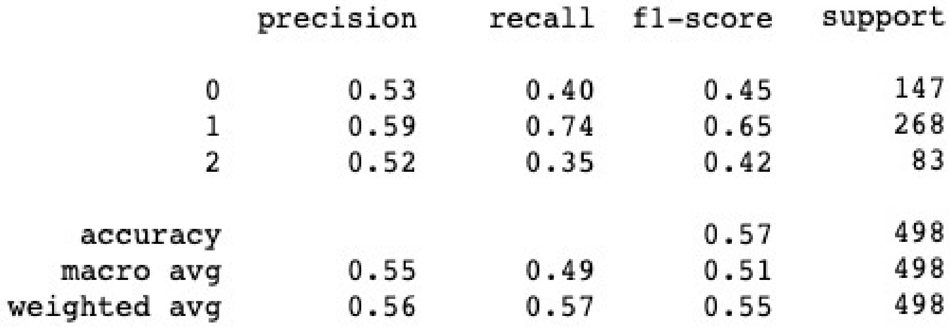
Class	BERT-Base PPV	Validation Set PPV
Social Issue	0.66	0.54
NPIs	0.78	0.77

We then performed active learning based on the small 100 tweets set for the two models. Each model were given labeled tweets for further training. Similar to the training, the active learning set was used to the model with 5 epochs. As a result, we were not able to observe performance gain with the active learning set, indicating the features of the training set has all been extracted, with active learning set not contributing to the model.

4.3.2 Sentiment Classification

Sentiment is also a important aspect of understanding the public's opinion towards certain topics. Similar to topic classification, we trained a sentiment classifier using BERT-Base and its 3-class classifier. We use similar hyper parameters to tune the BERT-Base classifier, with the only difference being that we used 8 epochs instead of 5 for the topic classification. This is due to sentiment is more difficult to model, and it takes more training case to update the model parameter to the optimal. In addition, it is more difficult to model the sentiment for tweets, as they can often be sarcastic or being informal. We compared the result of sentiment generated by Vader and sentiment generated using BERT-Base model. Note the label 0, 1, 2 means negative, neutral, positive.

Table 4.7: Sentiment Classification using VADER

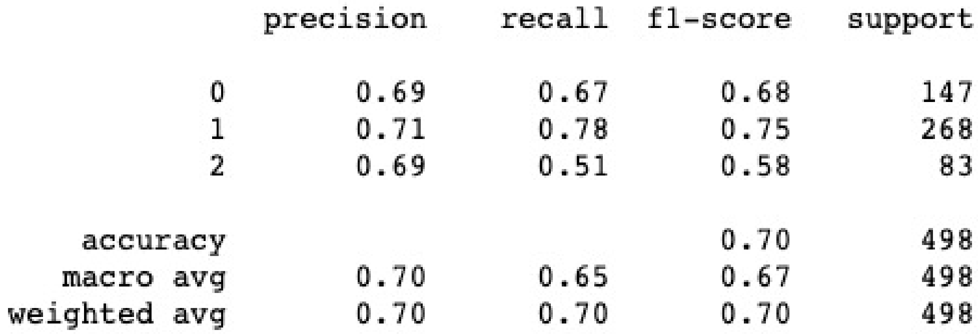


```
              precision    recall  f1-score   support

           0       0.53      0.40      0.45       147
           1       0.59      0.74      0.65       268
           2       0.52      0.35      0.42        83

    accuracy                           0.57       498
   macro avg       0.55      0.49      0.51       498
weighted avg       0.56      0.57      0.55       498
```

Table 4.8: Sentiment Classification using BERT-Base



```
              precision    recall  f1-score   support

           0       0.69      0.67      0.68       147
           1       0.71      0.78      0.75       268
           2       0.69      0.51      0.58        83

    accuracy                           0.70       498
   macro avg       0.70      0.65      0.67       498
weighted avg       0.70      0.70      0.70       498
```

From Table 4.7 and Table 4.8, we observed that BERT-Base model significantly outperformed VADER in all class labels. VADER model has an overall accuracy of 0.57, which is relatively low compared to other studies. We believe this is due to the annotation of the sentiment is done with crowd sourcing, making it relatively inconsistent. Overall, BERT-Base achieved an accuracy of 0.7 for a three-class sentiment classifier, outperformed previous state-of-the-art VADER, showing the capability of deep learning and transformer's context aware nature.

4.3.3 Trend Analysis

We applied the topic classification models and sentiment classification model on the 4 million tweets to understand the change in topic in the time and geospacial scale. We also compared the Twitter data we extracted, the set of tweets with Geo-tags and the set without Geo-tags, to see if they have similarity in trend. Note the trend analysis presented in this chapter are all being smoothed using Gaussian Smoothing with a 7 day window and standard deviation of 3.

4.3.3.1 Comparison between Geo-tagged and Non Geo-tagged tweets

We have gathered 6000 tweets per day for both geo-tagged tweets and Non geo-tagged tweets. We want to see if they are different in terms of trend and topic distribution.

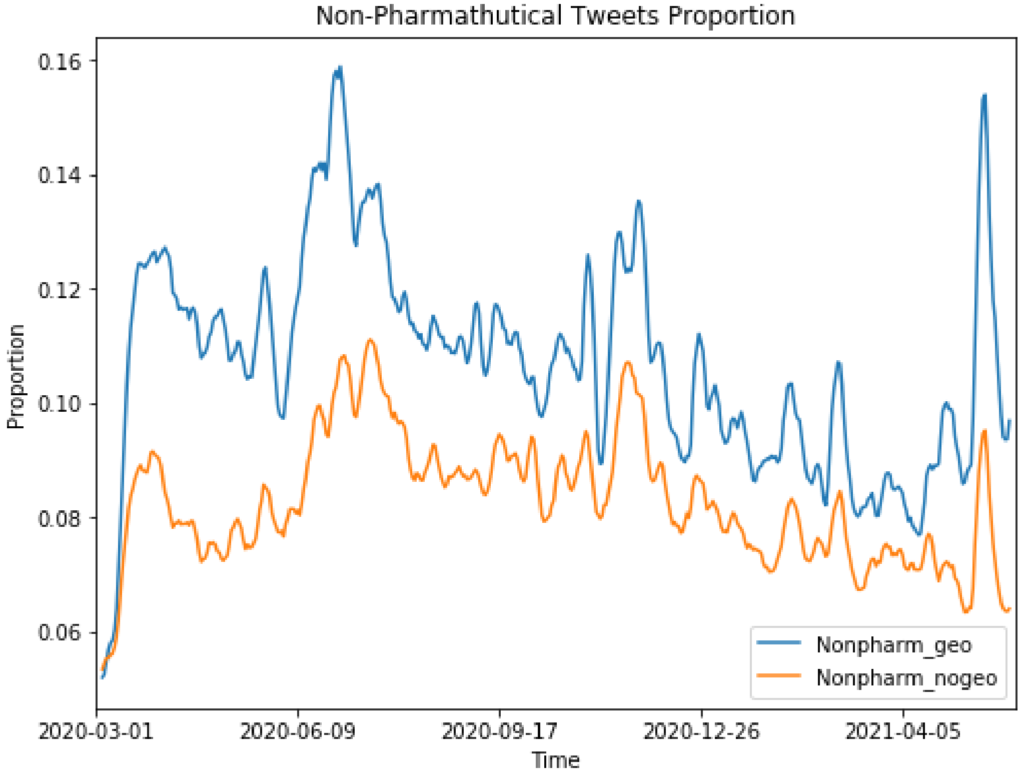


Figure 4.5: NPIs Topic Proportion Geo-Tagged vs Non Geo-Tagged

From Figure 4.5 and Figure 4.6, we found that generally the proportion of the topic

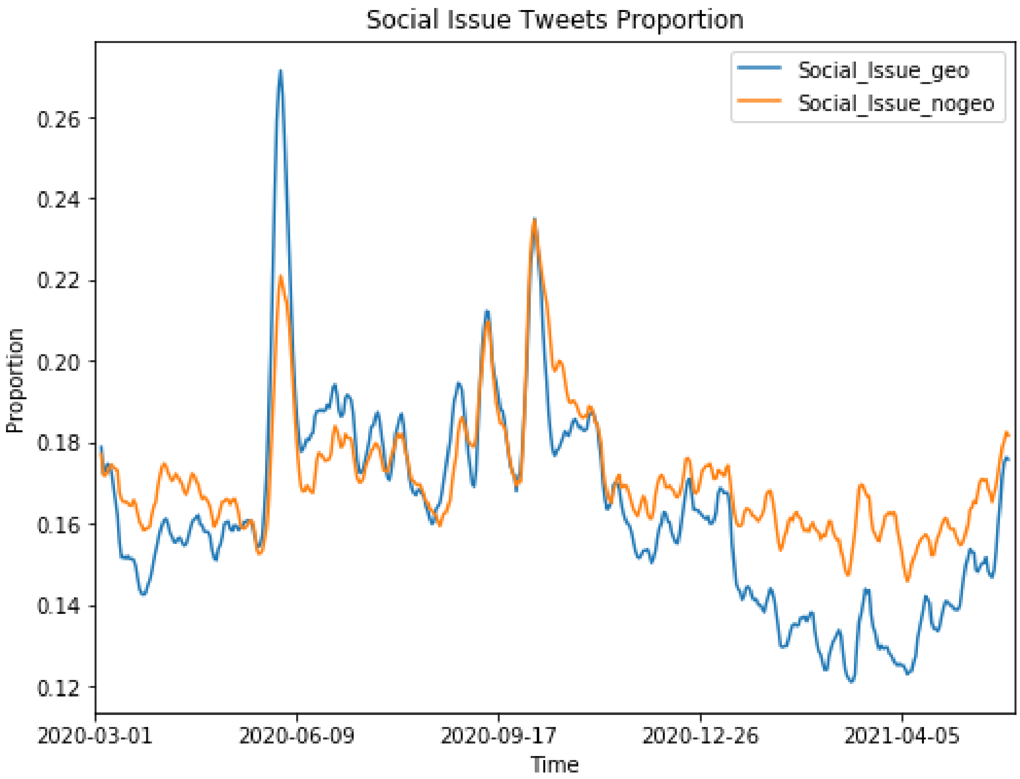


Figure 4.6: Social Issue Topic Proportion Geo-Tagged vs Non Geo-Tagged

has very high correlation. The Pearson's Correlation test result in a Person's R score of 0.79 and 0.8 for NPIs and Social Issues, showing the topic being discussed among geo-tagged tweets are highly correlated with non-geo-tagged tweets. In contrast, the Pearson's R score for comparing the geo-tagged tweets in NPI and geo-tagged tweets in social issue is 0.24, showing it is weakly correlated. We also find the geo-tagged tweets has significantly higher proportion in NPIs, showing people who shares geo-tags are more enthusiastic in discussion of NPIs related issues. On the other hand, for social-issue related topics, people without geo-tags show more interests in such topics.

We also compared the sentiment between tweets with geo-tags and tweets with no geo-tags. The sentiment is based on overall mean sentiment per day for the two datasets. The sentiment range from -1 to 1, while 0 indicate a neutral sentiment. The sentiment for all tweets combine as shown in Figure 4.7.

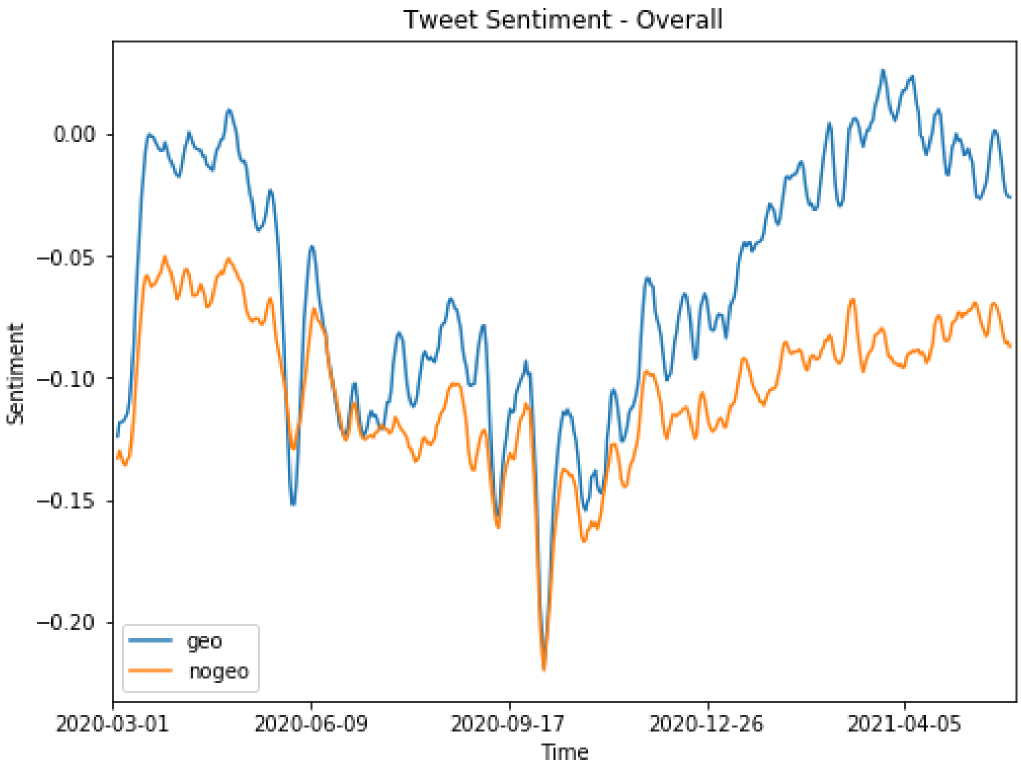


Figure 4.7: Overall daily sentiment

The Figure 4.7 shows there are significant difference between tweets with geo-tags and without geo-tags. Pearson's Correlation test shows Pearson's R score of 0.84, showing the sentiment for the two sets are highly correlated. Overall, tweets with geo-tags has significantly higher sentiment score than tweets without geo-tags.

We compared the sentiment in the topics of NPIs and social issues. We compared tweets with geo-tags and without geo-tags, and we also associated the trend of sentiment with certain key events. Figure 4.7 shows the sentiment on the topic of NPIs, tweets with geo-tags has higher sentiment compared with tweets without geo-tags. There are several sudden dive of the sentiment, and based on the time frame and sample tweets during the time frame, we found the dip in the sentiment were caused by President Trump test positive for COVID-19 and CDC update the guideline for masks. The two topics are highly associated with NPIs, showing the sentiment classification model is able to capture the change in the sentiment.

