



Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.



ELSEVIER

Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

An easy numeric data augmentation method for early-stage COVID-19 tweets exploration of participatory dynamics of public attention and news coverage

Yuan Chen^{*}, Zhisheng Zhang^{*}

School of Mechanical Engineering, Southeast University, Nanjing, China

ARTICLE INFO

Keywords:

Social media analysis
Data augmentation
COVID-19 outbreak
Public engagement
News coverage
Text classification

ABSTRACT

With the onset of COVID-19, the pandemic has aroused huge discussions on social media like Twitter, followed by many social media analyses concerning it. Despite such an abundance of studies, however, little work has been done on reactions from the public and officials on social networks and their associations, especially during the early outbreak stage. In this paper, a total of 9,259,861 COVID-19-related English tweets published from 31 December 2019 to 11 March 2020 are accumulated for exploring the participatory dynamics of public attention and news coverage during the early stage of the pandemic. An easy numeric data augmentation (ENDA) technique is proposed for generating new samples while preserving label validity. It attains superior performance on text classification tasks with deep models (BERT) than an easier data augmentation method. To demonstrate the efficacy of ENDA further, experiments and ablation studies have also been implemented on other benchmark datasets. The classification results of COVID-19 tweets show tweets peaks triggered by momentous events and a strong positive correlation between the daily number of personal narratives and news reports. We argue that there were three periods divided by the turning points on January 20 and February 23 and the low level of news coverage suggests the missed windows for government response in early January and February. Our study not only contributes to a deeper understanding of the dynamic patterns and relationships of public attention and news coverage on social media during the pandemic but also sheds light on early emergency management and government response on social media during global health crises.

1. Introduction

1.1. The COVID-19 pandemic

For the past two years, the COVID-19 pandemic has swept around the world and has had dire consequences for public health and the global economy, causing huge economic losses, and physical and mental traumas for people worldwide. Several cases of unknown respiratory illness were first found in December 2019 in Wuhan, Hubei Province, China. On 20 January 2020, Chinese authorities identified the new coronavirus and its human-to-human transmission. The virus has quickly spread over China and international cases

^{*} Corresponding authors.

E-mail addresses: 230159322@seu.edu.cn (Y. Chen), oldbc@seu.edu.cn (Z. Zhang).

<https://doi.org/10.1016/j.ipm.2022.103073>

Received 21 May 2022; Received in revised form 21 August 2022; Accepted 25 August 2022

Available online 29 August 2022

0306-4573/© 2022 Elsevier Ltd. All rights reserved.

have also been identified in many other countries and areas including Japan, France, Germany, India, the United States, Africa, etc. On January 30, after a third meeting, World Health Organization (WHO) declared the outbreak as a Public Health Emergency of International Concern (PHEIC) which is the sixth time that the PHEIC measure has been invoked since the H1N1 pandemic in 2009. The coronavirus was later officially named COVID-19 on February 11 by WHO. Concerned with the alarming levels of spread and severity (more than 118,000 confirmed cases in 114 countries and regions, and 4,291 deaths), WHO characterized the outbreak of COVID-19 as a pandemic on March 11 with the virus spreading worldwide increasingly. Till now, even with the advent of the vaccine, there remains an unknown but tough battle for us as the virus continues to threaten public health and testify government policy.

1.2. Social media and its application in the COVID-19 pandemic

Nowadays, more and more people tend to retrieve information from social websites rather than traditional media like television, broadcast, newspapers, and so on. Millions of social network users express and communicate their ideas through social media. Previous studies have demonstrated the value of internet data in such a big data age that they can be used in various aspects including crisis response and management (Alkhodair et al., 2020; Imran et al., 2020; Jamali et al., 2019; Kaufhold et al., 2020; Li et al., 2020), commercial fields (Kozłowski et al., 2020; Seki et al., 2022) like consumer attitudes and behaviors (Chen & Zhang, 2022; Laguna et al., 2020), and political activities (Li et al., 2022; Mohammad et al., 2015; Stamatelatos et al., 2020). In the public health area, social media data from Twitter have shown their effectiveness in supporting monitoring public health and identifying specific populations. Health-related tweets could be analyzed for drug adverts reporting (De Rosa et al., 2021; Freifeld et al., 2014), emergency monitoring (Li et al., 2020; Şerban et al., 2019), vaccine attitude exploration (Griffith et al., 2021; Wilson & Wiysonge, 2020) and so on.

With the onset of the COVID-19 outbreak, social media has served as an important tool for information generation, dissemination, and consumption, contributing to many high-quality COVID-19 related papers involving main aspects like infodemic, surveillance, and monitoring (Shorten et al., 2021; Tsao et al., 2021). Infodemic papers are mainly about misinformation (Agle & Xiao, 2021; Ahmed et al., 2020; Alaa et al., 2020), its detection (Ayoub et al., 2021; Kouzy et al., 2020; Kumari et al., 2021), influence, exposure and spread (Burel et al., 2020, 2021; Hanyin et al., 2021; Obadimu et al., 2021; Tang et al., 2021). Surveillance and monitoring papers cover a wide range including assessing public sentiments (Basiri et al., 2021; Blanco & Lourenço, 2022; Xuehua Han et al., 2020) like vaccine attitudes (Aygun et al., 2021; Griffith et al., 2021; Lazarus et al., 2021), mental health (Behl et al., 2021; Guntuku et al., 2020), and detecting or predicting COVID-19 trends (Huang et al., 2021) and cases (Shen et al., 2020). Some surveillance and monitoring papers dig up information and assess public attitudes beyond the COVID-19 itself, by measuring the influence of the pandemic on various aspects, like environment concern and climate policy (Drews et al., 2022; Savin et al., 2022), food priorities (Laguna et al., 2020) and drinking behavior (Rodrigues et al., 2022).

