AI

MDPI

Article

Public Awareness and Sentiment Analysis of COVID-Related Discussions Using BERT-Based Infoveillance

Tianyi Xie [1], Yaorong Ge [1], Qian Xu [2] and Shi Chen [3,*]

[1] Department of Software and Information Systems, University of North Carolina at Charlotte, Charlotte, NC 28223, USA
[2] School of Communications, Elon University, Elon, NC 27244, USA
[3] Department of Public Health Sciences, University of North Carolina at Charlotte, Charlotte, NC 28223, USA
* Correspondence: schen56@uncc.edu

Abstract: Understanding different aspects of public concerns and sentiments during large health emergencies, such as the COVID-19 pandemic, is essential for public health agencies to develop effective communication strategies, deliver up-to-date and accurate health information, and mitigate potential impacts of emerging misinformation. Current infoveillance systems generally focus on discussion intensity (i.e., number of relevant posts) as an approximation of public awareness, while largely ignoring the rich and diverse information in texts with granular information of varying public concerns and sentiments. In this study, we address this grand challenge by developing a novel natural language processing (NLP) infoveillance workflow based on bidirectional encoder representation from transformers (BERT). We first used a smaller COVID-19 tweet sample to develop a content classification and sentiment analysis model using COVID-Twitter-BERT. The classification accuracy was between 0.77 and 0.88 across the five identified topics. In the sentiment analysis with a three-class classification task (positive/negative/neutral), BERT achieved decent accuracy, 0.7. We then applied the content topic and sentiment classifiers to a much larger dataset with more than 4 million tweets in a 15-month period. We specifically analyzed non-pharmaceutical intervention (NPI) and social issue content topics. There were significant differences in terms of public awareness and sentiment towards the overall COVID-19, NPI, and social issue content topics across time and space. In addition, key events were also identified to associate with abrupt sentiment changes towards NPIs and social issues. This novel NLP-based AI workflow can be readily adopted for real-time granular content topic and sentiment infoveillance beyond the health context.

Keywords: public awareness; sentiment analysis; social media analytics; infoveillance; natural language processing

Citation: Xie, T.; Ge, Y.; Xu, Q.; Chen, S. Public Awareness and Sentiment Analysis of COVID-Related Discussions Using BERT-Based Infoveillance. *AI* **2023**, *4*, 333–347. https://doi.org/10.3390/ai4010016

Academic Editor: Rüdiger Buchkremer

Received: 29 January 2023
Revised: 5 March 2023
Accepted: 13 March 2023
Published: 17 March 2023

1. Introduction

1.1. Background

Social media have become the major avenue for the public to receive health information from health agencies and news outlets and to share their own opinions on emerging health issues, especially during pandemics such as the 2009 H1N1 pandemic influenza, 2014 Ebola, 2015 Zika, and COVID-19. They have also become an important source for various health agencies and researchers to understand the public opinion and promote certain health campaigns. During the pandemic of 2014 Ebola, researchers noticed the significant upward trend of Twitter posts and Google search in the USA [1,2]. Moreover, during the 2016 Zika pandemic, multiple health agencies started to use social media as communication channels and adopted effective communication strategies to improve the dissemination of public health-related issues [3]. COVID-19 has become one of the most discussed topics on social media platforms across the globe.

Pandemics always involve issues beyond medical and health aspects alone. They are often associated with cultural, social, economic, and political issues [4,5]. In the early stage

of COVID-19, the majority of the discussions and debates on social media were about intervention policies such as quarantine and social distancing. As the pandemic progressed, the discussion shifted towards mask wearing; the government's handling of the crisis; and vaccine development, roll-out, and mandates. 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 from and share their opinions on social media platforms.

1.2. Relevant Work

Research on the monitoring and surveillance of social media discussions about health issues, commonly known as health infoveillance, started in 2000. Current infoveillance is achieved with the combination of natural language processing (NLP), time-series analysis, and geospatial analysis techniques. Various NLP applications, including topic modeling, topic classification, sentiment analysis, and semantic analysis, can give a comprehensive understanding of the topic, sentiment, and semantic of public opinion and sentiment regarding a health issue. Monitoring the trend of certain topics helps predict the outbreak and progress of an epidemic, such as influenza [7–13], Zika virus [14], and the recent COVID-19 [15,16]. More specific topics are of interest in infoveillance, especially during the COVID-19 pandemic. Non-pharmaceutical interventions (NPIs), including social distancing, stay-at-home orders, quarantine, and mask wearing, have been effective yet controversial ways to reduce airborne disease transmission [17–21].

Large pandemics, including COVID-19, have never been an isolated medical or health issue and are always associated with multiple aspects beyond health. Current NLP-based infoveillance can generate more comprehensive characterizations of the diverse topics and sentiments using textual data based on the rich linguistic, sentiment, and semantic features. There are word frequency-based NLP approaches, such as term frequency–inverse document frequency (TF-IDF) and latent Dirichlet allocation (LDA) [22]. Another approach is to apply the encoding of text with pre-trained embeddings, including Word2Vec [23], GloVe [24], and BERT [25]. The embeddings are then fed into certain machine learning or deep learning methods, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), for downstream tasks.

In this study, we focused on word- or sentence-level embedding to understand the contextual information of online discussions on COVID-19 on Twitter. Word embedding is the process of transforming textual words into numerical vectors. There are traditional static word embeddings, such as Word2Vec, FastText [26], and GloVe, where the embedding is trained based on a large cohort of texts. However, this kind of static embedding cannot effectively reveal the true meanings of the word 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 embeddings often perform not as well in certain specific contexts. As shown in this study, the language used in social media can be very different from the corpus upon which these text embeddings are trained; thus, this can result in low performance in the topic modeling tasks.