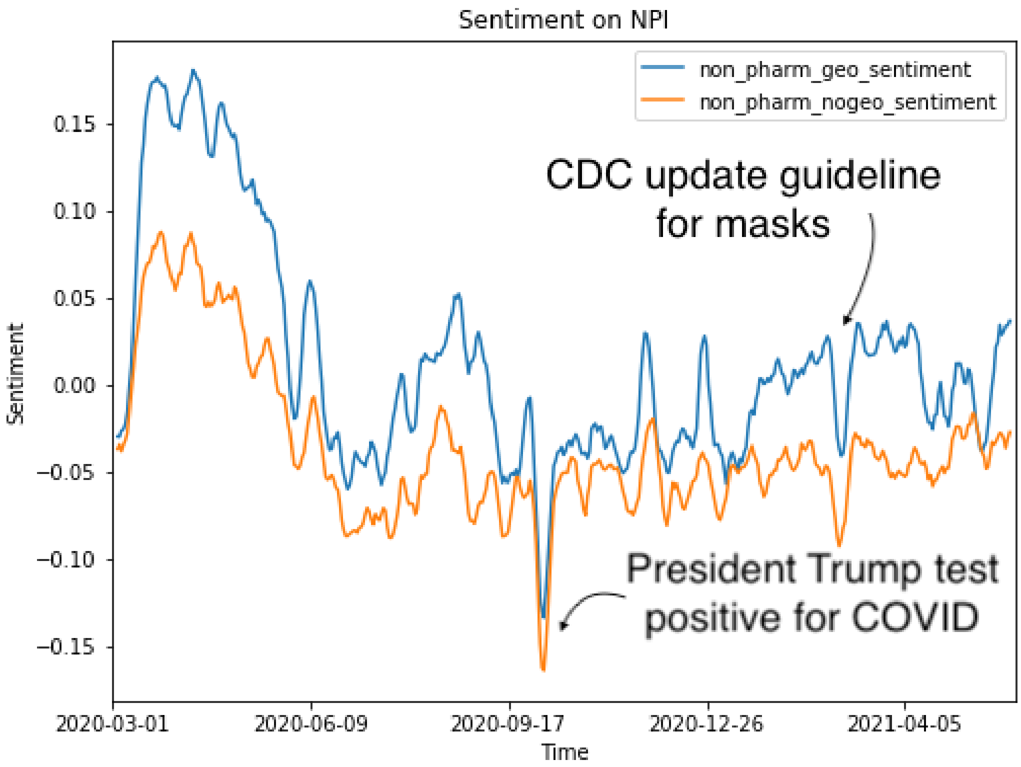


Figure 4.8: Sentiment on NPIs

Figure 4.8 shows the sentiment on social issues, sentiment shows tweets with geo-tags has lower sentiment than tweets without geo-tags. Compared with NPIs, the overall sentiment on social issues is -0.42, while the overall sentiment on NPIs is 0. The sentiment between the two topics are significantly different. And it fits our perception, as when someone talks about social issues, it is most likely to be negative. Similar to sentiment on NPIs, we are able to find the key events that caused the sudden dive of the sentiment. We found huge dips caused by Murder of George Floyd, President Trump admitted downplay of COVID threat, President Trump test positive for COVID, and US election. We also observed that some of these events are not reflected in the sentiment on NPIs, show the topic classification model is capable of separating the non-relevant tweets.

The trend analysis in comparing the geo-tagged tweets and non geo-tagged tweets shows although the two sets are fundamentally collected form different criteria, they

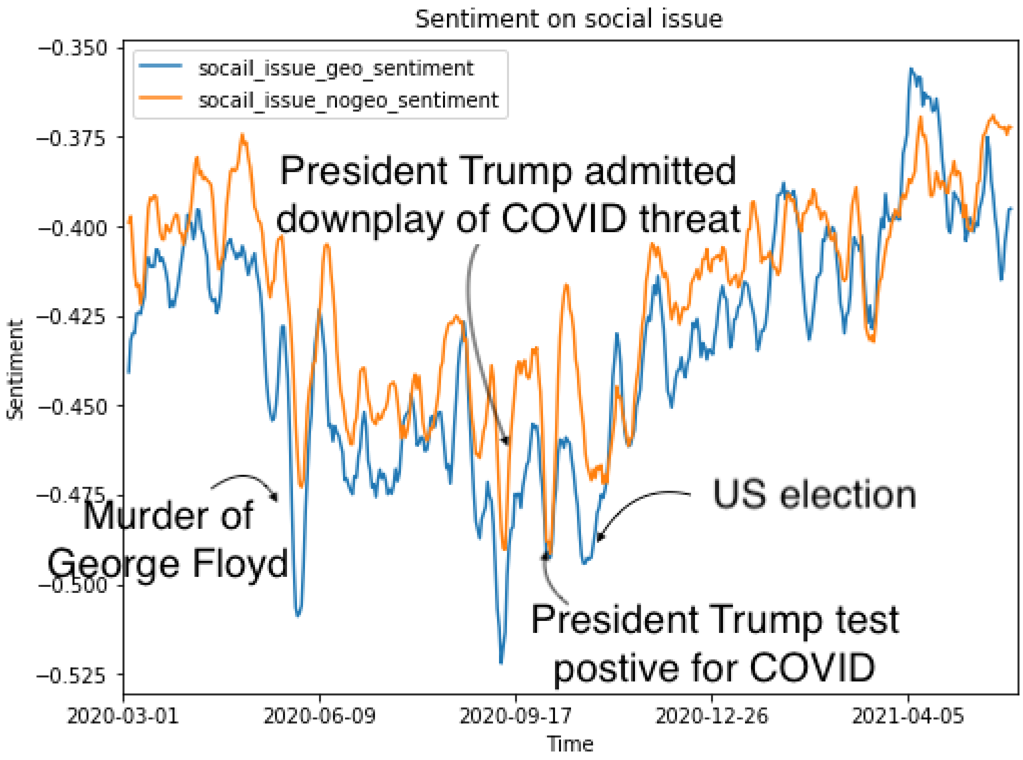


Figure 4.9: Sentiment on Social Issues

are highly correlated in trend, and shows difference in the proportion of topics and sentiment scores. We also observed certain key events that will drive the sentiment down during certain time period, showing the sentiment classification model is sensitive enough to capture the trend. Similarly, the topic classification model can classify the tweets into different topics correctly, resulting key events to be isolated and only observed in certain classes.

4.3.3.2 Comparison between Top 50 cities and the rest

In this section, we mainly focus on the comparison in top trend and sentiment trend for top 50 cities with most tweets and the rest of the tweet set. This allows us to understand if there is difference between people who lives in big cities urban areas and people who lives suburban and rural areas. The comparison was done using geotags that associate with the tweets. In total, 13,299 cities were found in the 2 million tweets, with top 50 cities contributing 37% of the total tweets. The Figure 4.10 shows the cumulative distribution when considering the top 200 cities with tweets, with top 200 cities contributing 50% of the tweets and top 50 contributing 37% of the tweets. We decided to use the top 50 cities as it is more representative of the people who lives in urban areas. The top 50 cities with most tweets are listed in Table 4.9.

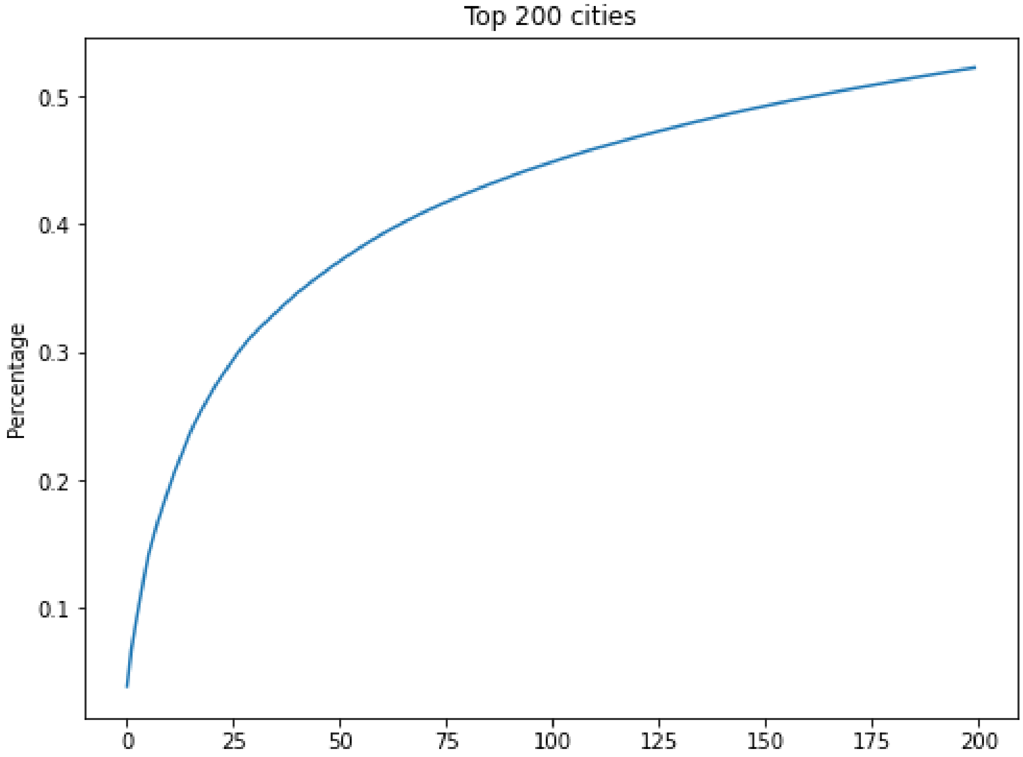


Figure 4.10: Tweet count proportion top 200 cities - cumulative

Table 4.9: Top 50 cities with most tweets

Rank	City	Num of Tweets	Rank	City	Num of Tweets	Rank	City	Num of Tweets
1	Los Angeles, CA	68337	18	Portland, OR	11166	35	Kansas City, MO	5567
2	Manhattan, NY	51047	19	Nashville, TN	10919	36	Tampa, FL	5536
3	Chicago, IL	35804	20	Denver, CO	10406	37	Raleigh, NC	5530
4	Brooklyn, NY	32850	21	Charlotte, NC	10388	38	Pittsburgh, PA	5464
5	Houston, TX	31695	22	Columbus, OH	9977	39	Oklahoma City, OK	5406
6	Washington, DC	30025	23	Minneapolis, MN	8573	40	Memphis, TN	5199
7	Philadelphia, PA	21751	24	Bronx, NY	8546	41	Long Beach, CA	5198
8	San Francisco, CA	20660	25	Baltimore, MD	8541	42	St Louis, MO	4690
9	Queens, NY	18181	26	Indianapolis, IN	8481	43	Tucson, AZ	4681
10	Austin, TX	17881	27	Las Vegas, NV	8153	44	Detroit, MI	4617
11	Seattle, WA	17203	28	New Orleans, LA	8015	45	Milwaukee, WI	4492
12	San Diego, CA	17103	29	Oakland, CA	6938	46	Arlington, VA	4418
13	Dallas, TX	14937	30	Miami, FL	6841	47	Cleveland, OH	4391
14	Atlanta, GA	14805	31	San Jose, CA	6367	48	Albuquerque, NM	4340
15	Phoenix, AZ	14747	32	Sacramento, CA	5935	49	Omaha, NE	4320
16	San Antonio, TX	14478	33	Fort Worth, TX	5639	50	Honolulu, HI	4147
17	Boston, MA	12289	34	Orlando, FL	5622			

First, we compared the topic proportion on NPIs and social issues in the top 50 cities and the rest. As Figure4.11 shows, the proportion of NPIs related tweets are generally around 11% of the overall tweets. We also noticed that in the beginning of the pandemic, during April 2020 and August 2020, people who lives in the top 50 cities are more likely to have discussion on NPIs than the rest of the people. We also observed that after September of 2020, the discussion on NPIs are relatively similar between people who lives in top 50 citis and the rest. This matches the trend of disease that major cities are impacted the most at the beginning of the pandemic, thus people who lives in those cities are more likely to discuss about NPIs related issues.

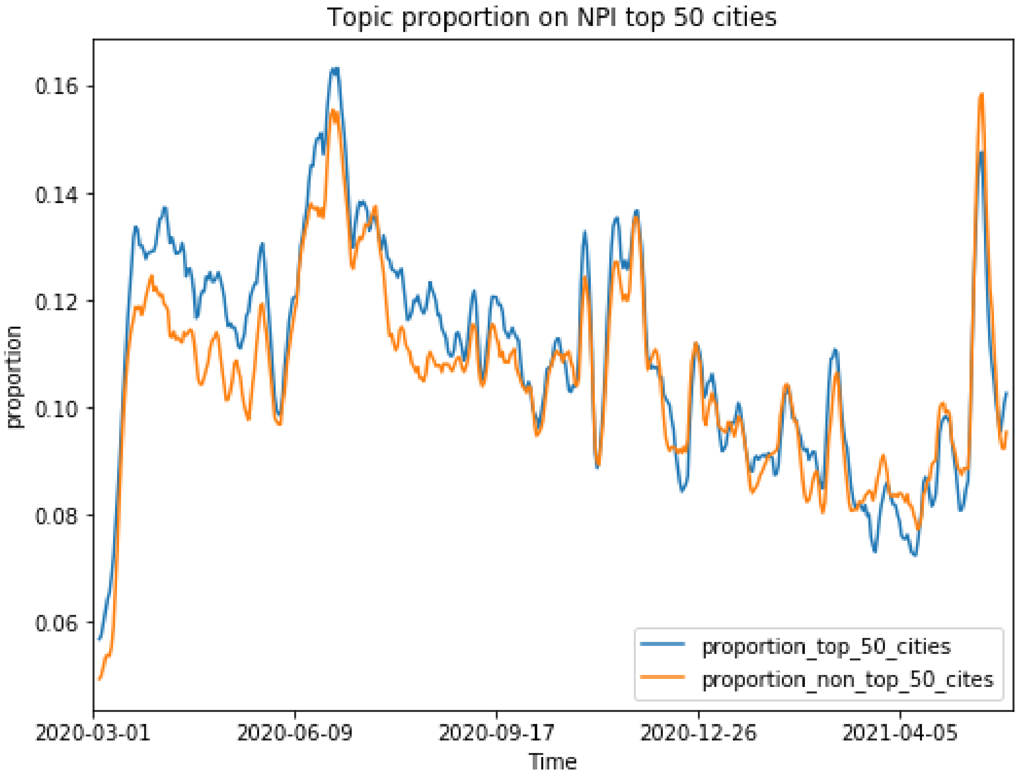


Figure 4.11: NPIs Topic Proportion Top 50 cities vs the rest

For the social issue, as Figure4.12 shows, the proportion of social issue related tweets are generally around 16% of the overall tweets, while it can abruptly rise when certain social issue related events happen. In contrast to the NPIs, we found tweets that are generated in top 50 cities has lower proportion on social issue compared with the rest. But it does have higher peak on late May 2020 with the event of the Murder of George Floyd.