But very little work has been done on how the public and officials react on social media platforms and if there is an association between public attention and news coverage from government and media. Amid the unprecedented health crisis, news coverage of COVID-19 on Twitter is of great significance to public health as previous findings suggest that during the pandemic, numerous citizens join Twitter to seek information on health matters (Haman, 2021). Understanding the flows of public engagement and news coverage in social networks in the context of COVID-19, an unprecedented global pandemic, would be of timely practical interest to health crisis management at present and in the future. However, a scarcity of studies addressing tweets of personal narratives and official news in the pandemic, especially during the early COVID-19 outbreak has been found.

Some papers involve government responses by digging up the participation of government or official accounts on social media platforms. Rufai & Bunce (2020) explore the role that world leaders from the Group of Seven (G7) play in response to COVID-19 by conducting the content analysis of their tweets qualitatively. Similarly, Wang et al. (2021) discover the discourses on the political leaders' Twitter accounts from "Five Eyes" nations¹ but use their tweets frequency and sentiment distribution using a lexicon-based algorithm with three traditional machine learning models: Bernoulli Naive Bayes model, linear Support Vector model, and logistic regression model. Luo et al. (2022) and Muqsith et al. (2021) focus on the tweets from former US President Donald Trump. Beyond the government leaders, Merkley et al. (2020) evaluate the response to COVID-19 from Canadian political elites and the mass public, focusing on differences between the elite and public on COVID-19 issues using social media accounts of Federal Members of Parliament, Google trends, and public opinion surveys. Wang et al. (2021) investigate the pandemic-related tweets from 67 federal and state-level agencies and stakeholders in the U.S. with text mining techniques and dynamic network analysis. There are also studies on other official accounts of public health authorities (Alhassan & Aldossary, 2021; Li et al., 2021; Raamkumar et al., 2020; Tang et al., 2021; Xi et al., 2022) and experts (Knox & Hara, 2021).

Despite all the inspiring work they have done, these papers all focus on a limited group of political and government actors by studying specific Twitter or Facebook accounts, and a broader scope of government and media response in comparison with public engagement quantitatively is absent. We decide to fill the gap and place a particular emphasis on how the public and officials participate in social networks and the association between their activities by focusing on the digital trace of Twitter data. All the early-stage tweets accumulated would be classified into two groups: personal narratives and news reports to discover the dynamic patterns of the public attention and official engagement from the government or media. The selected period starts from the very beginning: 31 December 2019 to the day that 'Pandemic' was announced by WHO: 11 March 2020 when a global alert was officially declared.

¹ The U.S., United Kingdom, Australia, New Zealand, and Canada.

1.3. Research objectives

Our paper sets out to contribute to this line of research for the first time: the early participatory dynamics of public and government on the social networks by analyzing COVID-19 related tweets in order to provide more insights on future government social media responses and policymaking in emergencies. In addition, we also introduce a novel easy numeric data augmentation (ENDA) technique for our text classification task as well as other NLP tasks. The ENDA method is applied to benchmark datasets: SST-2 and TREC datasets for text classification tasks and the ablation study is conducted. Our main contributions are as follows:

- A COVID-19-related dataset with over 9 million tweets during the early COVID-19 outbreak (31 December 2019-11 March 2020) is established and analyzed.
- A novel easy numeric data augmentation (ENDA) method is proposed for improving the performance of deep models on text classification task.
- Providing an overview of the participatory dynamics of public attention and government responses in the initial phase of the outbreak.
- Discovering the relationship between public engagement and news coverage through the daily number of personal narratives and news reports.

In the remaining sections, we present the literature overview of data augmentation methods in Section 2, introduce our dataset: the early-stage COVID-19 dataset in Section 3, illustrate the methodology including the proposed easy numeric data augmentation (ENDA) approach in Section 4, examine the label validity and conduct ablation study on benchmark datasets in Section 5, analyze the tweets classification result in Section 6, discuss the results, limitations, future directions, and theoretical and practical implications in Section 7. We end our paper by highlighting the main findings and conclusions in Section 8.

2. Data augmentation background

In the data mining field, data have always been the fundamental part. With the help of artificial learning, valuable information could be extracted and mined now in a much easier way. But neural networks especially deep learning methods often over fit and their high generalization performances require large amount of qualitative data which are scarcely to obtain in real cases. Hence, many researchers turn to data augmentation for improving model performance in natural language processing (Shleifer, 2019; Yu et al., 2018), image (Krizhevsky et al., 2017; Szegedy et al., 2015) and speech processing (Cui et al., 2015; Ko et al., 2015; Wang et al., 2021) field since qualitative labeled data are hard to get and time-consuming.