To address these problems, pre-trained embedding models such as BERT, ELMO [27], XLNet [28], and GPT-2 [29] have been developed to provide richer and more dynamic context-dependent information. BERT is one of these pre-trained embedding models for various NLP applications. BERT learns context from the input textual data with its initial embedding and positional information. Most importantly, BERT is able to infer a word's distinct meanings in different contexts by providing unequal vector representations, which static embeddings are not capable of achieving. BERT makes it possible to pre-train the model on the specific domain, such as health, using transfer learning techniques. Transfer learning usually ensures a better representation of the specific domain that the model is fine-tuned upon and leads to better performance in downstream tasks. Regarding medical and health domains, BioBERT [30], BlueBERT [31], and Med-BERT [32] are a few examples that have been pre-trained on biomedical publications and electronic health records. Regarding social media applications, examples include BERTweet and the more specific COVID-

Twitter-BERT [6], trained on COVID-related tweets. These more specifically pre-trained BERT variants show substantial performance improvements over the original BERT model. In addition to token-level embedding, there have been semantic embeddings for sentences, such as SentenceBERT [33].

2. Method

2.1. Data Source and Sampling

Twitter is one of the most popular social media platforms for online discussions about COVID-19. In this study, we used Twitter samples to analyze the trend and sentiment of COVID-related topics in the USA.

First, we used a relatively small tweet sample to develop the topic classification model. We randomly sampled 2000 tweets from 2020 using the keywords listed in Table 1. A filter was applied during the sampling process to ensure that the tweets had a geolocation tag in the USA. For this task, only English tweets were collected. In addition, we also excluded tweets that had fewer than 10 tokens for better semantic meaning and more accurate BERT classification. We also ensured that each user could only be sampled once. This criterion avoided the potential sampling bias of a few active users or bots who excessively tweeted about COVID-19. The key terms for sampling are provided in Table 1. Note that certain terms were discriminatory (e.g., China virus). However, we still included these inappropriate terms to increase sampling coverage for research purposes.

Table 1. Key terms for COVID-19-related tweet extraction.

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

Based on the sampled tweets, our team with a domain expert in COVID-19 developed the codebook in Table 2. After high inter-coder reliability was established, the final codebook covered 5 major topic categories, and a single tweet could belong to multiple topics and multiple sub-topics. Each sampled tweet was annotated by at least two annotators, and if discrepancies occurred, the tweet was then sent to the domain expert for the final determination of the topic category.

Table 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, and health personnel
Countermeasures	Masks, other PPE, disinfection, food, exposure, contact tracing, technology, research, and online resource
Policies and 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, and 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, and ecosystem
Social problems	Disc. country, disc. region, disc. ethnicity, disc. profession, disc. gender, disc. age, disc. religion, disc. food, violence, profanity, and misinformation

For analyzing the trends of topics and sentiments of COVID-related tweets, we used a larger dataset than the previous dataset for topic identification. We randomly collected 12,000 English tweets from 1-March-2020 to 31-May-2021 with COVID-19-related terms using Twitter's Academic API V2. In total, 6000 of the 120,000 daily tweets were geo-tagged, with their geolocation being in the USA The remaining 6000 daily tweets were without geo-tags for comparison.

2.2. Preprocessing

Prior to training the BERT topic classification model, each tweet went through a series of preprocessing steps. User names and URLs in the tweet text were replaced with a common text token. We also replaced all emojis or emoticons with textual representations using the Python emoji library. The title of the URLs and hashtags were preserved as additional features in addition to the tweet text. Each tweet was treated as a text input and then fed into the BERT model. The 280-character limitation of a tweet was within the longest sequence input limitation of the BERT model.

2.3. Text Embedding

Text embedding was an essential part of BERT in this project to reflect the contextual, sentiment, and semantic features of the text. The accurate embedding of the text resulted in a better representation of the text and subsequently more accurate topic modeling. In order to further increase model performance and efficiency, we adopted COVID-Twitter-BERT, which was specifically pre-trained on COVID-19-related tweets and aligned with the tasks in this study. Our preliminary analysis showed that COVID-Twitter-BERT had substantial performance improvement over the generic BERT-Base model.

2.4. Topic Classification

Once the tweet was embedded, we then used the embedding to develop a multi-label (multinomial) machine learning classification model that was able to accurately identify the topics of each tweet. Since each tweet could have multiple topic labels out of a total of five possible topics, we further turned this multi-label classification task into 5 independent binary classification tasks. Five different binary classifiers were trained to identify the topic of each tweet. During the training stage, imbalanced issues were present, as the classifier used one class against the remaining four classes. The weight of each classifier was further fine-tuned to ensure that the classifiers were able to generate tweet topic labels that reflected the true percentage of tweets in the dataset.

The performance of topic classification based on the text embedding of the BERT model and traditional logistic regression was evaluated. In addition, we also compared the classification performance of the generic BERT-Base model against the specifically pre-trained COVID-Twitter-BERT.

2.5. Sentiment Analysis

After the content topics were identified, we further evaluated the sentiments of the tweets. Sentiment analyses based on VADER (Valence Aware Dictionary and sEntiment Reasoner) and BERT were performed. VADER is a lexicon- and rule-based sentiment analysis tool specifically tuned to sentiments expressed in social media. VADER not only identifies the binary positive or negative sentiment of a tweet but also quantifies the degree of the positive or negative sentiment of the post. Similar to the topic classification task, BERT was also used to train a sentiment classifier. In this study, BERT was applied to develop a 3-class sentiment classifier: positive, neutral, or negative sentiment of a tweet. In this study, the sentiment of a tweet was assumed to be mutually exclusive, that is, each tweet could only have one specific sentiment. This assumption could be relaxed in future studies.