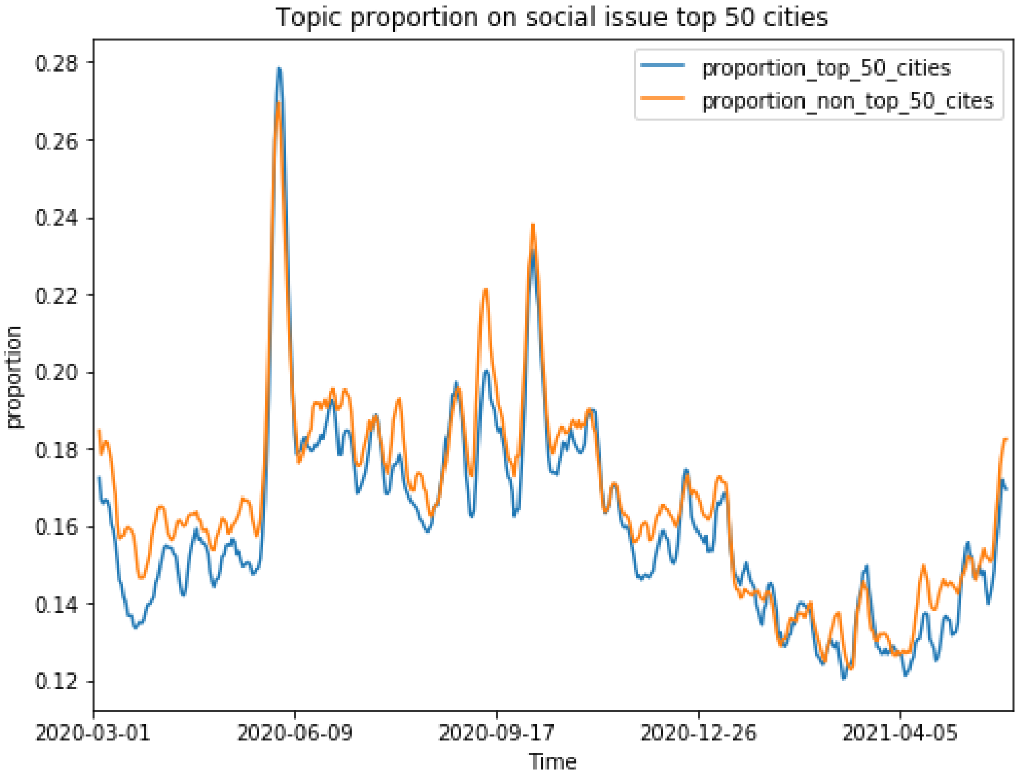


Figure 4.12: Social Issue Topic Proportion Top 50 cities vs the rest

We then focused on comparing the sentiment between the tweets generated from the top 50 cites and tweets generated from the rest of the cities. We compared the overall sentiment of tweets, that are COVID-19 related. As Figure4.13 shows, there are a consistent difference throughout in the sentiment, as the tweets generated from top 50 cities has a sentiment above the tweets generated from the rest of cities. The difference is highly correlated, with a Pearson R score of 0.92, showing the sentiment are highly correlated. However, there are a consistent 0.03 difference in the sentiment, as tweets from top 50 cities are generally more positive.

We also compared the sentiment on the tweets in NPIs and social issues. As Figure4.14 and Figure4.15 show, the sentiment of tweets from top 50 cities are higher

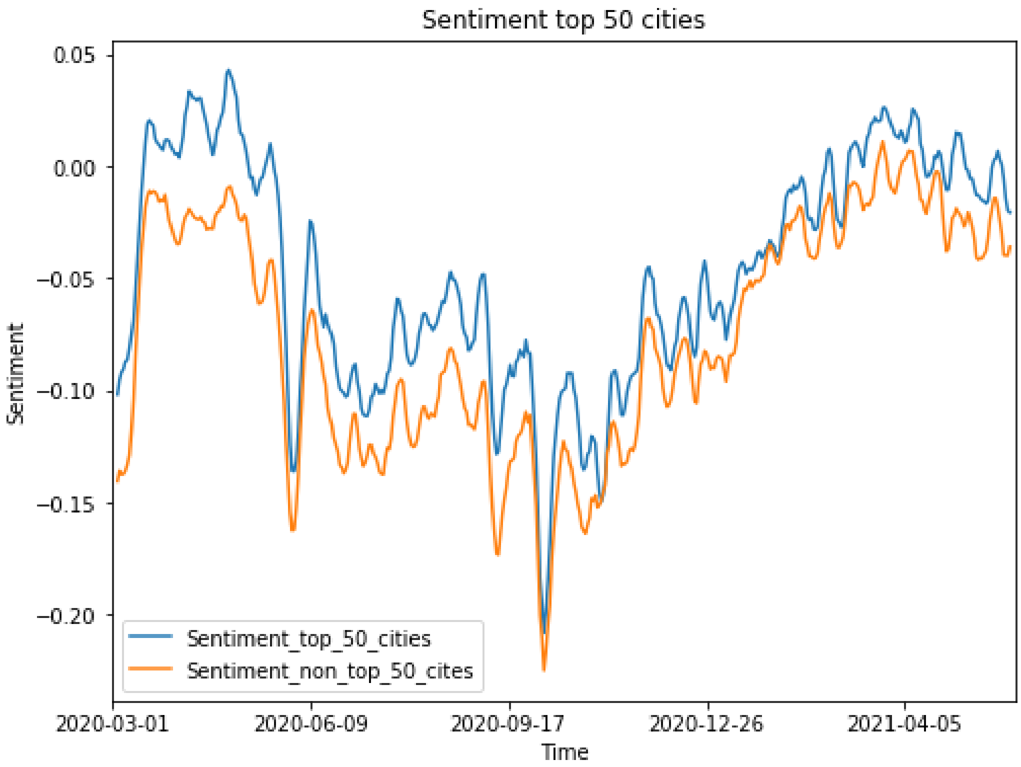


Figure 4.13: Sentiments Top 50 cities vs the rest

than the tweets form the rest cities. We also noticed the sentiment on NPIs was relatively high during March 2020 and June 2020, then the sentiment has went down and stay around 0. We believe this is due to the unclear message that CDC send in the beginning of the pandemic. Then with the requirement of mask wearing, the public begin to have negative comment towards NPIs. It is also observed that in June 2020, the proportion of tweets have topic of NPIs begin to pick up. For the social issues, we observed a trend that sentiment of tweets from top 50 cities are consistently higher than the rest of cities. However, based on Pearson's correlation test, the Pearson's R score is only 0.51 for social issues, which is the lowest in all the comparisons we observed. On the other hand, the Pearson's R score is 0.72 on the sentiment comparison for NPIs. This indicate the correlation trend itself many not be as obvious in the sentiment of specific topics.

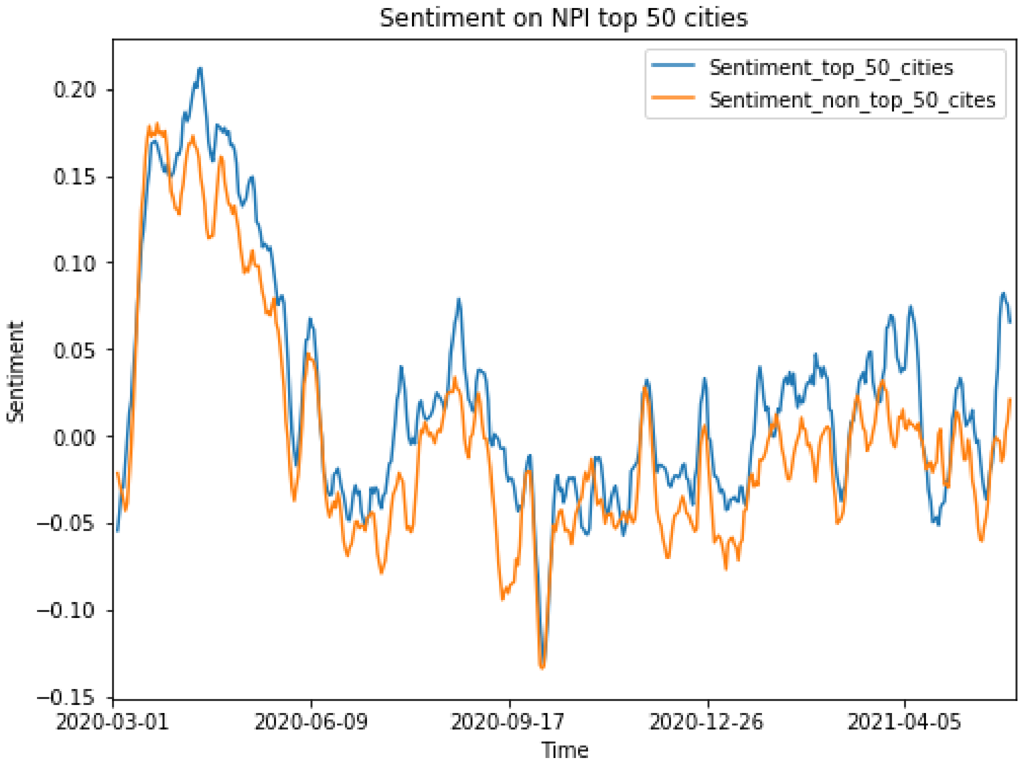


Figure 4.14: Sentiments on NPIs Top 50 cities vs the rest

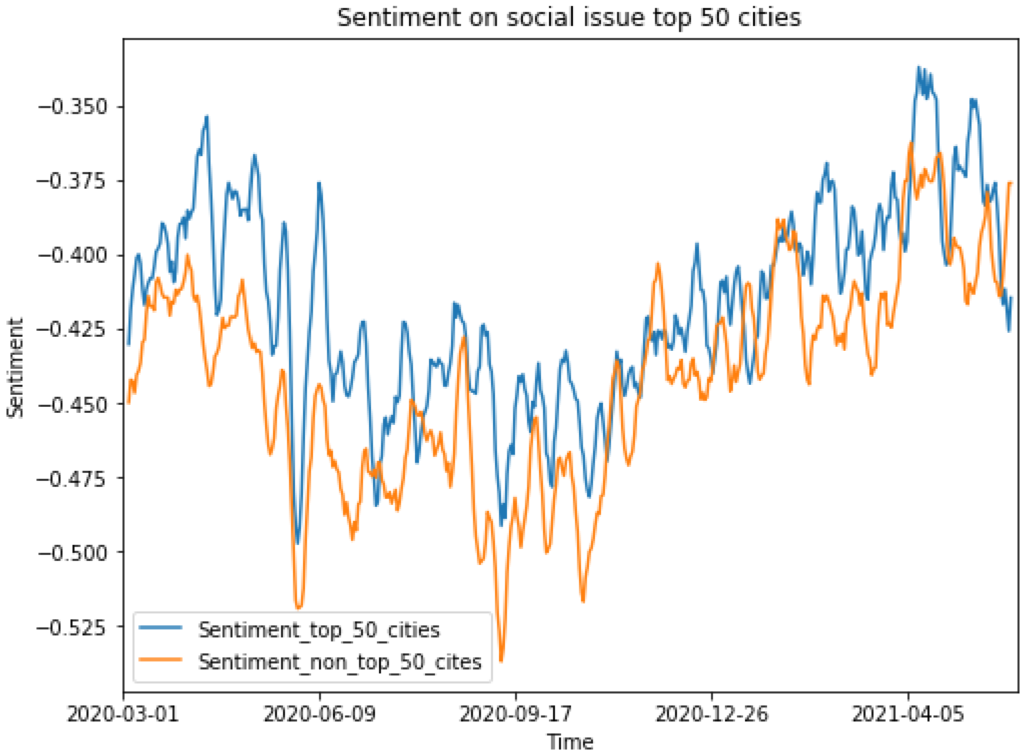


Figure 4.15: Sentiments on Socail Issues Top 50 cities vs the rest

4.3.3.3 Conclusion of Trend analysis

Overall, using the topic classification models and sentiment classification models, we are able to observe the similarity and difference in the trend. We believe the models are sensitive enough to capture certain key events during the pandemic, and also is capable of showing difference in topic proportions and sentiment on specific topics. We observed the trend in topic proportion and sentiment are highly correlated between geo-tagged tweets and non geo-tagged tweets. And we also observed the difference in topic proportion and sentiment between tweets from top 50 cities and tweets from the rest of cities. We believed the approach and the model can be used to monitor the change in the trend on social media, to better understand what is being discussed and the sentiment of the discussion.

4.3.4 Geospatial analysis

In addition to the trend analysis we performed in the previous section, we also performed an analysis based on Geo-tagged data. Specifically, we want to compare the state level difference in the proportion of topics being discussed and the sentiment. It allows us to understand the state level attitude towards certain topics, and if certain state are behaving different than other states.

4.3.4.1 Number of tweets

We first start by comparing people in which state is more likely to post tweets. We calculated the number of tweets generated from each state, and then normalized using the population estimate for each state. Then we are able to graph the number of tweets regarding COVID send from the different state per year per 1,000 capita.

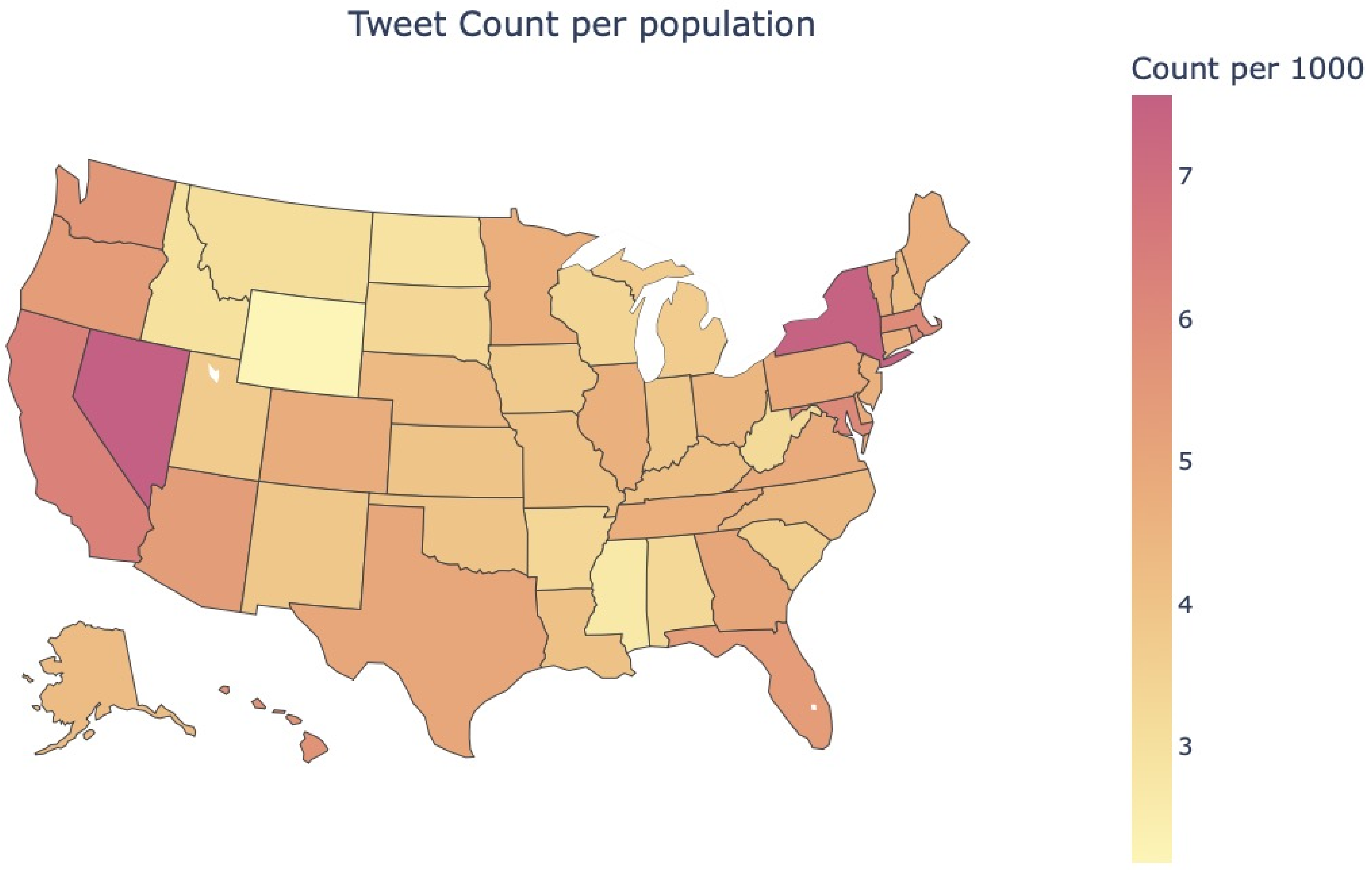


Figure 4.16: Tweet count per population by state

As Figure 4.16 shows, the states on the east coast and west coast, along with some state in the south, are more willing to post tweets. Particularly, the District of Columbia has way higher tweet rate than rest of the state. Based on our dataset, we are recording 34.1 tweets per 1000 population annually for the District of Columbia, compared with national average of 5.1. For the rest of states, we are observing over 3 times the difference of number of tweets per 100 population relatively, showing the state level difference. We found Nevada and New York has the highest tweeting rate per capita, followed by California and Maryland. On the other hand, Wyoming, Mississippi, and North Dakota has the lowest tweeting rate per capita.