Data augmentation means methods applied for increasing the data amount by adding slightly modified copies of original data or newly created synthetic data from the original one. Generally speaking, there are three main ways for data augmentation: paraphrasing-based methods, noising-based methods, and sampling-based methods (Li et al., 2022). Paraphrasing-based methods produce data that convey similar information with the limited semantic difference from the original data, based on proper and re-strained changes to sentences (Madnani & Dorr, 2010). Sampling-based methods are based on the data distributions and sample novel points within them. On the contrary, the noising-based methods manage to add minor noises into original data to create new ones which could improve model robustness at the same time. Methods include swapping, deletion, insertion, the substitution of words, and a more difficult one: mix-up. Paraphrasing-based methods and sampling-based methods techniques are often based on deep learning models which is rather costly relative to the performance gain. The noising-based method is easy to use in most cases and could improve model robustness but the generated augmentation data are rather limited and have the potential risk of changing original syntax and semantics in the data.

Specifically, in real problems, frequently used methods include the paraphrasing-based methods: easy data augmentation (EDA) (Wei & Zou, 2019), back-translation; the noising-based methods: an easier data augmentation (AEDA) method (Karimi et al., 2021), noising; and the emerging one: Generative Adversarial Networks (GAN)-based. Among them, EDA and its variant AEDA are the easiest and universal methods in real applications. Easy data augmentation (EDA) demonstrates a simple but effective way of augmenting text data. It randomly chooses and performs one of the four following operations: synonym replacement, random insertion, random swap, and random deletion on the original data to create new text data for boosting performance on text classification. Synonym was proposed and used in previous papers (Zhang et al., 2015), but other techniques like word insertion, swap, and deletion are used for the first time. In EDA, α is a parameter for the percent of the changed words in a sentence and n_{aug} for the number of augmented sentences in each original sentence. Another effective method AEDA consists of inserting a certain number of punctuation marks from list {".", ";", "!", ":", "!", " ", " " } into original sentences and the number of insertion is determined between 1 and one-third of the length of the sequence randomly. Both EDA and AEDA validate their performance on deep learning models: RNN and CNN with Glove embedding and the results indicate accuracy improvement, especially for small-size datasets. But there is no systematic study on their performance on deep models like BERT (Devlin et al., 2019). Only Karimi et al. (2021) do a trial on BERT on two datasets for one augmentation, and the result shows that AEDA improves the model performance on SST-2 and TREC datasets while EDA decreases instead.

3. Early-stage COVID-19 (ESCD) dataset

With web crawl², we retrieved English tweets from 31 December 2019 to 11 March 2020 on Twitter by using search words. We use search terms like "China pneumonia", "Wuhan outbreak", "coronavirus", and "COVID-19" which were frequently used (see more details in supplementary material, Table S.1) to investigate online discourses on COVID-19. What's worth noticing is that some titles like "China virus" are not accurate and have been strongly rejected by WHO later but they were used for discussion and reports widely at the early beginning, so we keep them for preserving the data integrity. The crawled data are JSON files and contain information including text, tweet id, retweets, likes, etc. For every tweet, there is a unique tweet id. We apply data cleaning procedures including removing duplicate tweets and information missing or null tweets by identical tweet id. Finally, the total number of tweets is 9,259,861 spreading from the last day of 2019 to the day when the outbreak was characterized as a "pandemic". With over 9 million tweets in 72 days, this is by far the most complete COVID-19-related tweets dataset for providing insights for information discourses on Twitter during the early stage of the outbreak. This dataset contains the text, time, likes and retweets, and other metadata. Any identifiable information is dumped for personal privacy.

We select 9263 tweets out of the whole dataset for annotation randomly and use re (regular expression) module in python³ to remove website and picture links, blank lines, and any non-characters as preprocessing cleaning procedures. After annotating, the samples are divided into two categories: personal narratives and news reports with the numbers 5234 and 4029, respectively (We provide all the ids and labels of the sample in supplementary materials and GitHub <https://github.com/yuanchenroy/COVID-19-related>. All the tweets could be accessed by their ids through corresponding link <https://twitter.com/user/status/id>). Examples of annotation are shown in Table 1. We split the whole annotated dataset into training, validation, and test dataset with a ratio of 3:1:1 randomly.

4. Methodology

In this section, we demonstrate the proposed easy numeric data augmentation method for text classification and the deep model we use for experiments. The ablation study is conducted to investigate the performance gain for different augmentation numbers and dataset sizes with the deep model.

4.1. Easy numeric data augmentation (ENDA)

As mentioned above, most data augmentation methods share a universal theoretical foundation: either adding slightly modified copies of original data or creating new synthetic data from the original one. Basically, the paraphrasing-based methods like back translation, and sampling-based methods like EDA create synthetic data while noising-based methods like AEDA choose to add noises to modify the original data. Existing noising-based data augmentation methods choose variants of words like synonyms or punctuations to create new sentences.