2.6. Performance Evaluation

The performance of the classifiers was evaluated using the corresponding confusion matrix obtained by testing sets with four elements: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Classification performance metrics included accuracy ($ACC = \frac{TP+TN}{TP+TN+FP+FN}$), precision ($PPV = \frac{TP}{TP+FP}$), recall ($TPR = \frac{TP}{TP+FN}$), and $F_1 score = \frac{2TP}{2TP+FP+FN}$. High ACC, F_1, PPV, and TPR scores indicated robust model performance, indicating that the classification models were validated. These metrics also allowed us to compare different text embedding and classification models so that the most accurate and reliable models could be identified.

The complete analytical framework was written in Python 3.7 with necessary supporting NLP and machine learning libraries. The codes are freely available upon request.

3. Results

3.1. Topic Classification

We developed and compared the classification performance of the generic BERT-Base and COVID-Twitter-BERT models. Figure 1 shows that the optimal number of epochs to balance training loss and validation loss, as well as to reduce overfitting, was five.

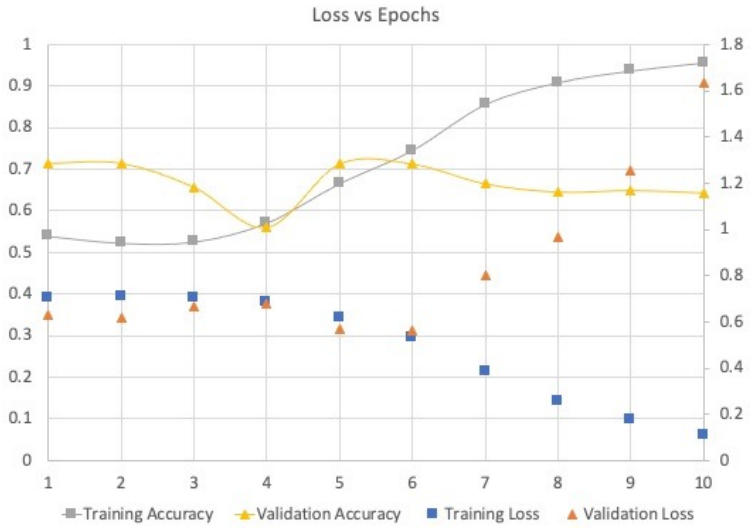

Figure 1. Training and validation loss vs. number of epochs.

The comparison among the different models showed that the deep learning-based BERT models significantly outperformed the traditional logistic regression models based on classification accuracy ($ACC = \frac{TP+TN}{TP+TN+FP+FN}$). In addition, COVID-Twitter-BERT also showed improved performance over the generic BERT-Base model. These results demonstrated the advantage of large-scale deep neural networks that are pre-trained on specific domain data (Table 3).

Table 3. Topic classification accuracy comparison.

Class	Logistic Regression	BERT-Base	CT-BERT
Clinical/epi	0.64	0.71	0.77
Countermeasures	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

In this study, we focused on two topics that were specifically related to the COVID-19 pandemic: confounded social issues and non-pharmaceutical interventions (NPIs). The NPI topic was the combination of certain sub-topics in the classes of countermeasures and

policies, and the related topics included masks, other PPE, disinfection, social distancing, stay-at-home, and shelter-in-place. The performance is shown in Tables 4 and 5. Overall, the two BERT classifications models for social issues and NPIs both showed excellent performance, with accuracy of over 87%, as well as high precision and recall.

Table 4. BERT topic classifier performance: social issues.

	Precision	Recall	F1-Score	Support
Negative	0.95	0.83	0.88	305
Positive	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

Table 5. BERT topic classifier performance: NPIs.

	Precision	Recall	F1-Score	Support
Negative	0.92	0.93	0.92	408
Positive	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

3.2. Sentiment Classification

Next, we investigated how BERT identified sentiments in COVID-19 discussions on Twitter. There were three classes: positive, neutral, and negative sentiments. For sentiment analysis, eight epochs were chosen instead of five as in the previous topic classification, because sentiments were more challenging to model and took more training to update the optimal model parameters. Practically, it was also more difficult to identify the sentiments of tweets, as online discussions could be frequently sarcastic or informal. We compared the sentiment analysis performance of the VADER and BERT models. Labels 0, 1, and 2 corresponded to negative, neutral, and positive sentiments.

The sentiment classification performance is shown in Tables 6 and 7. Overall, BERT was able to achieve the accuracy of 0.7 in the three-class sentiment classification task, significantly outperforming the previous benchmark method, VADER ($ACC < 0.6$). These results demonstrate the capability of NLP methods based on deep neural networks, especially transformers, which are able to further identify contexts in texts.

Table 6. Sentiment classifier using VADER.

	Precision	Recall	F1-Score	Support
Negative	0.53	0.40	0.45	147
Neutral	0.59	0.74	0.65	268
Positive	0.52	0.35	0.42	83
Accuracy			0.57	499
Macro avg.	0.55	0.49	0.51	499
Weighted avg.	0.56	0.57	0.55	499

Table 7. Sentiment classifier using BERT-Base.

	Precision	Recall	F1-Score	Support
Negative	0.69	0.67	0.68	147
Neutral	0.71	0.78	0.75	268
Positive	0.69	0.51	0.58	83
Accuracy			0.70	499
Macro avg.	0.70	0.65	0.67	499
Weighted avg.	0.70	0.70	0.70	499

3.3. Analysis of Topic Trends and Sentiments

Once accurate COVID-19 topic and sentiment classification models were developed using BERT, we further applied the topic and sentiment classifiers on a much larger scale, i.e., to a 4 million-tweet sample, to comprehensively understand the spatio-temporal variability of COVID-19 discussions on Twitter. The trend analysis data input was smoothed using a 7-day Gaussian smoother with standard deviation of 3.

3.3.1. Comparison between Geo-Tagged and Non-Geo-Tagged Tweets

A total of 6000 geo-tagged tweets per day and 6000 non-geo-tagged tweets were sampled and analyzed to evaluate differences in topic distributions and trends between the two groups.