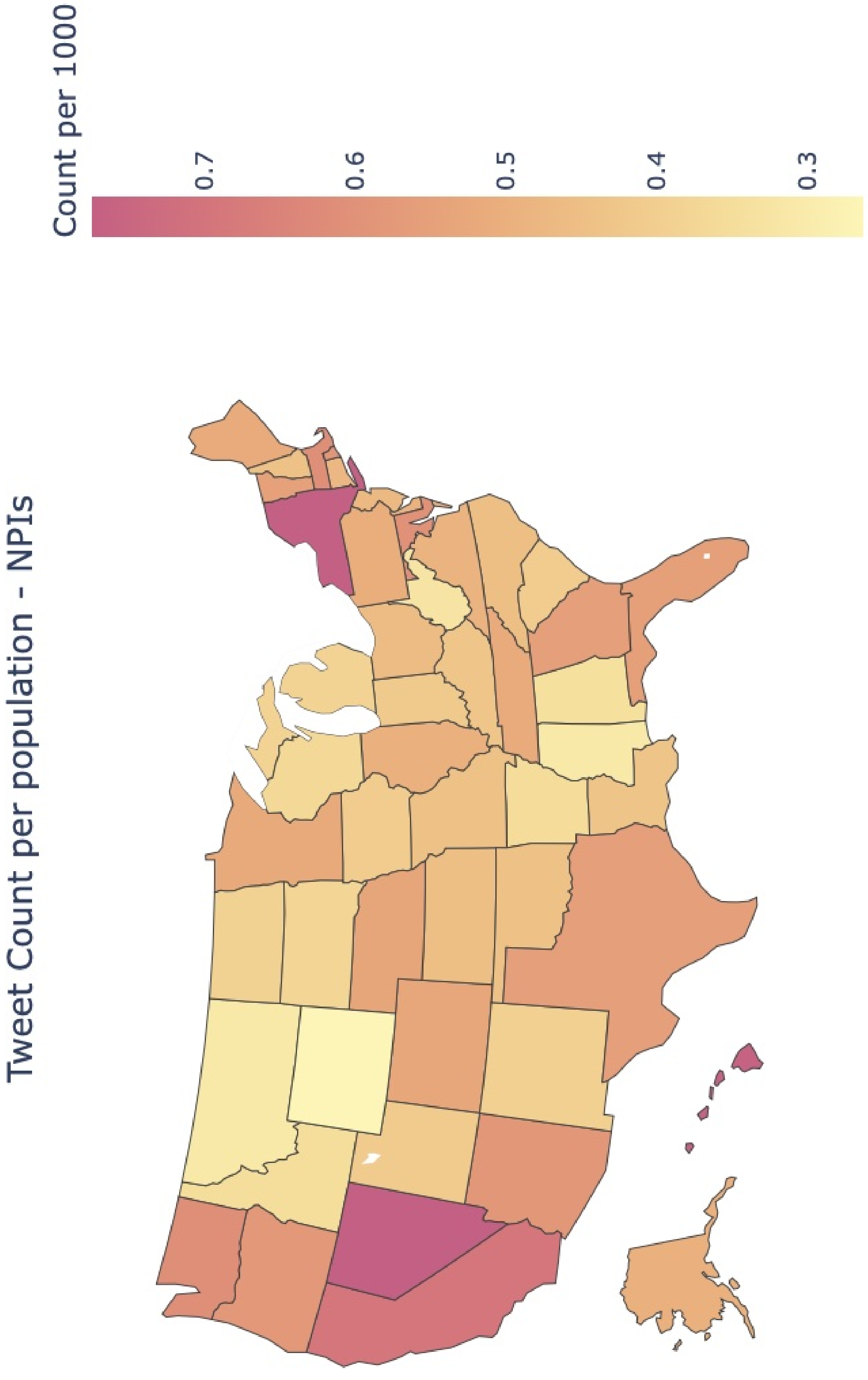


Figure 4.17: Tweet count per population by state - NPIs

As Figure 4.17 shows, there are marginal change in the number of tweets regarding the NPIs compared with number of COVID-related tweets overall. However, we noticed that the state of Hawaii has generated similar number of NPIs related tweets per capita, which is a large proportion consider the COVID-related tweets are on par with national average. In general, it seems people in Hawaii has more interest to post tweets regarding NPIs related issues.

For the topic of Social Issues, we observed modest change from the COVID-related tweets as a whole. It appears the central of US and north east of US has slightly lower willingness to post social issues related tweets, while the west and the south remains mostly same. We are not able to find any single state that has a signicant change.

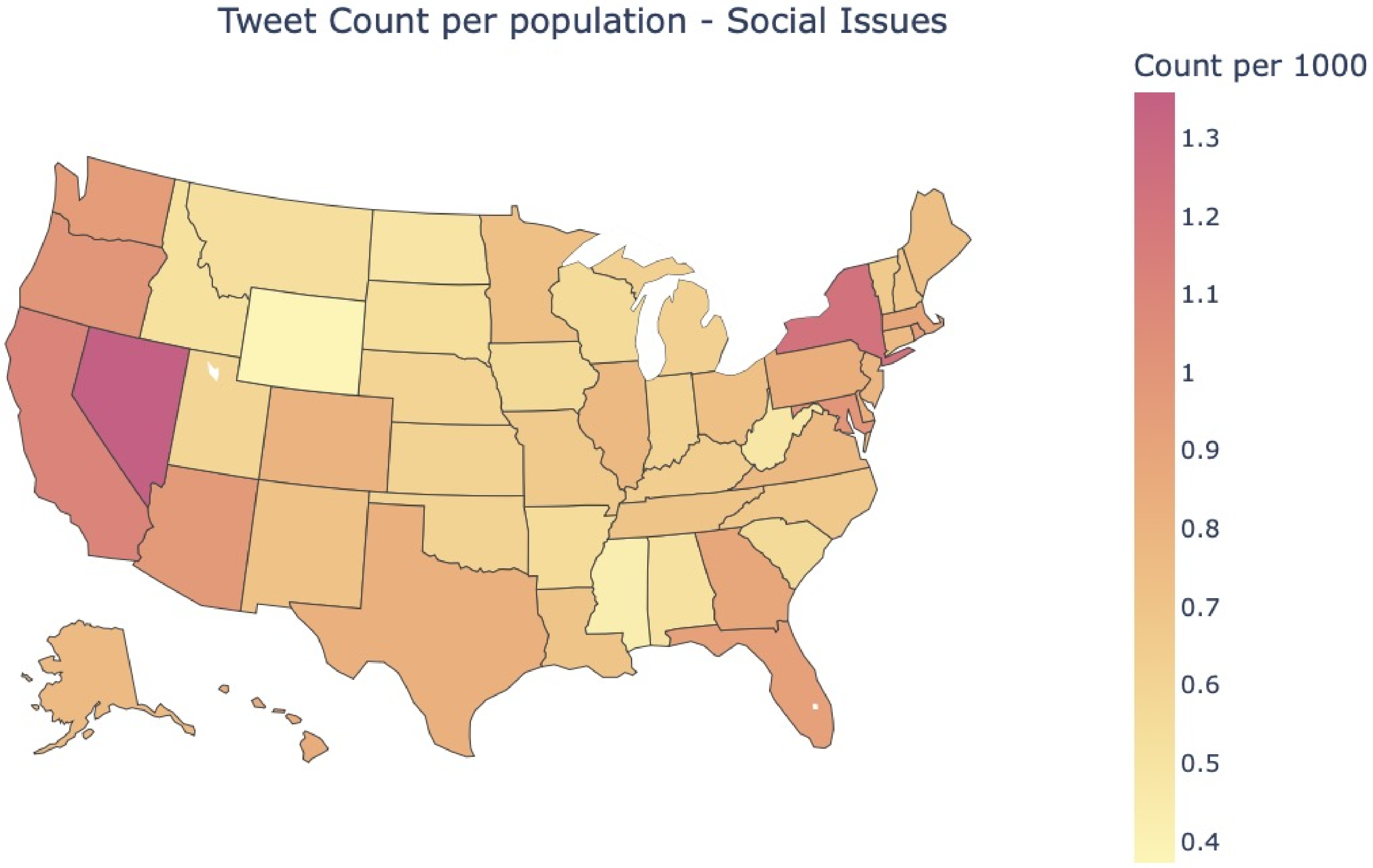


Figure 4.18: Tweet count per population by state - Social Issue

Overall, the state level analysis shows there are difference in tweeting behavior. The District of Columbia has significantly outpace the number of tweets per capita. We have also observe that the state of Nevada, New York, California, Maryland, and Massachusetts has higher tweets per capita compared to other states, We have also observed that the state of Hawaii has generated a good amount of tweets on NPIs,

4.3.4.2 Topic proportion of tweets

Followed by the number of tweets, we are also interested in the topic proportion of tweets in state level. Similar to the previous sections, we focused on the two topics, NPIs and social issues. We are interested to see if there are difference of the topics being discussed from state level.

As Figure 4.19 shows, about 10.7% of the tweets are about NPIs as the national average. We found some states include Hawaii, North Dakota, and Nebraska has tweet proportion on NPIs higher than the rest of states. In particular, Hawaii has 13.4% of the tweet on the topic of NPIs, which is significantly different then other states.