The ENDA is also motivated by the linguistic characteristic of text data and the requirement of label validness after augmentation. In data augmentation, original data are changed while class labels are maintained. But there are always concerns about the label validness after augmentation. Because if sentences are altered significantly, the original class labels may no longer be valid. Compared with the possible semantic changes brought by insertions of punctuations, we have assumption that our method could maintain label validness of the augmented text better. We presume punctuations could possibly damage the sentiment and meaning in the original sentences while numbers often will not. For example, in the case of our COVID-19 data set, personal narrative tweets could use more punctuations like '?', and '!' than news tweets statically. The descriptive statistics data of ESCD dataset are shown in Table 2. From the average and median value of numbers and punctuations per sample, it can be seen that punctuations occur much more frequently than numbers in all type of tweets. And the difference of punctuations between two categories is significant as revealed by the non-parametric Mann-Whitney U test ($p < 0.0001$) while the difference of numbers between two type of tweets is not significant ($p = 0.4949$). The results show that we may reject the null hypothesis and there is a significant and strong correlation between the occurrences of punctuations in tweets and the tweets type.

In this paper, we select numbers instead and propose an easy numeric data augmentation (ENDA) method to improve the performance of deep learning models for COVID-19 news classification. The ENDA method share the universal theoretical foundation that adding minor noises into original data to create new samples and to improve model robustness at the same time. The ENDA is quite simple without requiring any expert or extra knowledge and could be easily applied in many datasets. Similarly, we also perform a certain number of insertions but with numbers as noises instead. Specifically, we use a number list {"0", "1", "2", "3", "4", "5", "6", "7", "8", "9"}, insert n_{insert} numbers from number list into original sentence randomly, and repeat this process for n_{aug} times, where $n_{insert} \in \text{Range}(1, N = \alpha \ell)$ while ℓ denote the length of original sentence, α is the percentage parameter and n_{aug} is the parameter presents the number of augmented sentences per original sentence generates.

² <https://github.com/taspinar/twitterscraper>

³ The data cleaning code is available on GitHub <https://github.com/yuanchenroy/COVID-19-related>

Table 1
Annotated dataset of ESCD.

Classification Type	Number of tweets	Example Tweets
Personal Narratives	5234	<ol style="list-style-type: none"> 1 This is my biggest worry about getting potentially incapacitated by the coronavirus/flu/cold/etc., since anyone who cares for my cats would wind up being vulnerable to whatever virus I have. It's a bad situation that would just get worse, and there are not any easy answers. 2 #coronavirus has ruined football in a very short time than match fixing!! Scientists that studied on government money where u?
News and Reports	4029	<ol style="list-style-type: none"> 1 Italy puts 10 million on LOCKDOWN with even funerals banned to fight coronavirus spread http://shr.gs/Q2vyEnU BBC News - Coronavirus: Cases jump in Iran and Italy https://www.bbc.co.uk/news/world-middle-east-51783242

Table 2
Descriptive statistics data of ESCD dataset.

Tweets category	Occurrences of punctuations per sample				Occurrences of numbers per sample				
	average	Median	Max	Min	average	Median	Max	min	
Personal Narratives	5.84	5	46	0	1.43	0	57	0	
News and Reports	4.70	4	92	0	1.53	0	48	0	

4.2. Deep learning models for text classification and experiment set up

We choose Bidirectional Encoder Representation from Transformers (BERT) model as our experiment field for all the data augmentation techniques. Developed by Google, BERT is a transformer-based machine learning model pre-trained on the large-scale corpus for natural language processing (NLP). It has become a ubiquitous baseline in NLP experiments and has swept all other deep learning models like CNN, and RNN (Han et al., 2021). BERT is based on the stacked multi-layer bidirectional transformers with an encoder-decoder structure that applies a self-attention mechanism, which models the correlations between words in parallel. Compared with traditional word embedding methods like Glove or Word2Vec, BERT shows distinct advantages with its outstanding ability to grasp massive information such as shallow features, syntactic features, and semantic features from the text. For text classification tasks, the BERT model inserts a [CLS] token and a separator [SEP] token at the beginning and end of the input text. With sample size sequences of tokens with a maximum length as input, the output vector corresponding to the sequence is the semantic representation of the text for text classification. In this paper, we fine-tuned BERT-Base (English only, uncased, trained with WordPiece masking) (Devlin et al., 2019) which contains 12 transformer block layers, 768 hidden units, 12 heads, and 110 M parameters totally. We extract the output of the [CLS] token as the representation for sentence-level classification and feed it into the classifier of the Softmax layer to get to obtain the predictions of tweets type. We set the max length as 125 as it covers over 99% samples. The BERT model is trained on the training dataset with the early stopping of 3 epochs, the best-performing model on the validation data would be saved for predicting the test data to test its performance later. To decrease the effect of random initialization, we repeat the process with 6 different random seeds to get the average results.

We perform experiments using the data augmented by AEDA, ENDA, and the original data for evaluation. Furthermore, to investigate performance on the different dataset sizes, we delegate 5 different sizes of datasets by selecting a random subset of the full training dataset. The training data sizes include 500, 1000, 2000, 4000, and full size. We add 1, 2, 4, 8, and 16 augmentations for each sample for all experiments for ablation study on augmentation number.