Figures 2 and 3 show that the topics were very similar and highly correlated between the two groups. The Pearson correlation coefficients were 0.79 and 0.8 for the topics of NPIs and social issues, respectively, showing that the topic being discussed in geo-tagged tweets were highly correlated with tweets without geo-tags. We also found that the proportion of NPI topics was significantly higher in geo-tagged tweets than in non-geo-tagged tweets, indicating that users who shared their geo-tags were more engaged in discussing NPI-related issues. On the other hand, users without geo-tags showed more interests in social issue-related topics.

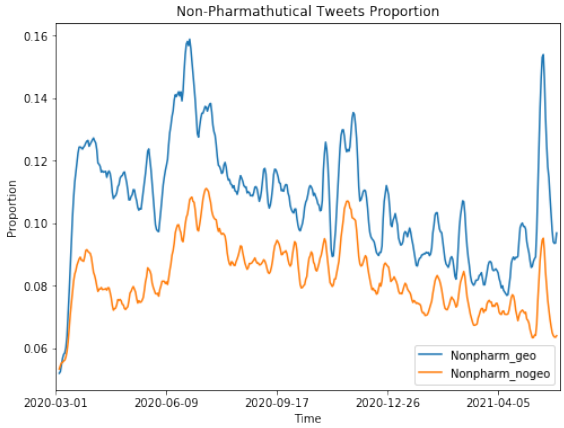

Figure 2. NPI topic proportions: geo-tagged vs. non-geo-tagged.

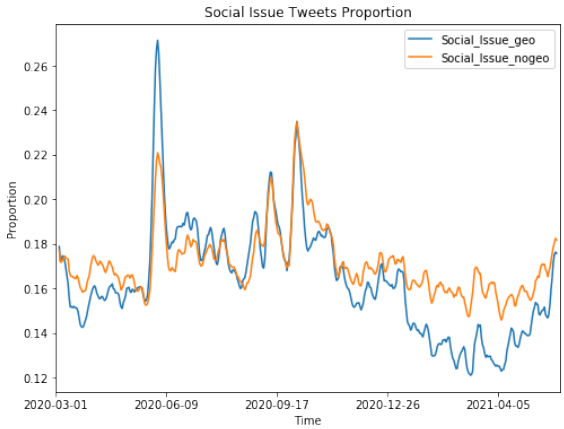

Figure 3. Social issue topic proportions: geo-tagged vs. non-geo-tagged.

We also compared sentiments in tweets with and without geo-tags. The overall sentiments were based on the arithmetic mean sentiment across all sampled tweets per day in the two groups. Overall sentiments ranged from -1 to 1, where 0 indicated a neutral sentiment. The sentiment trends in the two groups are shown in Figure 4.

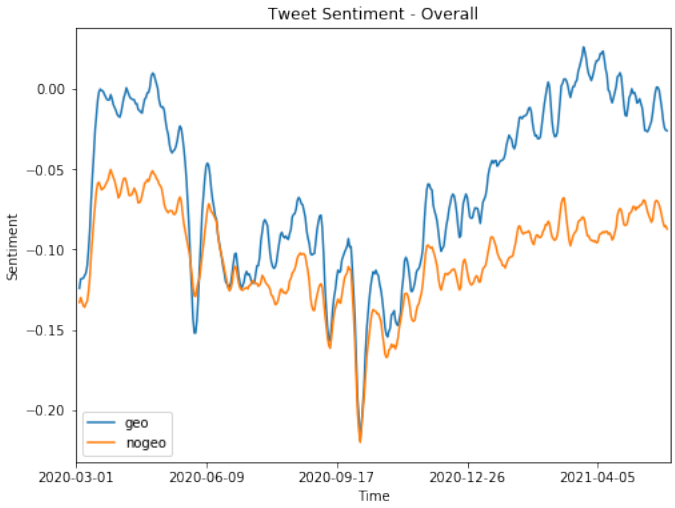

Figure 4. Overall daily sentiments.

Figure 4 shows substantial overall sentiment differences between tweets with geo-tags and tweets without geo-tags. Nevertheless, the Pearson correlation coefficient of 0.84 showed that the sentiments of the two sets were highly correlated. Overall, tweets with geo-tags had significantly higher sentiment scores (i.e., more positive sentiments) than tweets without geo-tags.

We further compared sentiments towards NPIs and social issues. Figure 5 shows sentiments towards the topic of NPIs. Tweets with geo-tags had more positive sentiments than tweets without geo-tags. There were several sudden changes in sentiment towards NPIs. Based on the time frame of these abrupt sentiment changes, we hypothesized that such changes were caused by the real-world events of former President Trump testing positive for COVID-19 and the CDC updating the guideline on mask mandates. The two topics were highly associated with NPIs, showing that our BERT sentiment classification model was able to successfully capture the changes.

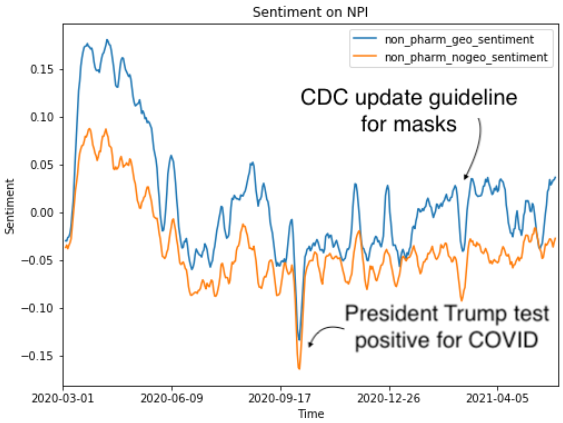

Figure 5. Sentiments towards NPIs.