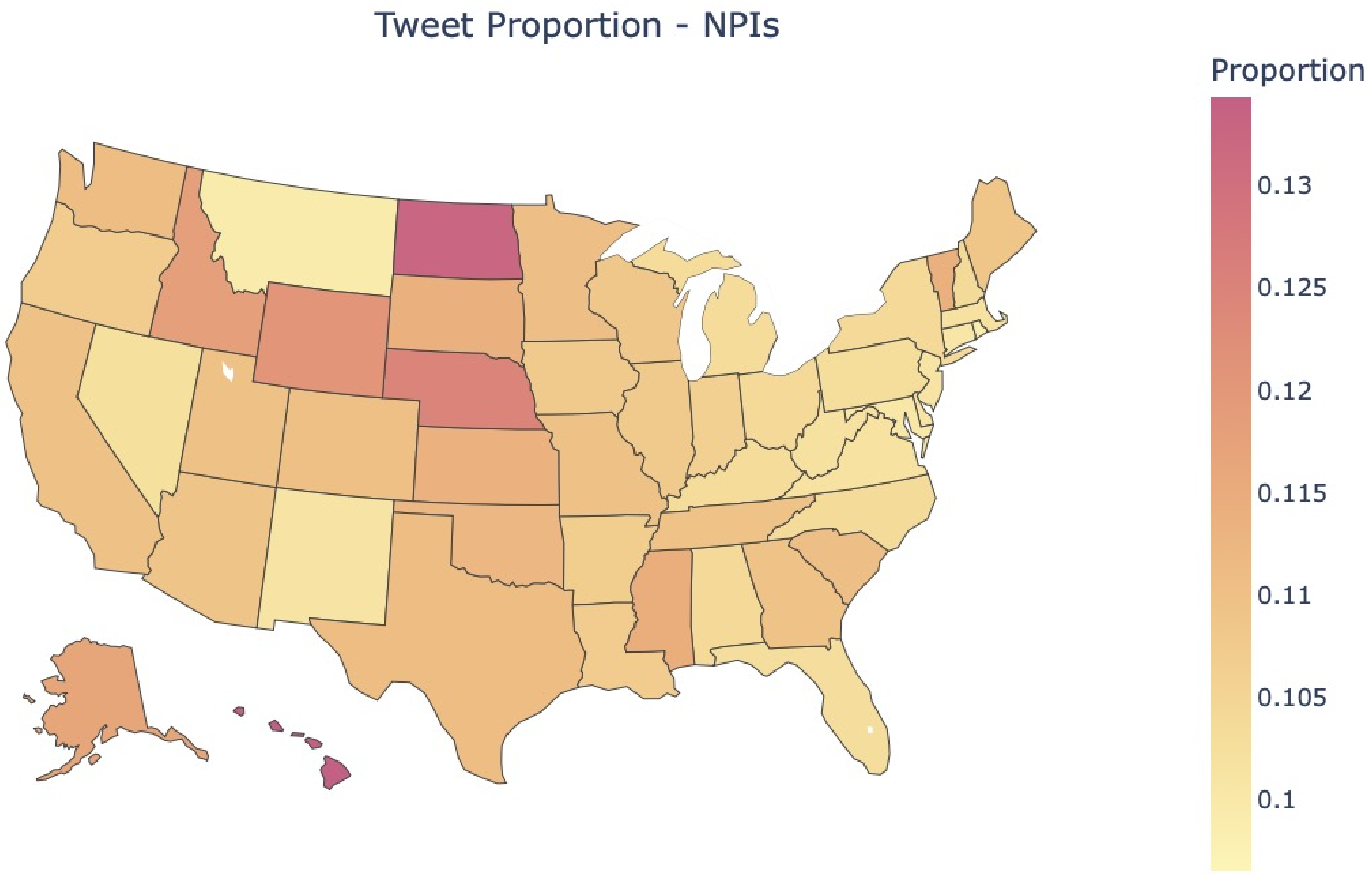


Figure 4.19: Tweet proportion by state - NPIs

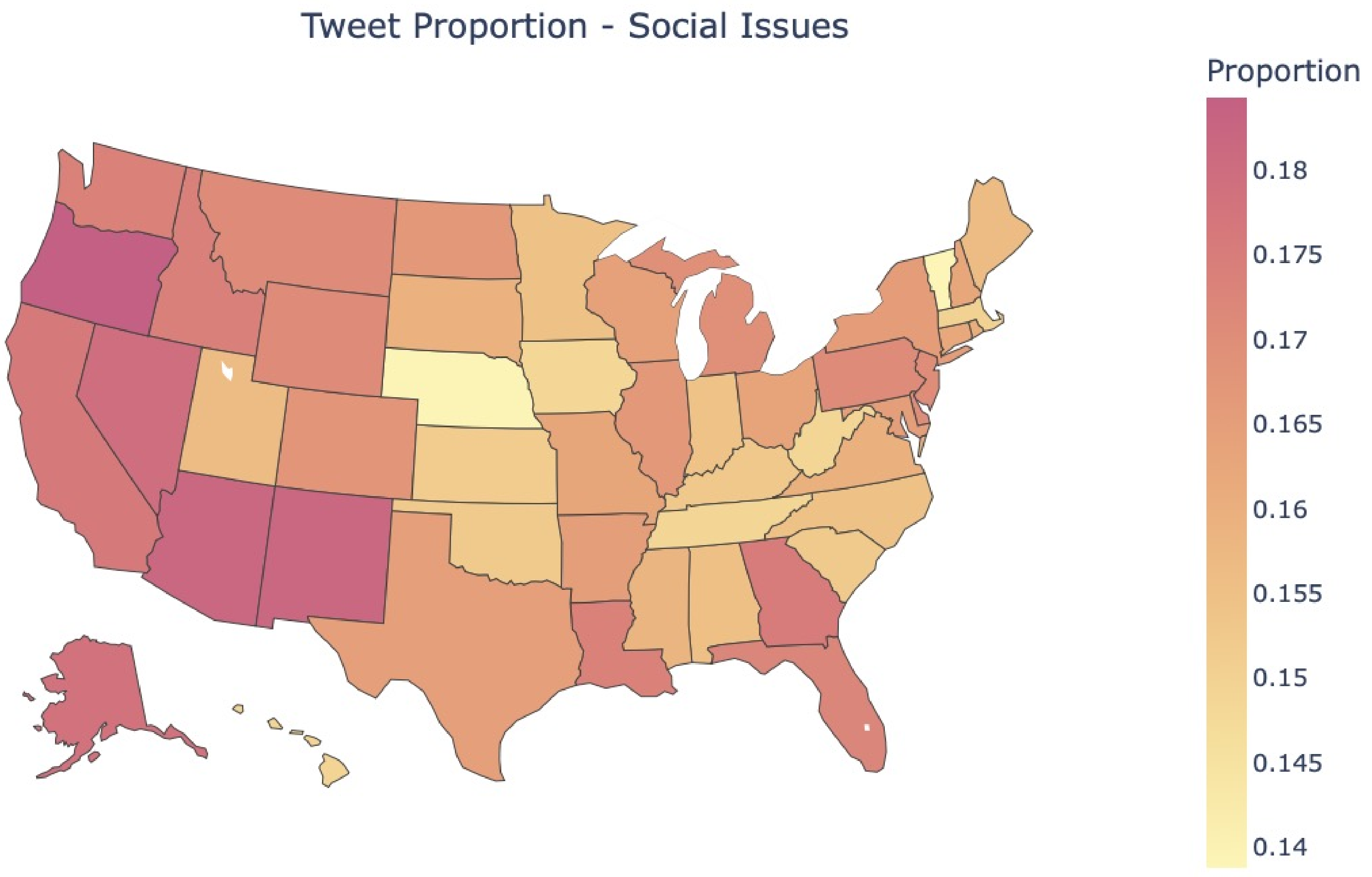


Figure 4.20: Tweet proportion by state - Social Issue

For the topic of social issues, we have observed an national average of 16.6%, showing that 1 in 6 tweets are discussing about social issues. We have observed difference in state level when talking about social issues, with the state of Oregon, Arizona, and Minnesota having the highest rate of social issue related discussion, with over 18% of tweets on such topics. On the other hand, the state of Vermont and Nebraska has less than 14% of tweets discussing social Issues.

We have observed difference in the proportion of topic being discussed in state level. And it seems in most states, the discussion on the two topics are exclusive to each other, with states prefer to discuss on NPIs and in contrast less discussion on social issue, or vise versa. While there are states like Idaho and North Dakota, which have relative high proportion of discussion on both topics.

4.3.4.3 Sentiment

An important task in our study is to trace the sentiment change in different degree. Previously, we showed the sentiment change in in the time scale, in addition, we are also interested in comparing the sentiment in state level. As Figure 4.21 shows, we found there are rather significant difference in the overall sentiment on state level. The overall sentiment for each state ranging from -0.13 to 0, showing the sentiment regarding the COVID-19 differs from state to state. Among all the states, we found Hawaii, Vermont, and Massachusetts have the overall highest sentiment, with the overall sentiment close to neutral. While the state of Wyoming, Arizona, and Oregon has the lowest sentiment score.

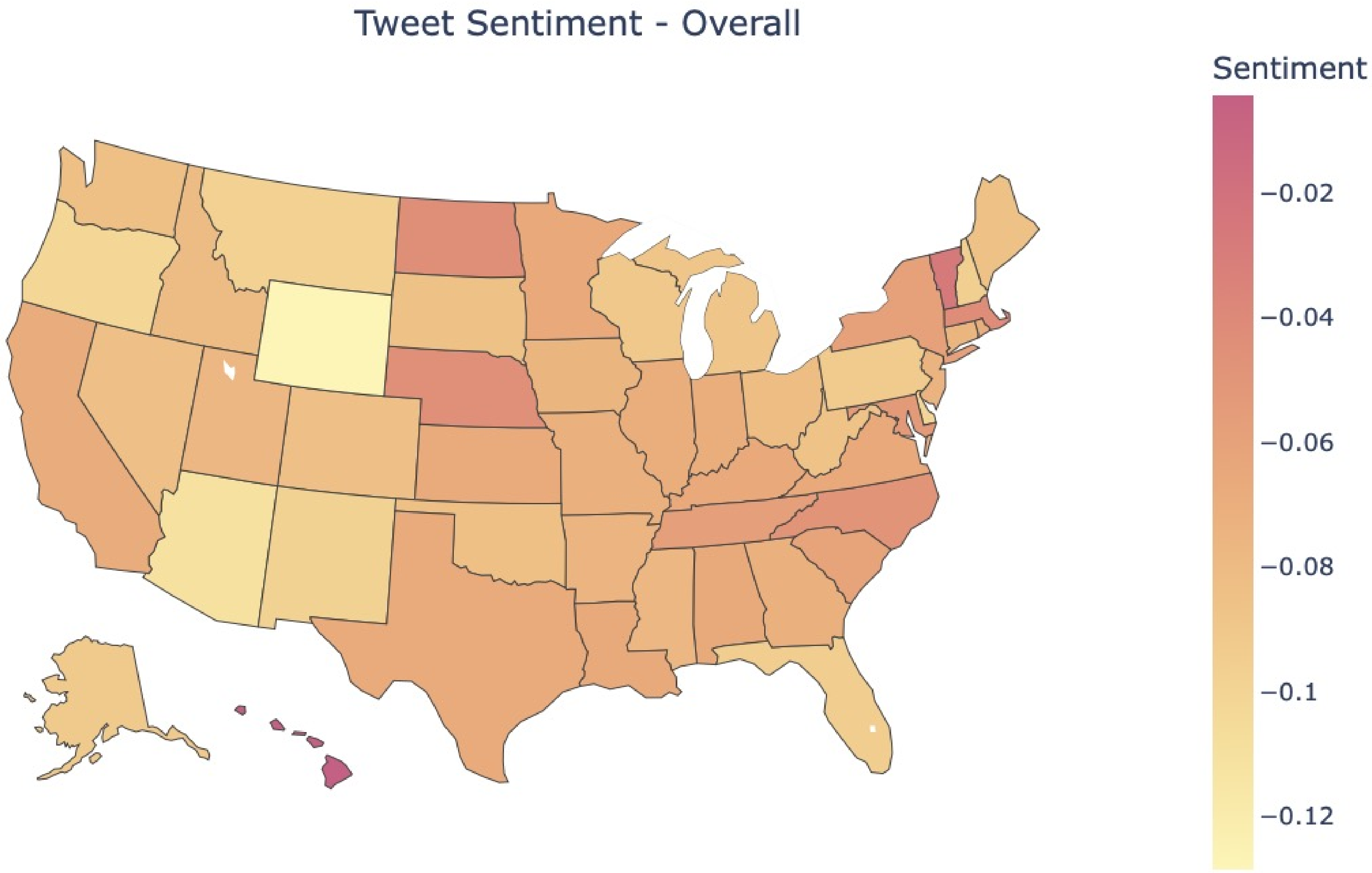


Figure 4.21: Tweet sentiment by state]

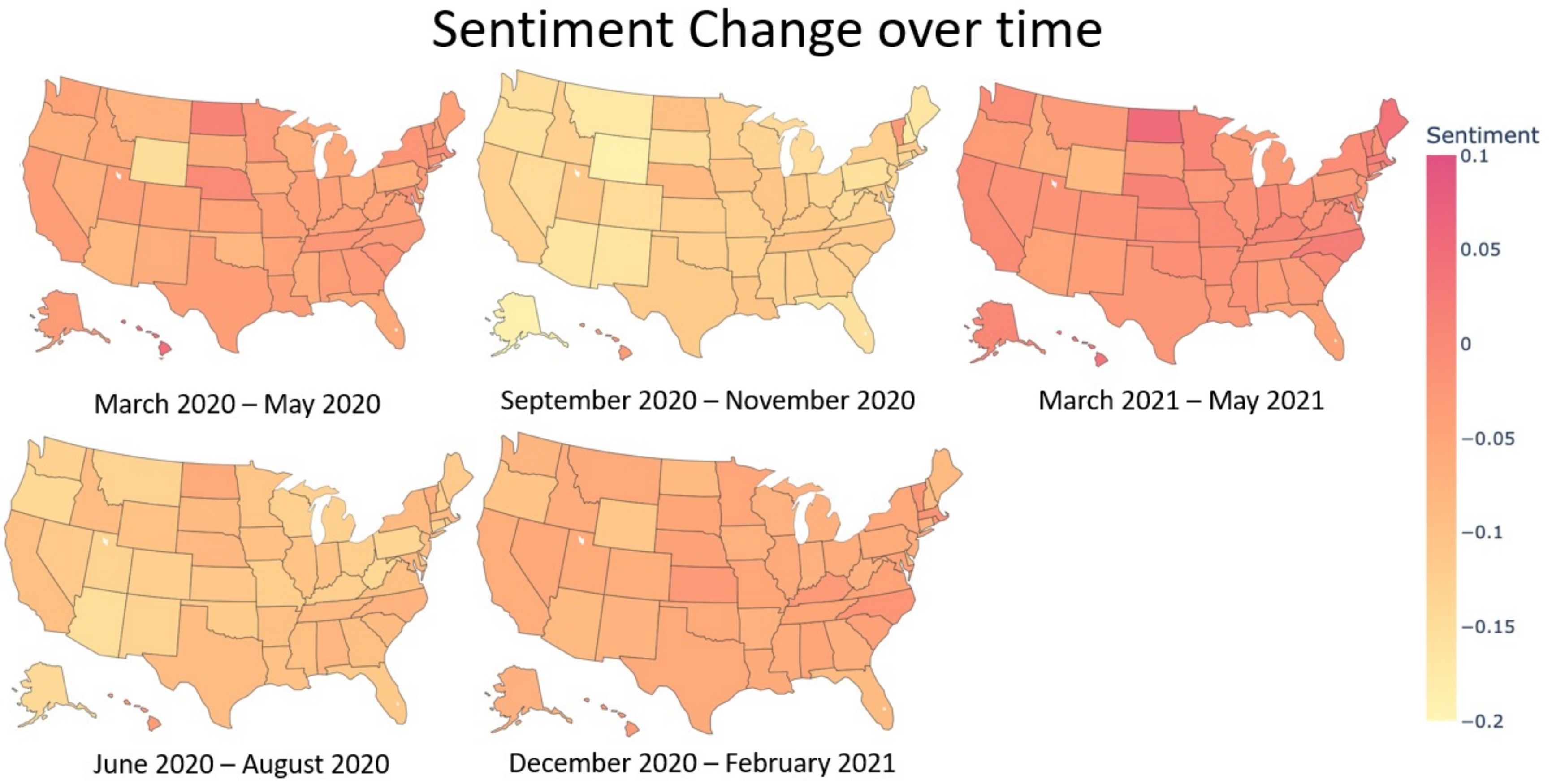


Figure 4.22: Tweet sentiment Change by state

In addition to the sentiment score across the whole time period, we also break down the whole 15-month period to five 3-month periods. We want to find which state has the highest sentiment change over time period. We performed a summation of the absolute difference between any of the two continuous periods. As the Figure **??** shows, we find Vermont, Delaware, and Iowa has the lowest sentiment change, with absolute change over time period of around 12%. On the flip side, we found Alaska, Maine, and New Hampshire have the highest change in sentiment, with over 30% of absolute value. Overall, we noticed a downshift of sentiment across the United State between September 2020 to November 2020, where the second wave of the pandemic and election happens. The overall sentiment gets better afterward after the approval and the mass shipping of COVID-19 vaccine.

We are also interested in the specific topics, the NPIs and social issues about how it differs from state to state. As Figure 4.23 shows, the overall sentiment on NPIs are neutral, with Vermont and Hawaii have relative higher sentiment of 0.08 compared to other states. And the state South Dakota and Montana has the lowest sentiment close to -0.05. The rest of the state are relatively close in sentiment.