We execute all the code using a Tesla V100 GPU with 16 GB memory. The Hyper-parameters obtained by fine-tuning are shown in Table 3. All codes are performed with Alibaba Cloud Elastic Compute Service (Tesla V100 GPU and 16G RAM) with Python 3.8.10 environment. The BERT model is implemented in Python using Keras⁸ with TensorFlow⁹ as backend.

4.3. ENDA outperforms AEDA

The experimental results with data sizes of 500, 1000, 2000, 4000 and full data augmentations of 1, 2, 4, 8, and 16 are plotted in Fig. 1. Both ENDA and AEDA work better with larger dataset sizes. For training dataset size of 1000, 2000, 4000, and full-size, both

Table 3
Hyper-parameters setting.

Parameters	Value
Learning rate	1e-5
Batch size	32
Optimizer	Adam

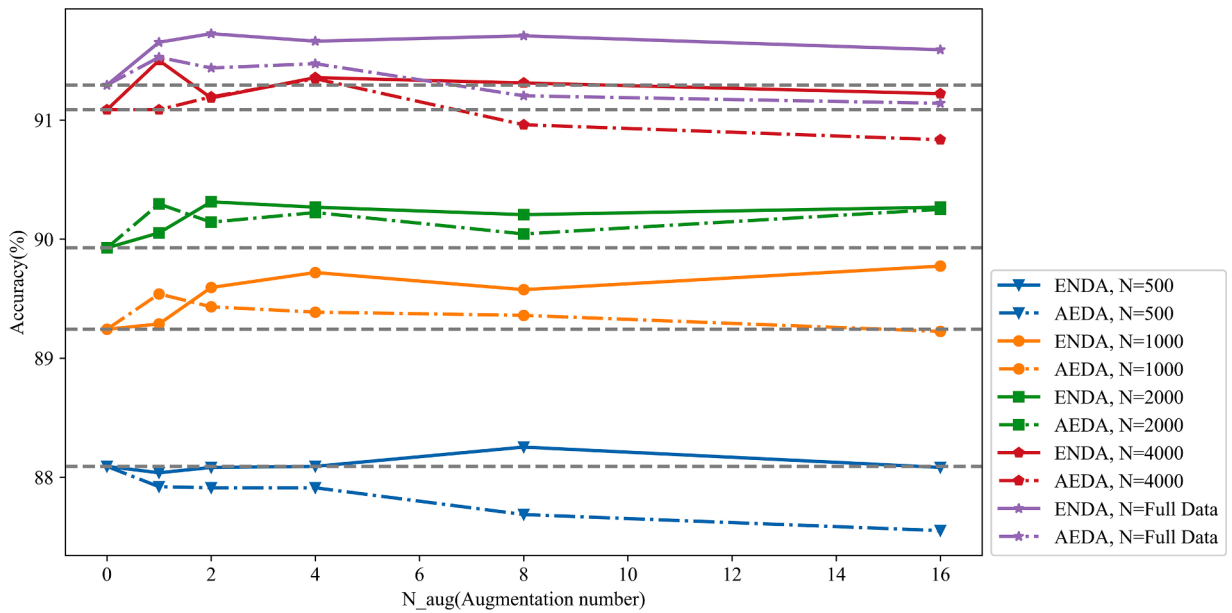


Fig. 1. Average Augmentation performance of AEDA and ENDA on sub sizes of early-stage COVID-19 dataset over 6 seeds, in line with original data.

ENDA and AEDA show improvements compared with the original data baseline. But ENDA gives a better booster than AEDA. For the size of 500, only ENDA increases the accuracy with the augmentation of 4 per sample. Instead of a booster, AEDA decreases the accuracy with all 5 augmentation numbers on the smallest dataset size. Overall, ENDA shows superior results compared with AEDA on all datasets. The results indicate that ENDA gives a better booster and outperforms AEDA for our tweets classification task of personal narratives and news reports.

Basically, both the improvements of AEDA and ENDA follow the trend of rising at first and then leveling or falling back with the increase of augmentation number. For most sizes, ENDA and AEDA arrive at high points with one or two augmentations except for the smallest dataset size where it takes ENDA 8 augmentations per sample to achieve an increase and continuous decrease for AEDA. Our recommendation on the ESCD dataset goes as follows: for datasets less than 1000, the recommended augmentation number is $n_{aug} = 8$;