Figure 6 shows that the overall sentiment towards social issues was more negative in tweets with geo-tags than in tweets without geo-tags. Compared with NPIs, the sentiment towards social issues was −0.42 (i.e., overall negative), while the sentiment towards NPIs was 0 (i.e., overall neutral). Therefore, overall public sentiments on social media significantly differed between the two topics. Similar to NPI sentiment changes, we were able to identify some key real-world events that caused the sudden changes in public sentiments towards social issues. Examples included the murder of George Floyd, former President Trump admitting downplay of COVID-19 threat, Trump being diagnosed with COVID-19, and the 2020 US election. Some of these events were not reflected in the sentiments towards

NPIs (e.g., murder of Floyd), indicating that the BERT model was capable of identifying and separating non-relevant tweets.

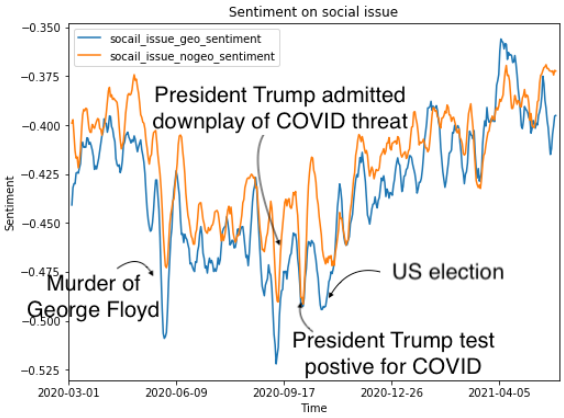

Figure 6. Sentiments towards social issues.

3.3.2. Comparison between Top 50 Cities and the Rest of the Country

In this section, we further present the comparison of content topic trends and sentiment trends between tweets geo-tagged in the top 50 most populous cities in the USA and the rest of the geo-tagged tweets. There were a total of 13,299 cities in the 2 million-geo-tagged-tweet sample, with the top 50 cities contributing 36.5% of the sample. The top 50 cities with their number of tweets are presented in Table 8.

Table 8. Top 50 cities with most tweets.

Rank	City	No. of Tweets	Rank	City	No. of Tweets	Rank	City	No. of Tweets
1	New York, NY	116,258	18	Seattle, WA	17,207	35	Sacramento, CA	5935
2	Los Angeles, CA	68,337	19	Denver, CO	10,406	36	Kansas City, MO	5571
3	Chicago, IL	35,804	20	Washington, DC	30,028	37	Mesa, AR	3061
4	Houston, TX	31,709	21	Nashville, TN	10,921	38	Atlanta, GA	14,805
5	Phoenix, AI	14,747	22	Oklahoma City, OK	5408	39	Omaha, NE	4320
6	Philadelphia, PA	21,751	23	El Paso, TX	4090	40	Colorado Springs, CO	2398
7	San Antonio, TX	14,478	24	Boston, MA	12,289	41	Raleigh, NC	5530
8	San Diego, CA	17,103	25	Portland, OR	11,166	42	Long Beach, CA	5198
9	Dallas, TX	14,937	26	Las Vegas, NV	8153	43	Virginia Beach, VA	3009
10	San Jose, CA	6367	27	Detroit, MI	4625	44	Miami, FL	6841
11	Austin, TX	17,881	28	Memphis, TN	5199	45	Oakland, CA	6938
12	Jacksonville, FL	10,850	29	Louisville, KS	3838	46	Minneapolis, MN	8573
13	Fort Worth, TX	5640	30	Baltimore, MD	8541	47	Tulsa, OK	3142
14	Columbus, OH	9977	31	Milwaukee, WI	4492	48	Bakersfield, CA	2182
15	Indianapolis, IN	8481	32	Albuquerque, NM	4340	49	Wichita, KS	2540
16	Charlotte, NC	10,394	33	Tucson, AR	4683	50	Arlington, TX	2427
17	San Francisco, CA	20,660	34	Fresno, CA	3583			

First, we compared the proportions of NPI and social issue topics in the top 50 cities and the rest of the country. Figure 7 shows that the proportion of NPI topics was around 11% of the overall tweets. At the beginning of the pandemic (April 2020 to August 2020) people who lived in the top 50 most populous cities were more likely to discuss NPIs on social media than people from less populous areas. We also observed a convergence in NPI discussions between populous, large metropolitan areas and less populous regions after September 2020. This matched the trajectory of the COVID-19 pandemic in the USA, as major metropolitan areas were impacted the most at the beginning; thus, people in these populous regions were more concerned about NPIs and engaged in NPI-related topics on social media.

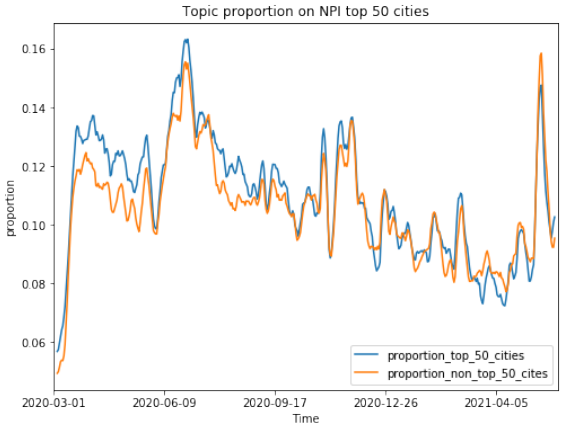

Figure 7. NPI topic proportions: top 50 cities vs. the rest of the country.

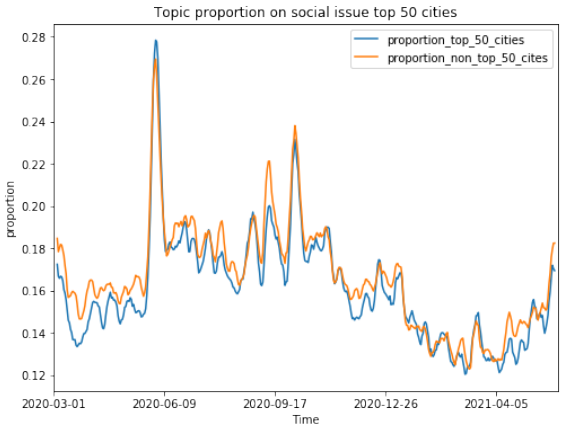

Figure 8. Social issue topic proportions: top 50 cities vs. the rest of the country.