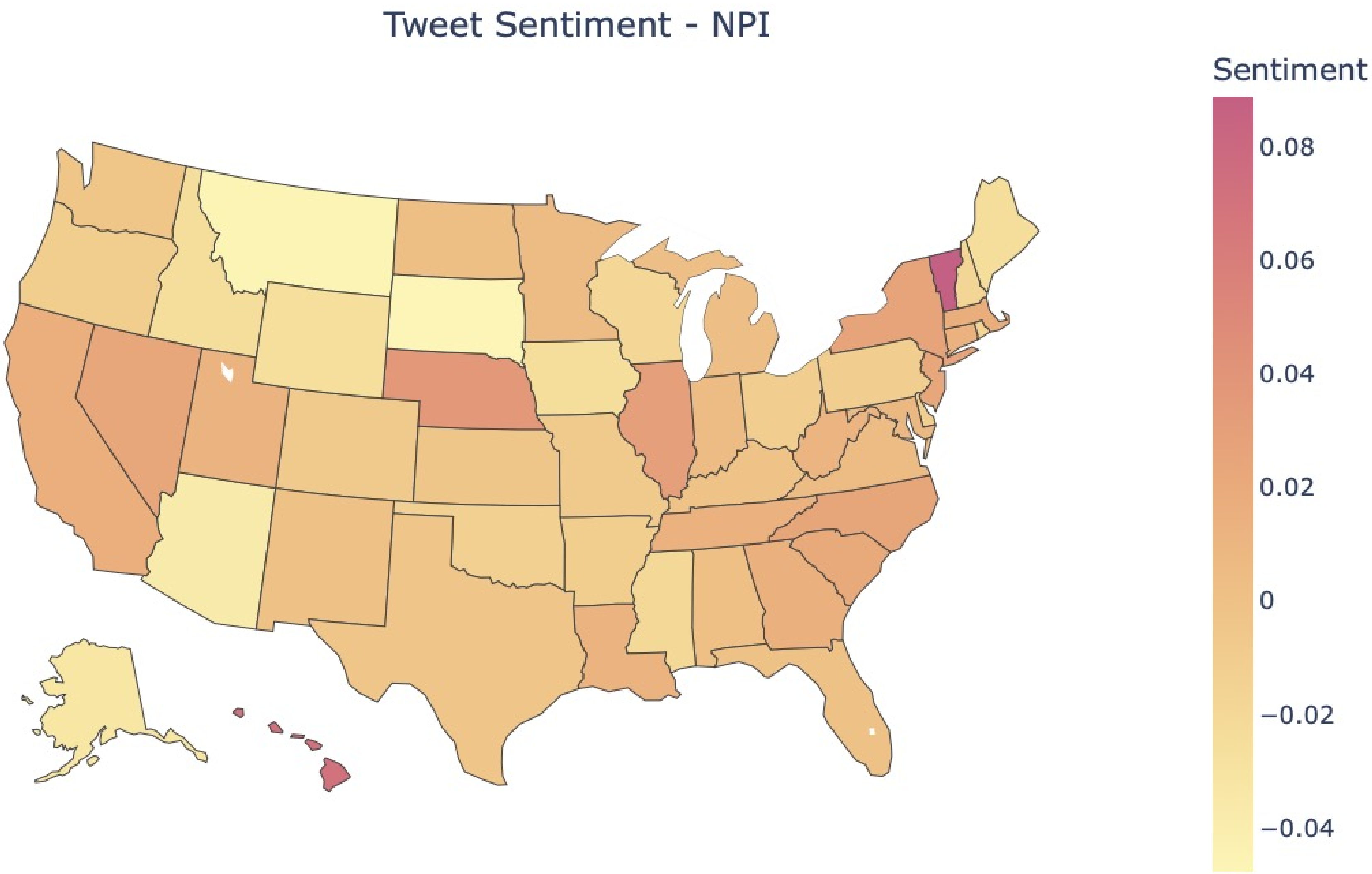


Figure 4.23: Tweet sentiment by state - NPIs

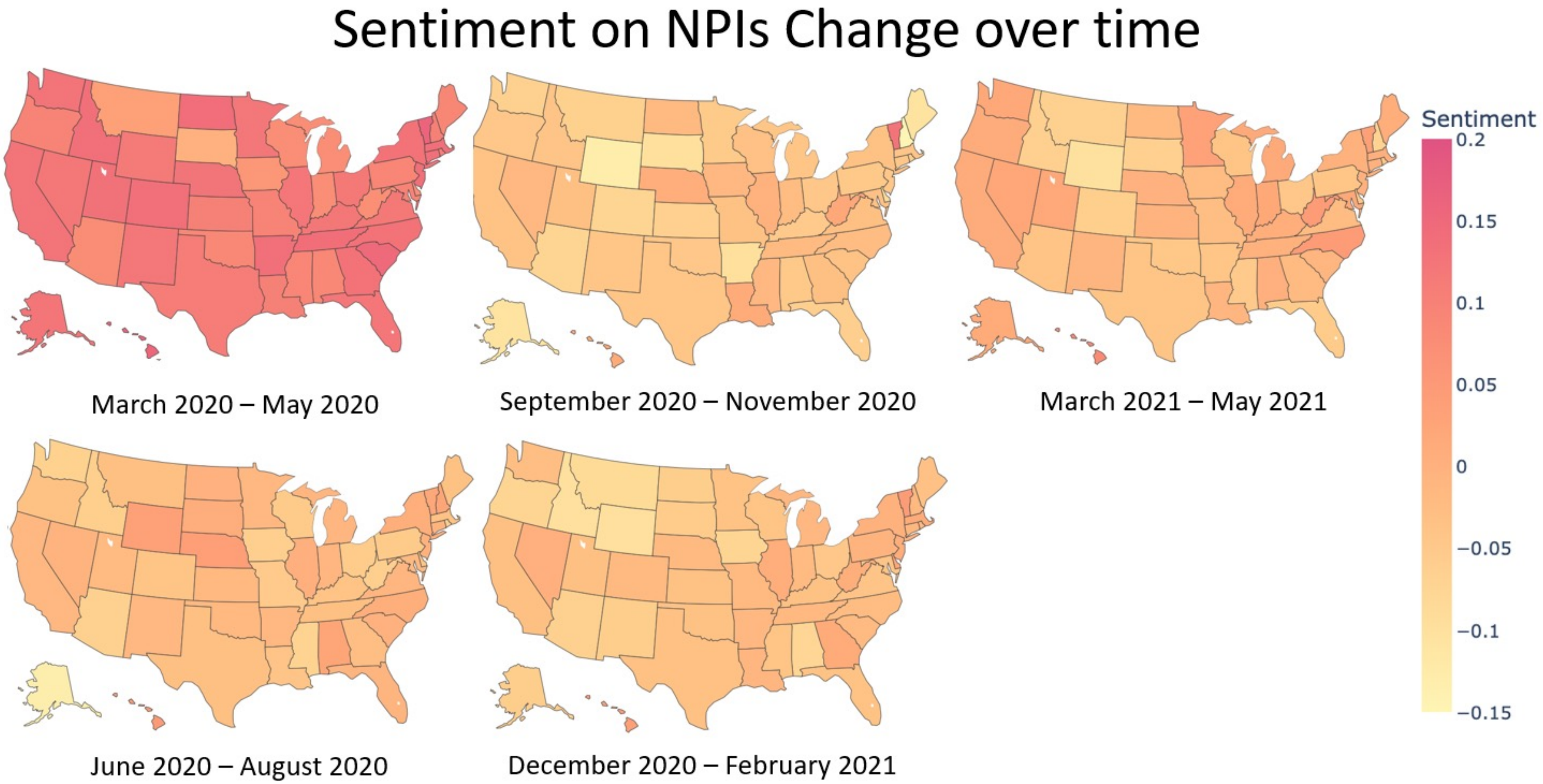


Figure 4.24: Tweet sentiment Change by state - NPIs

Similar to the change in over all sentiment across the states, we also performed a study on the change of sentiment in NPIs. Unlike the overall sentiment, as shown in Figure 4.24, we see a initial neutral to positive sentiment of 0.11 across the United States. For the rest time period from June 2020 to May 2021, the overall sentiment is -0.02 across the nation. This is possibly due to the mask mandate that certain group of people are not willing to comply. Overall, the sentiment seems not having significant pattern geospatially.

In terms of the sentiment on social issues, as shown in Figure 4.25 we find in different states, the overall sentiment range from -0.41 to -0.49 showing that the public generally have a negative sentiment on social issues. Specifically, we find Mississippi, Maryland, and Georgia ranked top in terms of the sentiment on social issues, and Connecticut, New Hampshire, and Wyoming ranked the lowest.

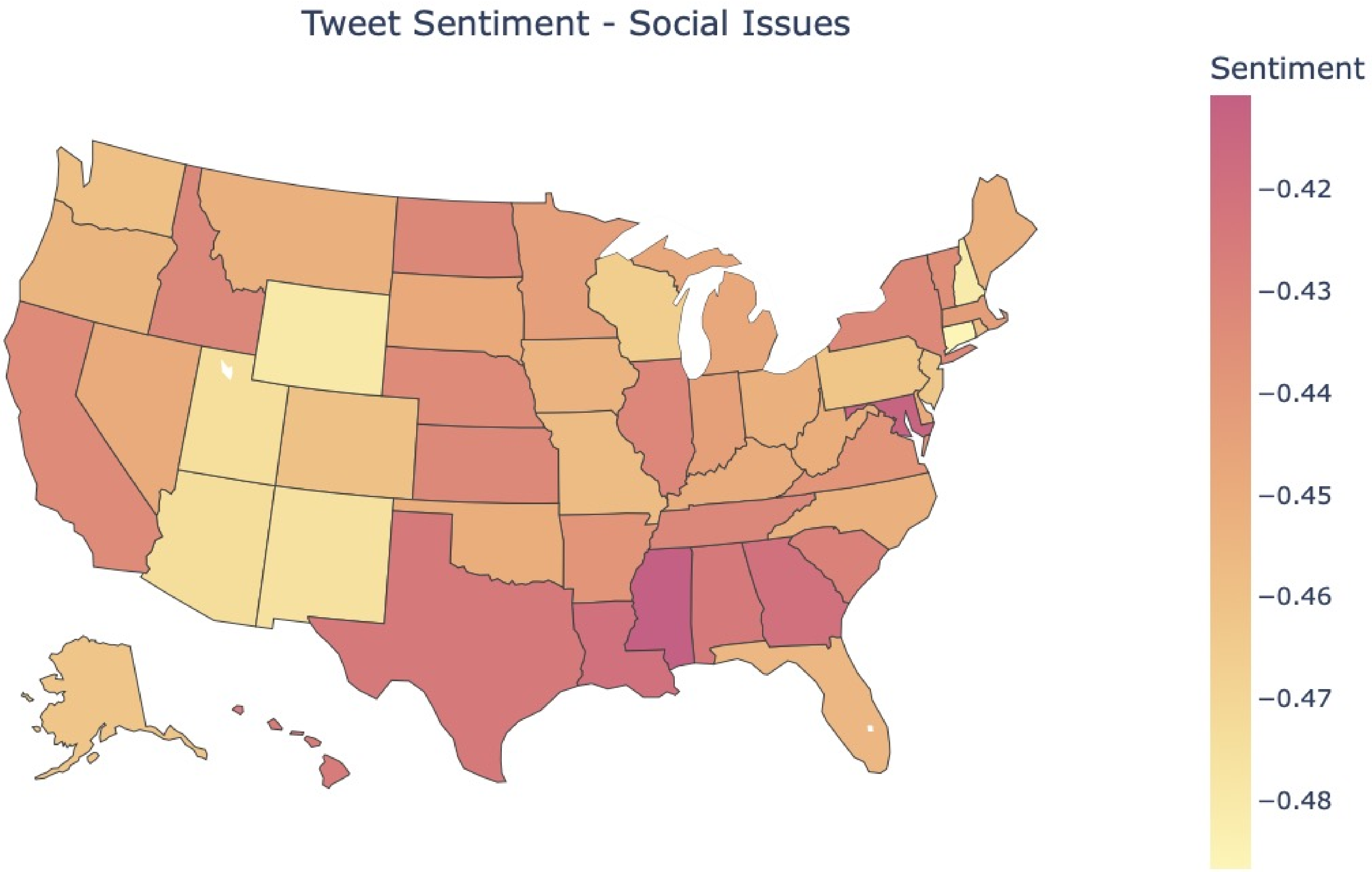


Figure 4.25: Tweet sentiment by state - Social Issue

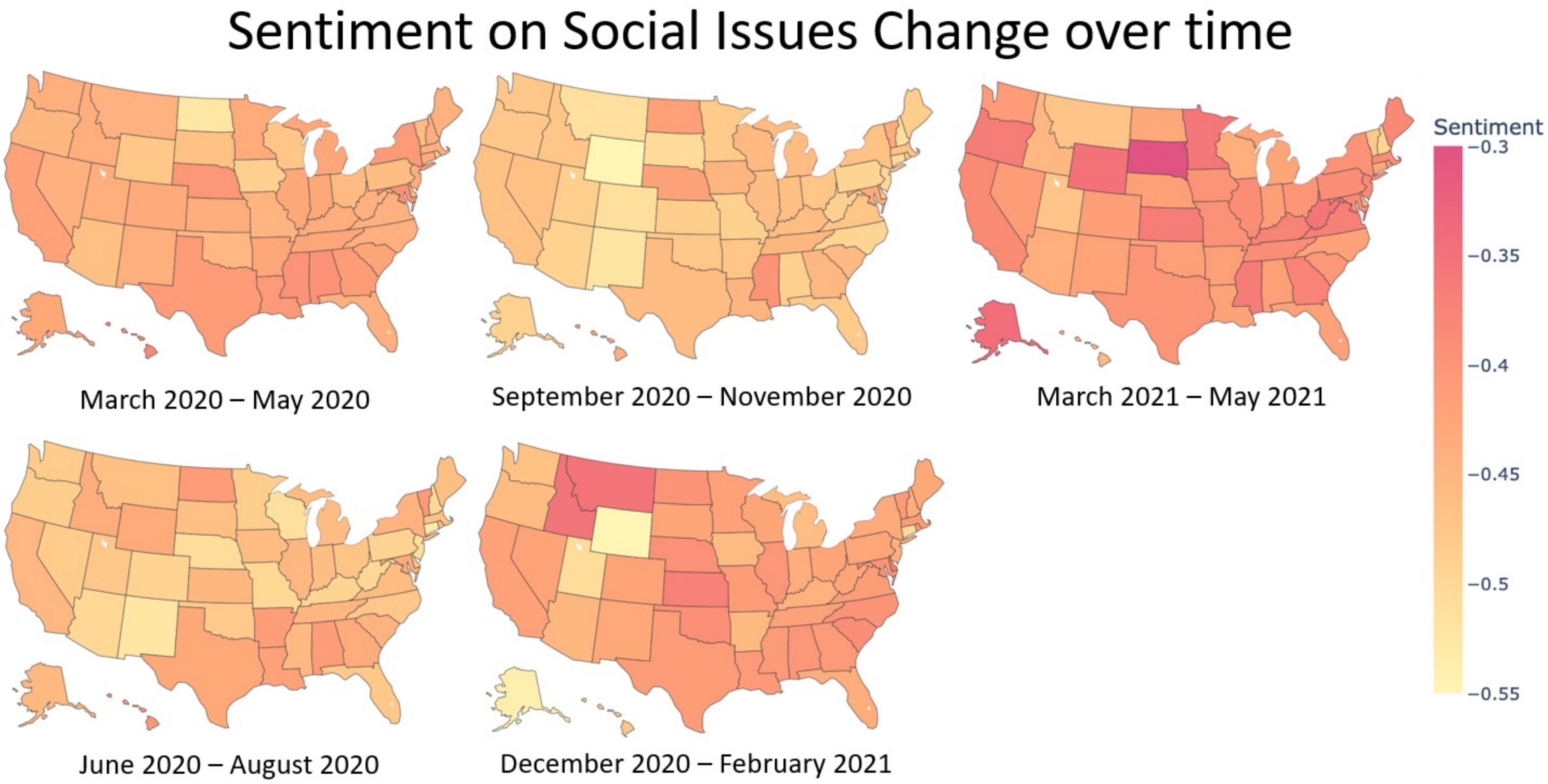

Sentiment on Social Issues Change over time
March 2020 – May 2020
September 2020 – November 2020
March 2021 – May 2021
June 2020 – August 2020
December 2020 – February 2021
Sentiment
−0.3
−0.35
−0.4
−0.45
−0.5
−0.55

Figure 4.26: Tweet sentiment Change by state - Social Issue

The Figure 4.26 shows the sentiment change on social issues for each 3-month periods. We have observed the change and noticed the change that during September 2020 to November 2020, the sentiment on social issues sees the dip the most, with an overall sentiment of -0.47. This means the overall discussion on social issues are dominated with neutral and negative sentiments. In contrast, the overall sentiment on social issues from December 2020 to May 2021 has an overall sentiment of -0.41. Although it is still relatively negative, it has been a major improvement over the previous period. We believe this is due to the aftermath of election seasons, and the pandemic has peaked, with the release of vaccine and number of cases going down.