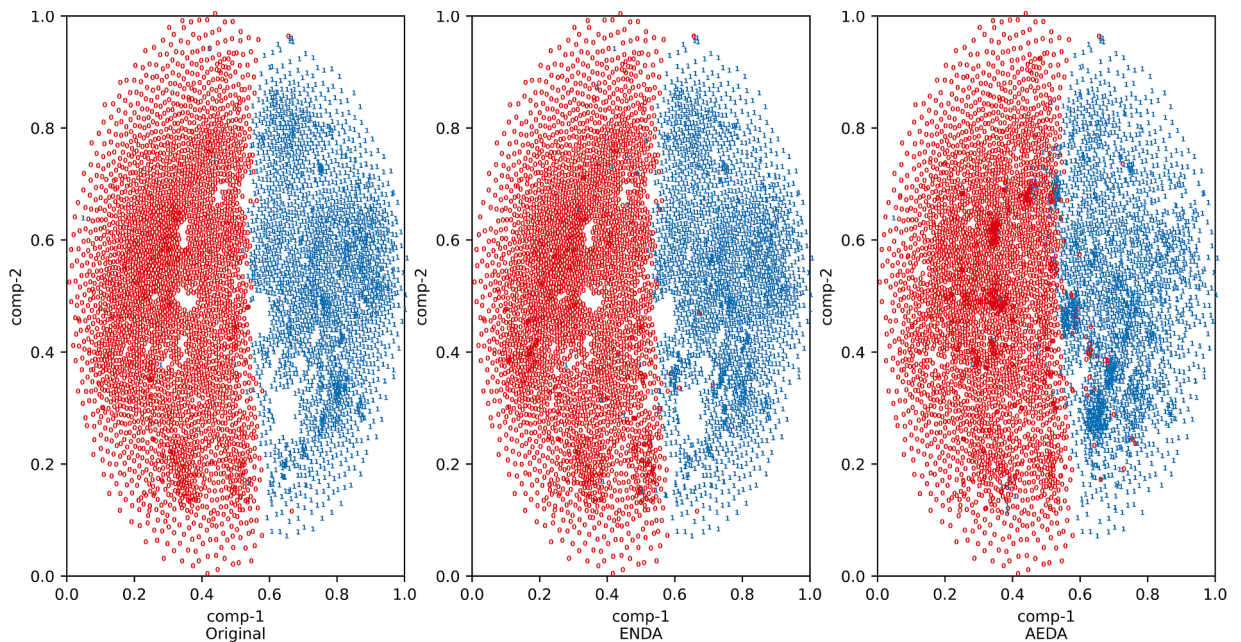


Fig. 2. Visualization of dimensionality reduction features of the original data, data augmented by ENDA, and AEDA (For interpretation of the references to color in this figure, the reader is referred to the web version of this article).

for datasets between 1000 and 4000, then $n_{aug} = 2$ to 4; for datasets more than 4000, then $n_{aug} = 1$ or 2.

5. Experiments

Results in Section 4 show that the proposed data augmentation method attains superior performance than AEDA, proving its efficiency in handling the classification problem of news and personal narrative tweets. To further demonstrate the efficacy of the proposed technique for generality, we have examined the label validity of augmented text and applied ENDA on two other benchmark datasets: sst-2 and TREC datasets.

5.1. Label validity

We reckon that ENDA could create new samples for training while preserving the original label of text after augmentation. We examine the label validity of the augmented text by extracting the high-dimensional layer output vectors from pre-trained models and visualized their latent space representations by t-sne (t-distributed stochastic neighbor embedding) approach.

We start with training a BERT-based deep model on the original training dataset of the ESCD, then we apply AEDA and ENDA to generate augmented text with one augmentation per sample. These augmented and original text are fed into the deep model we trained and the output of BERT-base: a 798-dimension layer is extracted and visualized by t-sne approach. These high-dimensional vectors are projected into a 2-d space and plotted in Fig. 2. We use scatters in red and blue to represent different labels. It can be seen that the ENDA dots surround the original dots more closely, following a similar pattern. With less overlap, the ENDA dots of different categories are also more easily separated than AEDA, suggesting that sentences augmented with ENDA could conserve the label of the original text better.

5.2. ENDA on SST-2 and TREC

We choose two benchmark datasets: sst-2: stanford sentiment treebank (Socher et al., 2013), and text retrieval conference (TREC) question classification (Li & Roth, 2002) to testify our proposed technique. We use the split training and test dataset from Karimi et al. (2021) (see details in supplementary material). Similarly, we repeat the experiments with the BERT models in the methodology part over 11 seeds to get the average results with 1, 2, 4, 8, and 16 augmentations per sample separately. We also use different fractions of the training data: .. For our experiments. Then the selected dataset would be split into train and validation datasets with a ratio of 9:1. The results of ENDA, AEDA, and original data with augmentation numbers of 1, 2, 4, 8, and 16 per sample under different fractions of