Regarding the social issue topic, as Figure 8 shows, the proportion was generally around 16% of the overall tweets. Topics on social issues could abruptly arise when some real-world events happened, as we discuss above. In contrast to NPI topics, users in the top 50 most populous cities showed lower interest in social issues during the pandemic than the rest of the country. Nevertheless, people in large metropolitan regions discussed social issues more than other regions around late May 2020, when George Floyd was murdered.

We compared overall sentiments between the tweets generated in the top 50 most populous cites and tweets from the rest of the country. As Figure 9 shows, there was a clear and consistent difference throughout the study period, as users from the top 50 cities generally expressed more positivity than users from the rest of the country. While the overall sentiments between the two groups were highly correlated, with a Pearson correlation coefficient of 0.92, there was a substantial 0.03 sentiment difference. Tweets sent from the top 50 cities were generally 22% more positive regarding the pandemic.

We also compared the sentiments specifically regarding NPIs and social issues between the two regions. As Figures 10 and 11 show, the sentiments of tweets from the top 50 cities were more positive towards NPIs than those of tweets from the rest of the country. We also observed a substantial drop in sentiments towards NPIs around September 2020, which was probably due to the unclear messages that the CDC sent regarding mask mandates. The public then began to show negative sentiments towards NPIs.

Regarding social issues, the sentiments of tweets from the top 50 cities were consistently more positive than those of tweets from the rest of the USA. However, the Pearson correlation coefficient was only 0.51 for sentiments towards social issues. On the other hand, the Pearson correlation coefficient was 0.72 for the comparison of NPI sentiments between the top cities and the rest of the USA.

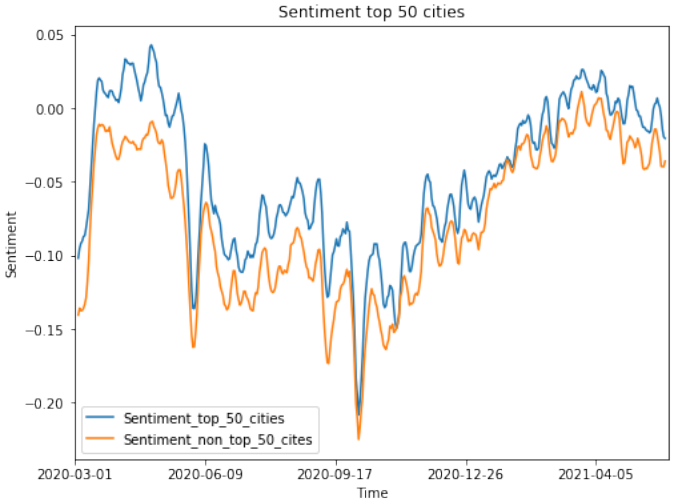

Figure 9. Sentiments: top 50 cities vs. the rest of the country.

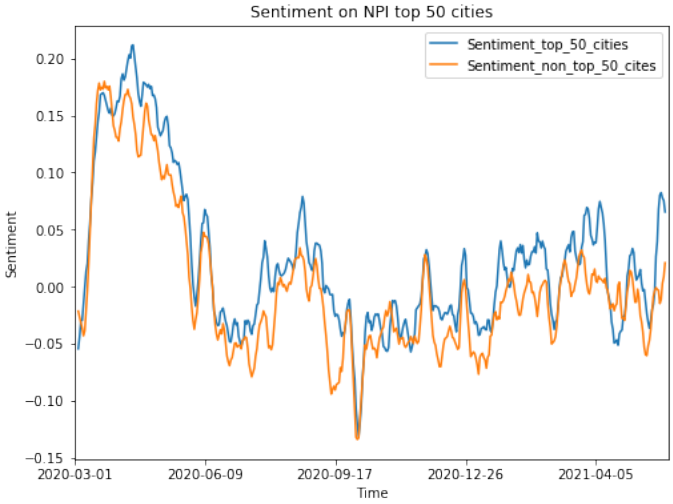

Figure 10. Sentiments towards NPIs: top 50 cities vs. the rest of the country.

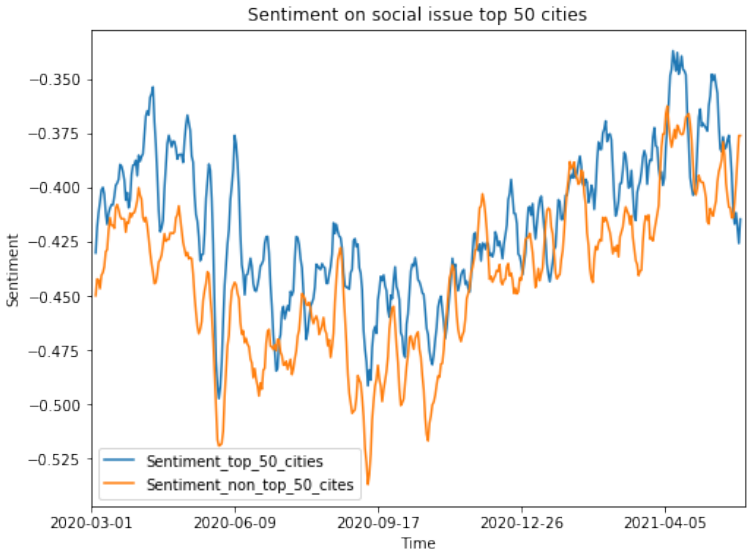

Figure 11. Sentiments towards social issues: top 50 cities vs. the rest of the country.