4.3.4.4 Analysis on 4 populous state

In this section, we analysis the tweeting behavior in 4 of the state with the most population. We try to find if there are certain state policies that will shift the topic being discussed in social media as well as the sentiment. If such case can be found, it shows our model is capable of find the state level change, and can be potentially applied by state level health agencies to refine the guideline and communicate with the general public.

First we start by comparing the difference in the topic proportion of NPIs. As Figure 4.27 shows, there are certain separation of trend in the topic of NPIs between the 4 states. We notice a peak in the state of Texas on March of 2021, which fit the timeline of Texas's reopen plan announced on March 2021. Similarly, there is a peak in the state of Florida on late April 2021, which fit the timeline of Florida's rule to not allowing mandatory mask wearing in public schools.

For the social issues, as Figure 4.28 shows, there are some difference between the state, but since it is following the similar overall trend, we are not able to locate any special events that can shift the discussion on social issues.

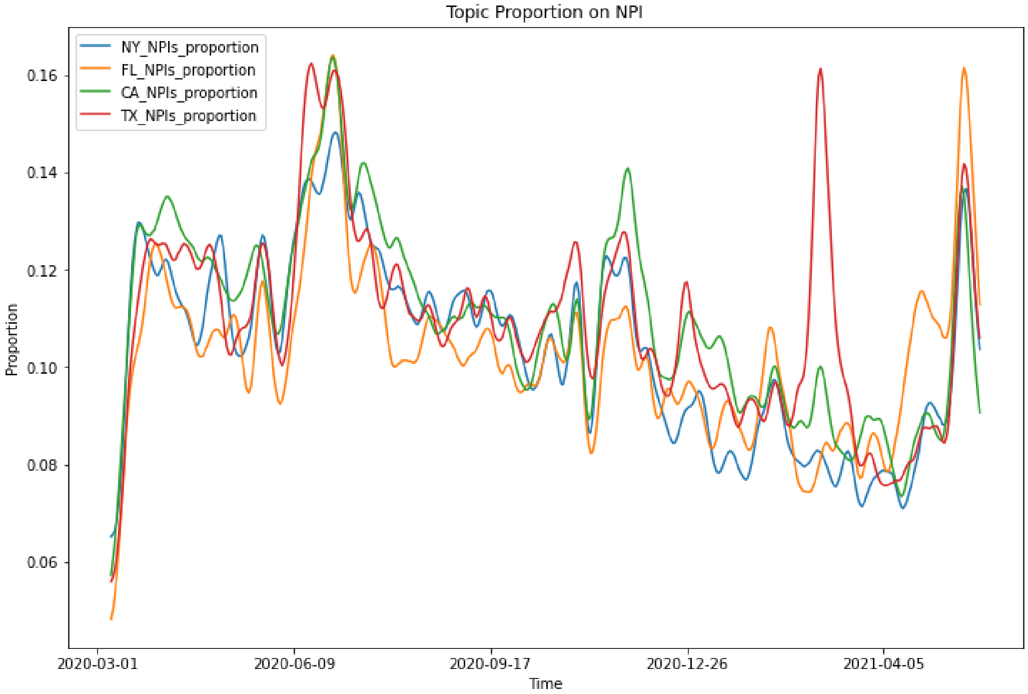


Figure 4.27: NPIs related tweets proportion in 4 states

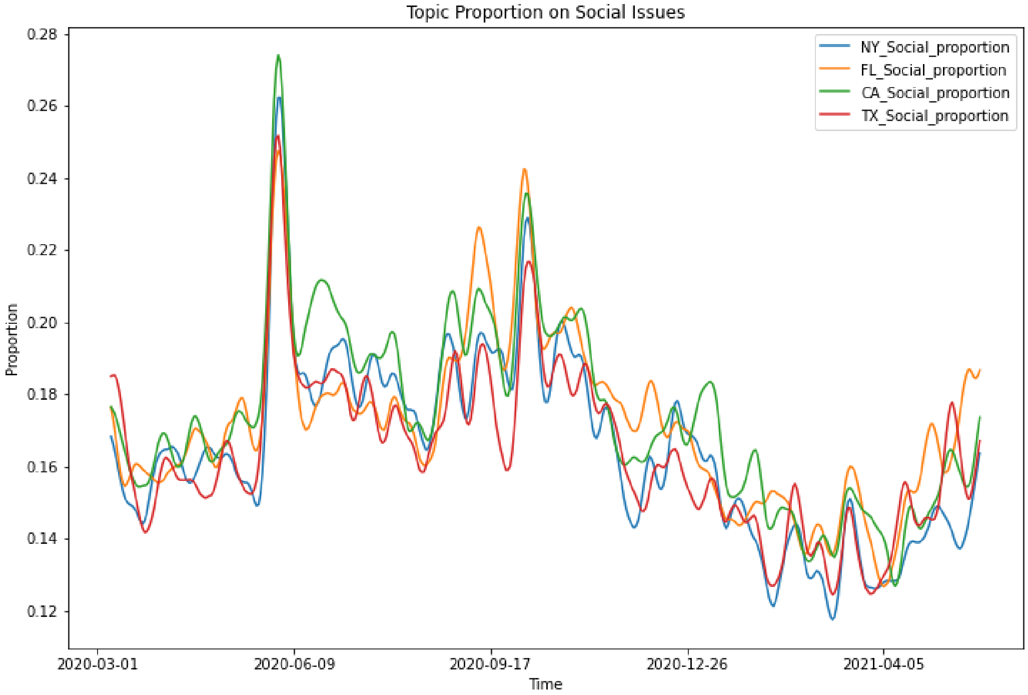


Figure 4.28: Social Issues related tweets proportion in 4 states

We also performed an analysis on the sentiment of the 4 states. As the overall sentiment (Figure 4.29) shows the state of Florida has an overall sentiment lower than the rest of the state, while the other 3 states shows similar value and trend in sentiment. Overall, the trend on the sentiment is most similar among the 4 states.

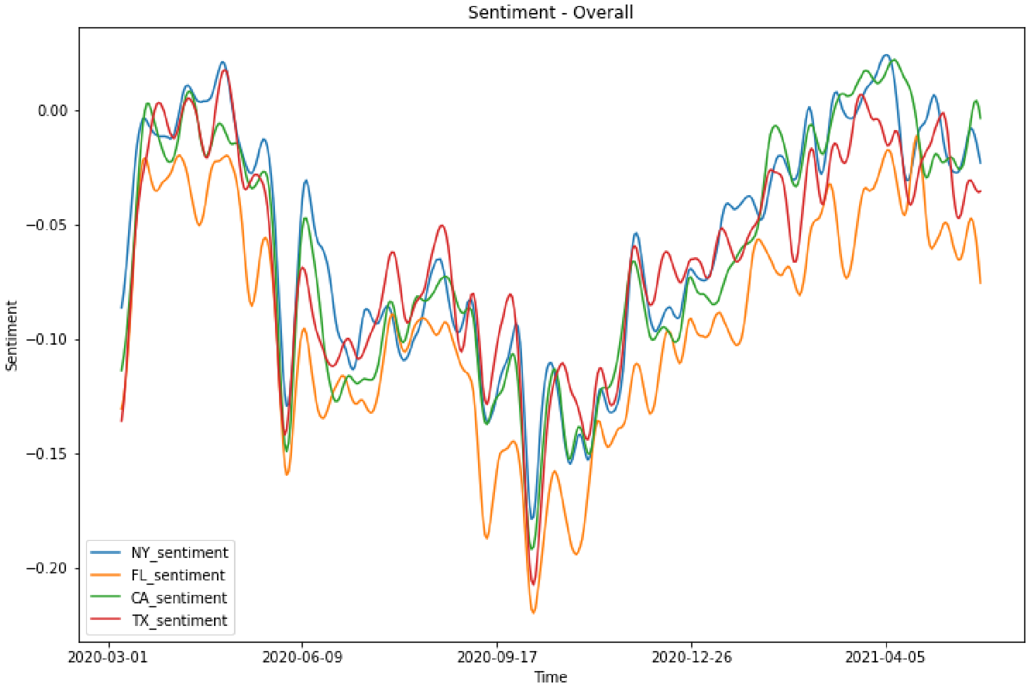


Figure 4.29: Tweet sentiment 4 states

Figure 4.30 and Figure 4.31 shows although the 4 state have shared trend in some vases, they show difference and random patterns. We believe this is because the discussion is on certain topics that associated with the state, causing each state has it's own pattern. We also notice a dip in sentiment for the state of Florida on social issues on April 2021, we believe this may be caused by the upcoming spring break as a lot of people will visit Florida.

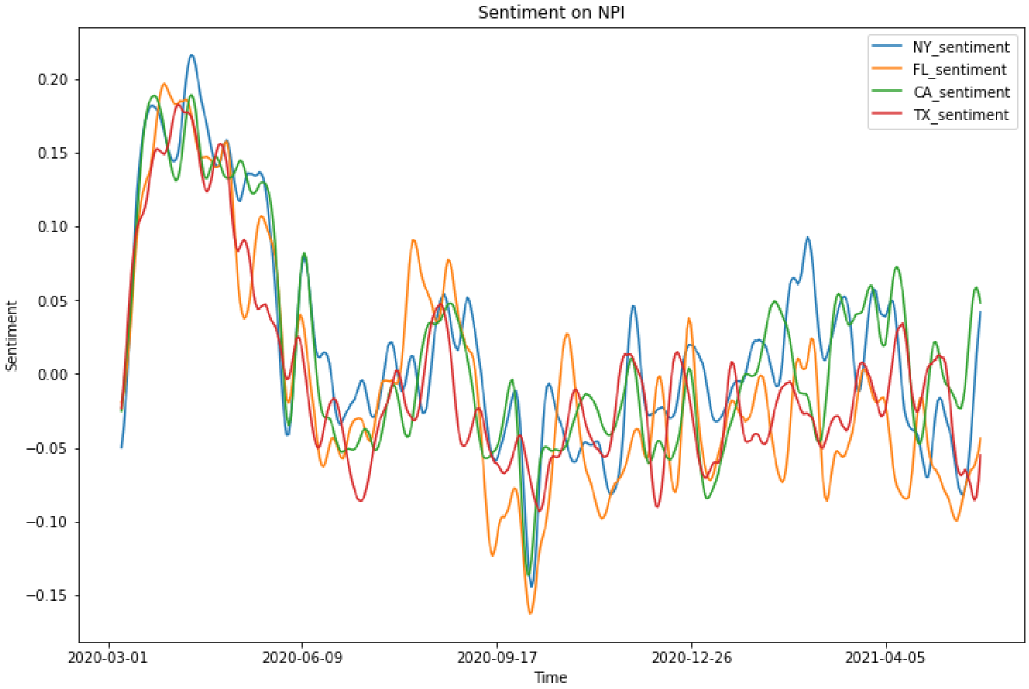


Figure 4.30: Tweet sentiment 4 states - NPIs

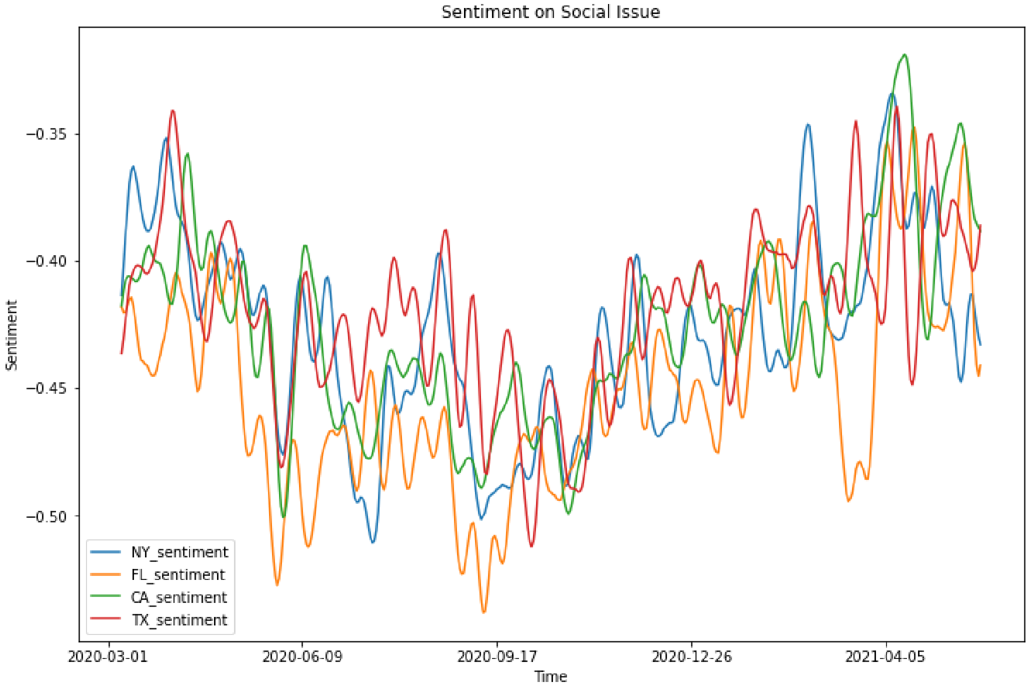


Figure 4.31: Tweet sentiment 4 states - Social Issue

The comparison of the 4 states shows the model is able to capture certain movement in the COVID-19 related topics on social media. In particular, we have observed the rise of NPIs related discussion in Texas and Florida, which matches the announcement of NPIs related guidelines.

4.3.4.5 Conclusion of Geospatial analysis

Overall, we have observed the state-level difference regarding the number, the topic proportion, and the sentiment of the COVID-related tweets. In particular, we are able to find some trend in state-level comparison in the number and the topic proportion of tweets. It also shows the policy in different state will result people in different state to react differently. The state level difference also shows, our model is capable of capturing the difference in tweeting behavior in state level.

4.4 Conclusion

In this discussion, we established the workflow of using Twitter data to monitor the change in online discussion. By combine the topic classification, sentiment classification, and a fixed set of randomly pulled tweets, we are able to monitor the trend in online discussion, capture major events, and showing the similarity and difference between certain subset of the data.

In the data collection, we implement an solution that is different than previous methods. Instead of using the count of hashtags, we used a sampling method of extracting a fixed number of tweets daily, and then performed topic classification to monitor the change in the trend. In addition, it allows us to keep track of certain sub-topic like NPIs, countermeasures, social issues, policies, and public responses. If we see the raise of one topic, we can examine the tweets labeled in the topic and find the content that is driving the discussion. Thus, public health agencies can make better response and more effective communication with the public.