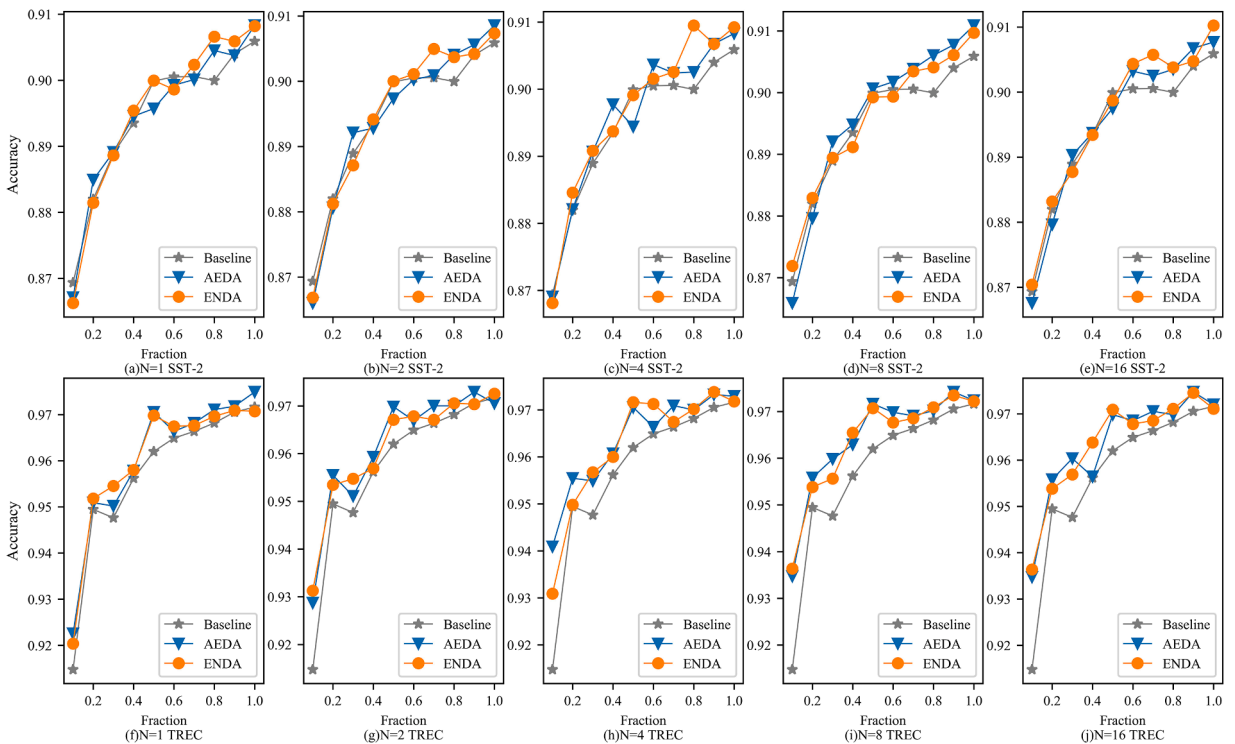


Fig. 3. Average augmentation performance of AEDA and ENDA over 11 seeds, in line with original data results over the same 11 random seeds on sst-2 and TREC dataset.

the training dataset for sst-2 and TREC are plotted in Fig. 3, respectively.

From Fig. 3, generally, both AEDA and ENDA could improve the performance of deep models as most points of ENDA, and AEDA are above those of the original data. But the dominant advantage of ENDA over AEDA is not that obvious compared to the results on the ESCD dataset. For sst-2, the improvements are much greater for large fractions while the small fractions of the TREC dataset have wider margins comparatively. Also, as the augmentation number increases, the performance improvement increases especially for lower fractions of the TREC training set. For sst-2 dataset, the improvement with the increase in augmentation number is more distinguished for larger fractions.

5.3. Ablation study on percentage parameter

Besides the fraction of the training dataset, the percentage parameter α for augmentation number is also worth exploring. In the above study, we simply follow the AEDA technique, setting the ratio as 0.3 by default. So in this part, we investigate the role of the percentage parameter α for augmentation number in ENDA by ablation study.

We conduct the experiments with different ratios $\alpha_{train} = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ and different fractions $Frac_{train} = \{0.1, 0.2, 0.4, 0.8, 1\}$. The augmentation numbers are set as 4. We use the average relative accuracy improvement $RelativeGain_{avg}$ on SST-2 and TREC for evaluation as shown in Eq. (1), where $Gain_{avg}, N_{acc}$ represents the accuracy obtained from the ENDA method, $Base_{acc}$ denotes the accuracy obtained with original data, and n is the runs of different seeds.

$$RelativeGain_{avg} = \frac{\sum_i^n (N_{acc} - Base_{acc})}{n} \tag{1}$$

The results are presented in Fig. 4.