4. Discussion

In this study, we developed an innovative BERT-based NLP workflow for effective content topic and sentiment infoveillance during the COVID-19 pandemic. We first developed a content topic classifier and a sentiment classifier based on a smaller sample of COVID-19-related tweets using the COVID-twitter-BERT variant. We compared the performance of the baseline BERT models and the more specifically tuned COVID-Twitter-BERT models. The COVID-Twitter-BERT models demonstrated higher performance in classifying content topics and sentiments than the baseline BERT-Base models and significantly outperformed non-deep learning logistic regression models.

We then applied the developed BERT topic classification and sentiment classification models to more than 4 million COVID-19-related English tweets over 15 months. We were able to characterize the overall temporal dynamics of COVID-19 discussions on Twitter, as well as the temporal dynamics of more specific content topics and sentiments. Using the NPI and social issue topics as examples, we were able to accurately characterize the dynamic changes in public awareness of these topics over time, as well as sentiment shifts during different stages of the pandemic. In general, we found that the public had an overall neutral sentiment towards NPIs, but an overall negative sentiment towards various social issues. Compared with many infoveillance studies during the COVID-19 pandemic, our study is one of the few that utilized advanced AI NLP techniques to identify the real-time content topics and sentiments of online discussions from massive social media data. In addition, we also developed a highly effective BERT-based content and sentiment classification model for health-related discussions.

Our granular-level intelligent infoveillance is based on the deep learning NLP technique BERT. It enables public health practitioners to perform scalable infoveillance to zoom in and zoom out of an issue of interest (e.g., the overall COVID-19 pandemic) and understand various content topics associated with the issue (e.g., different aspects of COVID-19, such as clinical/epidemiological information of the disease itself, NPIs, vaccination, policies and politics, social issues, etc.). By understanding how public awareness and sentiment vary across time and space during different stages of the pandemic, public health practitioners can develop more effective and targeted health communication strategies and better address public concerns towards specific content topics, such as vaccination, NPIs, and social issues, including health disparity and inequality during the pandemic and other health emergencies.

5. Future Work

An extension of this study using the current BERT-based NLP infoveillance workflow is to quantify the spatio-temporal variability of public sentiment towards vaccination, one of the most discussed topics during the COVID-19 pandemic in the USA and across the globe. We demonstrated that our infoveillance workflow could successfully monitor public awareness and sentiment towards NPIs. Similarly, public perception towards vaccination could also be explicitly evaluated. Similar to our study on NPIs, public health practitioners could quickly respond to abrupt drops in sentiment towards vaccination and effectively identify potential external influencing events to develop countermeasures.

Our infoveillance workflow is also spatially explicit. We compared tweets generated in the top 50 most populous cities and in the rest of the cities in the USA. We observed a substantial sentiment gap between these major metropolitan areas and less populated regions. Social media users in major metropolitan areas expressed more positive sentiments towards the pandemic, NPIs, and social issues in their tweets. Future improvement in this workflow could incorporate more scalable geospatial information, such as identifying content topics and sentiments across geospatial scales, from the county level to the nation level. Public health practitioners could not only zoom in and understand more granular content topics but also zoom across geospatial scales to understand the spatial heterogeneity of content topics and sentiments. By incorporating more spatially explicit variables, for instance, various social and structural determinants of health (SDOHs), public health

practitioners could identify key influencing factors for certain content topics at granular spatial scales.

Our BERT infoveillance workflow is modularly designed and is able to be integrated with other analytical techniques, e.g., time-series analysis and signal processing, to detect certain key events during the pandemic that could have driven the abrupt changes in public sentiment towards NPIs and social issues. The future version of this infoveillance system is expected to automatically detect the key turning points of public perception towards a specific content topic and effectively identify potential external real-world drivers of the sudden sentiment changes.

The modular design of our BERT-based NLP infoveillance workflow can also be adapted for future applications such as misinformation detection. Using NLP and other analytical techniques, we could quickly find potential misinformation content topics and promptly respond to emerging misinformation topics. More granular characterization of online discussion reveals more specific contents and sentiments that are highly associated with misinformation, similar to the "digital antigen". Therefore, the infoveillance workflow is also able to actively send alarms to public health practitioners when certain key content topics of emerging misinformation match the "digital antigen" of misinformation.

Another extension of our infoveillance workflow is to further investigate content topic and sentiment shifts in social networks, using graph and network analysis. For instance, we could collect all replies to a specific original post, construct the network of information dissemination, and evaluate potential content and sentiment shifts from the original post in the network. We could then identify key vertices in the network that contribute to sentiment shifts, i.e., online influencers. Network metrics, such as various centrality scores, can be used to quantify the potential effectiveness of influencers in driving online discussions on social media.

In summary, we successfully developed a highly effective BERT-based infoveillance workflow for content and sentiment analysis. The workflow serves as a cornerstone for more extensive research and applications of large-scale social media analytics beyond the public health context.

Author Contributions: Conceptualization, Y.G., Q.X., S.C.; writing—original draft preparation, T.X.; writing—review and editing, S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research is partially supported by the Models of Infectious Disease Agents Study (MIDAS) Network through NIH/NIGMS (award number MIDASUP05).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Twitter data used in this study are freely available upon request to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fung, I.C.; Tse, Z.T.; Cheung, C.N.; Miu, A.S.; Fu, K.W. Ebola and the social media. *Lancet* **2014**, *384*, 2207. [CrossRef]
2. Hossain, L.; Kam, D.; Kong, F.; Wigand, R.T.; Bossomaier, T. Social media in Ebola outbreak. *Epidemiol. Infect.* **2016**, *144*, 2136–2143. [CrossRef] [PubMed]
3. Gui, X.; Wang, Y.; Kou, Y.; Reynolds, T.L.; Chen, Y.; Mei, Q.; Zheng, K. Understanding the Patterns of Health Information Dissemination on Social Media during the Zika Outbreak. *AMIA Annu. Symp. Proc.* **2017**, *2017*, 820–829. [PubMed]
4. Karabag, S.F. *An Unprecedented Global Crisis! The Global, Regional, National, Political, Economic and Commercial Impact of the Coronavirus Pandemic*; Linkoping University: Linkoping, Sweden, 2020; pp. 1–6.
5. Dignum, F.; Dignum, V.; Davidsson, P.; Ghorbani, A.; Hurk, M.; Jensen, M.; Kammler, C.; Lorig, F.; Ludescher, L.; Melchior, A.; et al. Analysing the Combined Health, Social and Economic Impacts of the Corovanvirus Pandemic Using Agent-Based Social Simulation. *Minds Mach.* **2020**, *30*, 177–194. [CrossRef]