The topic classification and sentiment classification models showed they are performing as expected. We are able to train topic and sentiment classification models using a relatively small dataset. Even though we recon the dataset is not in the highest quality we expected, we are still able to build models that can capture the trend, key events, and showing the difference in the tweets. In this study, we chose BERT-Base as our production model to balance the training and validation cost. Still, the result shows such model can still significantly outperform the previous state-of-the-art models, and showed the power of deep-learning based NLP methods. We also tested the deep-learning model that are trained in the same context, and with more parameters. The result shows CT-BERT has the best performance overall, and we believe if computation resources allows, such models should be considered to pursuer the best performance.

By applying the topic classification and sentiment classification models to over 4 million tweets, we have graph an overall trend in terms of the topic proportion and sentiment over 15 months. We observed peaks and dips of NPIs, during the beginning of the pandemic, and the more recent upraise of the COVID-19 infection. The public has an overall neutral sentiment towards the NPIs, while we observed an apparent downturn when CDC published its guideline to mandate masks. On the topic of social issue, the overall sentiment is negative, which meets our perception. The model shows we captured certain key events during the pandemic, and reflected with sudden dive in the sentiment.

In addition to plotting the trend, we also performed comparison between subsets of the data. We first compared tweets with geo-tags and without geo-tags. We find they are highly correlated with the shift in the trend, while in terms of absolute sentiment and topics being discussed, they are quite different. We noticed a huge difference in the proportion of tweets that are NPIs related, showing tweets with geo-tags has more discussions on the NPI. However, for the proportion of social issues, the difference does not seem to be that significant. Similar situation applies to the sentiment, with geo-tagged tweets showing more positive on the sentiment overall and NPIs, the difference remains marginal on the topic of social issues. We then compared tweets generated from top 50 cities with most tweets and the tweets generated from the rest of cities. We observed difference in certain time periods, while the trend mostly remains identical. We also found the sentiment was generally more positive on the tweets from top 50 cites, with sentiment on social issue not highly correlated. In conclusion, we believe even using different subsets of tweets, the movement of trend can still be captured, and we can use the sudden peak and dip to detect certain ongoing events.

In this study, we take the advantage of state-of-the-art NLP methods and proposed a workflow that can monitor and capture the health-related topic and sentiment trend in social media. We used this approach to analyze the recent COVID-19 pandemic, and found certain trend on social media that matches our perception. We also observed some difference in the proportion of topics being discussed and the sentiment associated with those topics between certain subsets of the data we extracted. It is also our interest to propose a complete system with automatic detection of an upraise trend and a sudden divert of sentiment, and can automatically extract the related discussion in the future work.

4.5 Contribution

The major contribution of this part is we applied NLP to detect public health related trend on social media. The approach is able to detect the topic shift caused by major events during the pandemic, and also showed the difference in social media discussion between major cities and the rest. We also capture the sentiment of social media at state level. Such kind of system can be further developed as a info-surveillance system that detect the abrupt change in social media, allowing public health related agency to generate better response.

Charpter 4 Reference

[1] I. C. Fung et al. “Ebola and the social media”. In: *Lancet* 384.9961 (Dec. 2014), p. 2207.

[2] L. Hossain et al. “Social media in Ebola outbreak”. In: *Epidemiol Infect* 144.10 (July 2016), pp. 2136–2143.

[3] X. Gui et al. “Understanding the Patterns of Health Information Dissemination on Social Media during the Zika Outbreak”. In: *AMIA Annu Symp Proc* 2017 (2017), pp. 820–829.

[4] Solmaz Filiz Karabag. “An Unprecedented Global Crisis! The Global, Regional, National, Political, Economic and Commercial Impact of the Coronavirus Pandemic”. In: (Mar. 2020), pp. 1–6.

[5] Frank Dignum et al. “Analysing the Combined Health, Social and Economic Impacts of the Corovanvirus Pandemic Using Agent-Based Social Simulation”. In: *Minds and Machines* 30 (June 2020). DOI: 10.1007/s11023-020-09527-6.

[6] Martin Müller, Marcel Salathé, and Per E Kummervold. “COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter”. In: *arXiv preprint arXiv:2005.07503* (2020).

[7] C. D. Corley et al. “Text and structural data mining of influenza mentions in Web and social media”. In: *Int J Environ Res Public Health* 7.2 (Feb. 2010), pp. 596–615.

[8] D. A. Broniatowski, M. J. Paul, and M. Dredze. “National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic”. In: *PLoS One* 8.12 (2013), e83672.

[9] A. A. Aslam et al. “The reliability of tweets as a supplementary method of seasonal influenza surveillance”. In: *J Med Internet Res* 16.11 (Nov. 2014), e250.

[10] S. F. McGough et al. “Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data”. In: *PLoS Negl Trop Dis* 11.1 (Jan. 2017), e0005295.

[11] M. O. Lwin et al. “Global Sentiments Surrounding the COVID-19 Pandemic on Twitter: Analysis of Twitter Trends”. In: *JMIR Public Health Surveill* 6.2 (May 2020), e19447.

[12] A. Abd-Alrazaq et al. “Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study”. In: *J Med Internet Res* 22.4 (Apr. 2020), e19016.

[13] B. J. Cowling et al. “Impact assessment of non-pharmaceutical interventions against coronavirus disease 2019 and influenza in Hong Kong: an observational study”. In: *Lancet Public Health* 5.5 (May 2020), e279–e288.

[14] S. Lai et al. “Effect of non-pharmaceutical interventions to contain COVID-19 in China”. In: *Nature* 585.7825 (Sept. 2020), pp. 410–413.

[15] S. E. Eikenberry et al. “To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic”. In: *Infect Dis Model* 5 (2020), pp. 293–308.

[16] L. He et al. “Why do people oppose mask wearing? A comprehensive analysis of U.S. tweets during the COVID-19 pandemic”. In: *J Am Med Inform Assoc* 28.7 (July 2021), pp. 1564–1573.

[17] Abraham Sanders et al. “Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse”. In: (Sept. 2020). DOI: 10.1101/2020.08.28.20183863.

[18] Aron Culotta. “Towards detecting Influenza Epidemics by Analyzing Twitter Messages”. In: July 2010. ISBN: 978-1-4503-0217-3. DOI: 10.1145/1964858.1964874.

[19] Y. T. Yang, M. Horneffer, and N. DiLisio. “Mining social media and web searches for disease detection”. In: *J Public Health Res* 2.1 (Apr. 2013), pp. 17–21.

[20] C. W. Schmidt. “Trending now: using social media to predict and track disease outbreaks”. In: *Environ Health Perspect* 120.1 (Jan. 2012), A30–33.

[21] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. “Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter”. In: Jan. 2011, pp. 1568–1576.

[22] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: *the Journal of machine Learning research* 3 (2003), pp. 993–1022.

[23] Tomas Mikolov et al. “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781* (2013).

[24] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “Glove: Global vectors for word representation”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[25] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[26] Armand Joulin et al. “FastText.zip: Compressing text classification models”. In: *arXiv preprint arXiv:1612.03651* (2016).

[27] Matthew E. Peters et al. *Deep contextualized word representations*. 2018. arXiv: 1802.05365 [cs.CL].

[28] Zhilin Yang et al. *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. 2020. arXiv: 1906.08237 [cs.CL].

[29] Alec Radford et al. “Language Models are Unsupervised Multitask Learners”. In: (2019).

[30] Jinhyuk Lee et al. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. In: *Bioinformatics* (Sept. 2019). Ed. by JonathanEditor Wren. ISSN: 1460-2059. DOI: `10.1093/bioinformatics/btz682`. URL: `http://dx.doi.org/10.1093/bioinformatics/btz682`.

[31] Yifan Peng, Shankai Yan, and Zhiyong Lu. “Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets”. In: *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*. 2019, pp. 58–65.

[32] Laila Rasmy et al. “Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction”. In: (May 2020).

[33] Nils Reimers and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. 2019. arXiv: `1908.10084 [cs.CL]`.

CHAPTER 5: Conclusion

In this big data era, we believe the modern NLP methods can help to understand and extract useful information form health-related texts. There are still plenty of challenged to be tackled, posted by the variation of health-related texts and different tasks that are desired to be achieved. Over the years, various NLP tools and methods have been developed, in machine learning and deep learning, to address various kind of NLP related questions. This provide us an opportunity to streamline the process to achieve the goal of understand and extract information from health-related texts.

In this dissertation, we explored three different type of health-related texts, and implemented NLP methods with automation to address the most desired problem in each kind. The first type is EHRs, we proposed a deep learning based method to extract blood pressure readings from the clinical narratives. By combining regular expression and deep learning based methods, we reduced the annotation work load and implemented a model with no domain knowledge. It outperformed the previous state-of-the-art methods in extracting and phenotyping the blood pressure readings. We also showed such method can be set up easily and the model can be applied in a wide variety of tasks.

The second type is scientific publications, we proposed a BERT based self-supervised extractive text summarization tool for biomedical literatures. We used abstract in the publications to find information rich sentences in the body text of the literatures. We then used BERT to classify the sentence and decide whether certain sentences should be included in the summary. The novelty of this approach is the overall process does not require any human annotation, as the whole process is self-supervised. The result shows we are able to find sentences in the body text that are contextually similar to the sentences in the abstract, and generate a one to two page summary that cover the content in the abstract. Such summary can be useful to the clinicians, and can help them to catch up with the latest research without reading the whole paper, while getting much more information than abstract only. The model can also be seen as a base model for other transfer learning studies, as in certain situations abstract is not available in articles.

The third type is social media, we proposed a BERT based topic classification and sentiment classification tool to understand the public's attitude about COVID-19. We extracted COVID-19 related tweets from Twitter using their academic API. Instead of counting the number of hashtags or using google's search trend like previous works, we propose to extract fix number of tweets daily and apply topic classification and sentiment classification on those tweets. We trained and compared several BERT based models, and chose BERT-Base as our model in application for the balance of performance and efficiency. We applied the model on over 4 million tweets over 15 months, and observed the change in topic and sentiment trend. By comparing the trend with certain major events in the pandemic, we believe our model is sensitive enough to capture the change in the trend. We also found the similarity and difference in the proportion and the sentiment of certain topics. We believe this model can be applied in health-related social media posts, and capture the sudden topic and sentiment change. With such information, health agencies can react efficiently and communicate effectively with public on the trending issues.

In this dissertation, we used several Deep Learning based NLP methods. The models including the combination of static encoding model like GloVe with CNN, the context aware model like BERT with conventional machine learning model, BERT with Deep Neural Network classifier, and CT-BERT that are pre-trained in domain specific context. The result shows Deep Learning is capable of handling different NLP task, and has the versatility that can be trained on different sample size. With the design of certain element to automate the workflow especially the annotation, we are able to perform the studies in relative fast pace, while still retain good performance that allow us to extract the useful information or understand the trend in the data.

Bibliography

[1] Valentina Casola et al. “Healthcare-related data in the cloud: Challenges and opportunities”. In: *IEEE cloud computing* 3.6 (2016), pp. 10–14.

[2] Hongfang Liu et al. “An information extraction framework for cohort identification using electronic health records”. In: *AMIA Summits on Translational Science Proceedings* 2013 (2013), p. 149.

[3] Vasudevan Jagannathan et al. “Assessment of commercial NLP engines for medication information extraction from dictated clinical notes”. In: *International journal of medical informatics* 78.4 (2009), pp. 284–291.

[4] Oliver Baclic et al. “Artificial intelligence in public health: Challenges and opportunities for public health made possible by advances in natural language processing”. In: *Canada Communicable Disease Report* 46.6 (2020), p. 161.

[5] Ergin Soysal et al. “CLAMP â a toolkit for efficiently building customized clinical natural language processing pipelines”. In: *Journal of the American Medical Informatics Association* (Nov. 2017), ocx132. DOI: `10.1093/jamia/ocx132`.

[6] Guergana K Savova et al. “Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications”. In: *Journal of the American Medical Informatics Association* 17.5 (Sept. 2010), pp. 507–513. ISSN: 1527-974X. DOI: `10.1136/jamia.2009.001560`. eprint: `http://oup.prod.sis.lan/jamia/article-pdf/17/5/507/5940551/17-5-507.pdf`. URL: `https://doi.org/10.1136/jamia.2009.001560`.

[7] Olivier Bodenreider. “The Unified Medical Language System (UMLS): Integrating Biomedical Terminology”. In: *Nucleic acids research* 32 (Feb. 2004), pp. D267–70. DOI: `10.1093/nar/gkh061`.

[8] Oliver Ferschke, Johannes Daxenberger, and Iryna Gurevych. “A survey of nlp methods and resources for analyzing the collaborative writing process in wikipedia”. In: *The Peopleâs Web Meets NLP*. Springer, 2013, pp. 121–160.

[9] Johan Bos and Katja Markert. “Combining shallow and deep NLP methods for recognizing textual entailment”. In: *Proc. of the PASCAL RTE Challenge*. 2005.

[10] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv: `1301.3781 [cs.CL]`.

[11] Uday Kamath, John Liu, and James Whitaker. *Deep learning for NLP and speech recognition*. Vol. 84. Springer, 2019.

[12] Marta Sabou et al. “Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines.” In: *LREC*. Citeseer. 2014, pp. 859–866.

[13] Pontus Stenetorp et al. “BRAT: a web-based tool for NLP-assisted text annotation”. In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 2012, pp. 102–107.

[14] Jinhyuk Lee et al. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. In: *Bioinformatics* (Sept. 2019). btz682. ISSN: 1367-4803. DOI: `10.1093/bioinformatics/btz682`. eprint: `http://oup.prod.sis.lan/bioinformatics/advance-article-pdf/doi/10.1093/bioinformatics/btz682/30132027/btz682.pdf`. URL: `https://doi.org/10.1093/bioinformatics/btz682`.

[15] Yukun Chen et al. “A study of active learning methods for named entity recognition in clinical text”. In: *Journal of biomedical informatics* 58 (2015), pp. 11–18.

[16] Zhenzhong Lan et al. “Albert: A lite bert for self-supervised learning of language representations”. In: *arXiv preprint arXiv:1909.11942* (2019).

[17] John Hornberger. "Electronic Health Records: A Guide for Clinicians and Administrators". In: *JAMA* 301 (Jan. 2009), pp. 110–110. DOI: `10.1001/jama.2008.910`.

[18] Taxiarchis Botsis et al. "Secondary use of EHR: data quality issues and informatics opportunities". In: *Summit on Translational Bioinformatics* 2010 (2010), p. 1.

[19] Aleksandar Kovačević et al. "Mining methodologies from NLP publications: A case study in automatic terminology recognition". In: *Computer Speech & Language* 26.2 (2012), pp. 105–126.

[20] Michael J Paul et al. "Social media mining for public health monitoring and surveillance". In: *Biocomputing 2016: Proceedings of the Pacific symposium.* World Scientific. 2016, pp. 468–479.