6. Müller, M.; Salathé, M.; Kummervold, P.E. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. *arXiv* **2020**, arXiv:2005.07503.
7. Culotta, A. Towards detecting Influenza Epidemics by Analyzing Twitter Messages. *arXiv* **2010**, arXiv:1007.4748.
8. Yang, Y.T.; Horneffer, M.; DiLisio, N. Mining social media and web searches for disease detection. *J. Public Health Res.* **2013**, *2*, 17–21. [CrossRef]
9. Schmidt, C.W. Trending now: Using social media to predict and track disease outbreaks. *Environ. Health Perspect.* **2012**, *120*, A30–A33. [CrossRef]
10. Corley, C.D.; Cook, D.J.; Mikler, A.R.; Singh, K.P. Text and structural data mining of influenza mentions in Web and social media. *Int. J. Environ. Res. Public Health* **2010**, *7*, 596–615. [CrossRef]
11. Broniatowski, D.A.; Paul, M.J.; Dredze, M. National and local influenza surveillance through Twitter: An analysis of the 2012–2013 influenza epidemic. *PLoS ONE* **2013**, *8*, e83672. [CrossRef]
12. Aslam, A.A.; Tsou, M.H.; Spitzberg, B.H.; An, L.; Gawron, J.M.; Gupta, D.K.; Peddecord, K.M.; Nagel, A.C.; Allen, C.; Yang, J.A.; et al. The reliability of tweets as a supplementary method of seasonal influenza surveillance. *J. Med. Internet Res.* **2014**, *16*, e250. [CrossRef]
13. Aramaki, E.; Maskawa, S.; Morita, M. *Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter*; Association for Computational Linguistics: Edinburgh, UK, 2011; pp. 1568–1576.
14. McGough, S.F.; Brownstein, J.S.; Hawkins, J.B.; Santillana, M. Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data. *PLoS Negl. Trop. Dis.* **2017**, *11*, e0005295. [CrossRef] [PubMed]
15. Lwin, M.O.; Lu, J.; Sheldenkar, A.; Schulz, P.J.; Shin, W.; Gupta, R.; Yang, Y. Global Sentiments Surrounding the COVID-19 Pandemic on Twitter: Analysis of Twitter Trends. *JMIR Public Health Surveill.* **2020**, *6*, e19447. [CrossRef] [PubMed]
16. Abd-Alrazaq, A.; Alhuwail, D.; Househ, M.; Hamdi, M.; Shah, Z. Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study. *J. Med. Internet Res.* **2020**, *22*, e19016. [CrossRef]
17. Cowling, B.J.; Ali, S.T.; Ng, T.W.Y.; Tsang, T.K.; Li, J.C.M.; Fong, M.W.; Liao, Q.; Kwan, M.Y.; Lee, S.L.; Chiu, S.S.; et al. Impact assessment of non-pharmaceutical interventions against coronavirus disease 2019 and influenza in Hong Kong: An observational study. *Lancet Public Health* **2020**, *5*, e279–e288. [CrossRef]
18. Lai, S.; Ruktanonchai, N.W.; Zhou, L.; Prosper, O.; Luo, W.; Floyd, J.R.; Wesolowski, A.; Santillana, M.; Zhang, C.; Du, X.; et al. Effect of non-pharmaceutical interventions to contain COVID-19 in China. *Nature* **2020**, *585*, 410–413. [CrossRef]
19. Eikenberry, S.E.; Mancuso, M.; Iboi, E.; Phan, T.; Eikenberry, K.; Kuang, Y.; Kostelich, E.; Gumel, A.B. To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. *Infect. Dis. Model* **2020**, *5*, 293–308. [CrossRef] [PubMed]
20. He, L.; He, C.; Reynolds, T.L.; Bai, Q.; Huang, Y.; Li, C.; Zheng, K.; Chen, Y. Why do people oppose mask wearing? A comprehensive analysis of U.S. tweets during the COVID-19 pandemic. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 1564–1573. [CrossRef]
21. Sanders, A.; White, R.; Severson, L.; Ma, R.; McQueen, R.; Paulo, H.; Zhang, Y.; Erickson, J.; Bennett, K. Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse. *AMIA Jt. Summits Transl. Sci. Proc.* **2020**, *2021*, 555–564. [CrossRef]
22. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
23. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
24. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 26–28 October 2014; pp. 1532–1543.
25. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
26. Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; Mikolov, T. FastText.zip: Compressing text classification models. *arXiv* **2016**, arXiv:1612.03651.
27. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
28. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv* **2020**, arXiv:1906.08237.
29. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* **2019**, *1*, 9.
30. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2019**, *36*, 1234–1240. [CrossRef] [PubMed]
31. Peng, Y.; Yan, S.; Lu, Z. 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), Florence, Italy, 1 August 2019; pp. 58–65.

32. Rasmy, L.; Xiang, Y.; Xie, Z.; Tao, C.; Zhi, D. Med-BERT: Pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.* **2020**, *4*, 86. [CrossRef]
33. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv* **2019**, arXiv:1908.10084.