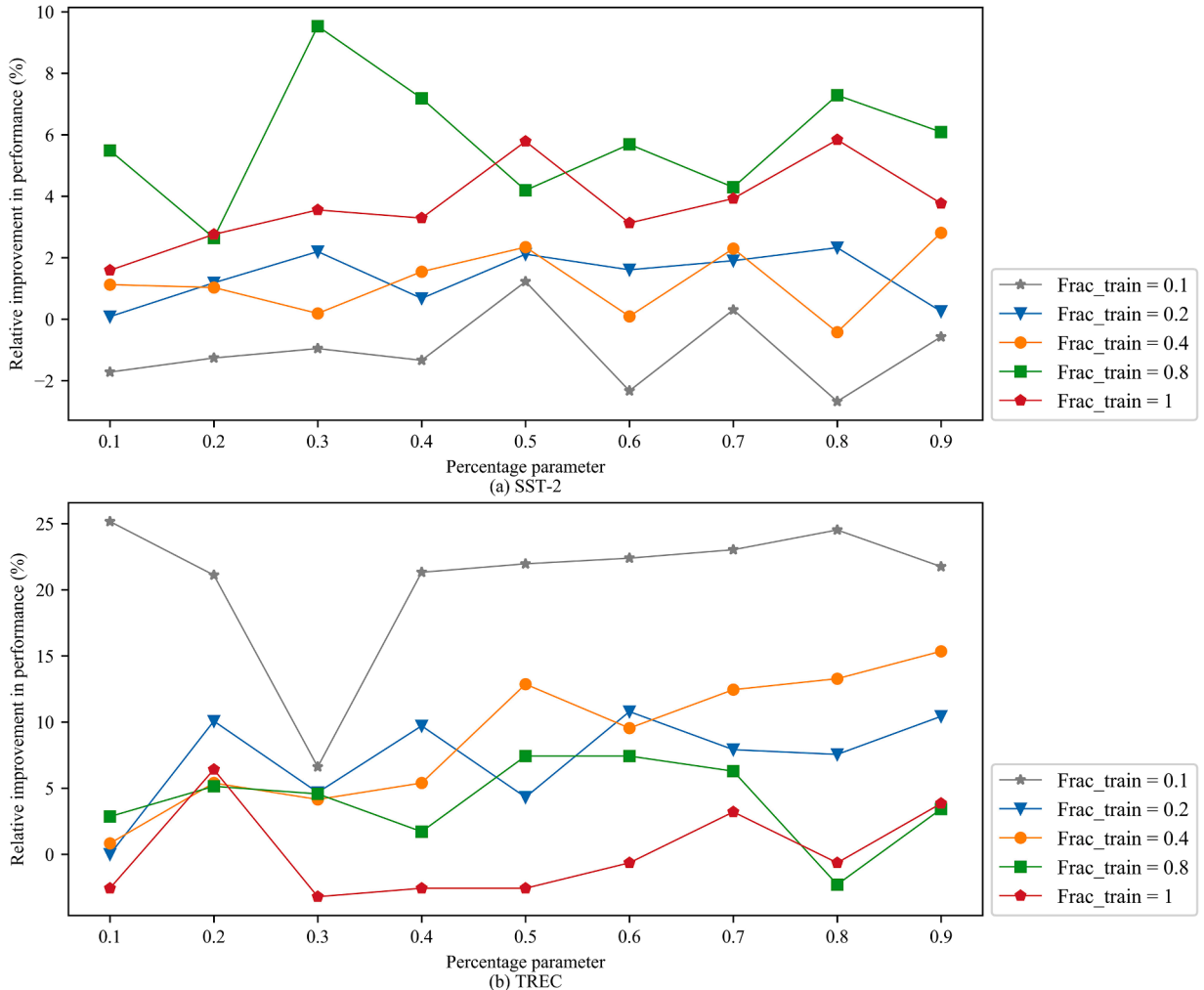


Fig. 4. Relative improvement in performance of ENDA over 11 seeds on SST-2 and TREC dataset with augmentations of 4.

Unlike augmentation number and training size, which impact the performance, the ratio is more domain-invariant as the fluctuations among different ratios suggest no regular tendency. It is hard to pick up the best value or range for percentage parameters under different fractions. The insertion times per sample is selected from zero to n_{aug} while the range of insertion number is determined by the length l and the percentage parameter α , $n_{aug} = \alpha l$. This indirect relation may weaken the impact of the parameter, resulting in limited and nonlinear sensitivity to percentage parameter change.

6. Results analysis

Based on the above study, we choose one augmentation with ENDA for the Early-stage COVID-19 dataset. We visualize the predicted results on the whole dataset to show the dynamic patterns of two general types of tweets over time. Fig. 5 presents the daily number of personal narratives, news reports, and the total tweets during the outbreak period.

As in Fig. 5, there had been a turning point for both news and personal narratives on 20 January 2020 when they started to grow significantly in accordance with the announcement of human transmission from the Chinese government on that day. Following this, both public attention and news coverage kept increasing and remained high ever since. In particular, peak levels of news reports and personal narratives appear around when WHO declared a PHEIC on January 30, when the first COVID-19-related death was reported in the U.S. on February 28, and when WHO announced a global pandemic on March 11.

Specifically, before January 20, the daily number of personal narratives was less than 500, hard to notice even in the zoomed view from Fig.5. More broadly, emerging on January 20, the personal narratives maintained a similar level of news coverage until February 25, when personal narratives tweets started to suppress news coverage, reaching the highest peak on March 11 at the very end. Thus, we conclude there were three periods separated by January 20 and February 23. In the first period, there was little news related to COVID-19 and scarcely any personal narrative tweets. In the second period, both news coverage and personal narrative remained at a moderate and similar level for around one month. For the third period, we see a fast-growing trend for two tweet categories, especially for personal narrative ones which exceed the other in just two days.

We further plot the daily change in the number of news reports and personal narratives tweets per day in Fig. 6. As the trends suggest the possibility of correlation, thus, to testify the dependency between personal narrative tweets and news reports tweets, two null hypotheses are considered. Hypothesis 1: the daily amount of personal narrative tweets is not related to news coverage per day: the daily number of news and reports. Hypothesis 2: the daily change in the number of personal narrative tweets per day are not related to the daily change in news coverage per day. We calculate the Spearman correlation coefficients to testify our null hypotheses on the daily number of tweets and daily change in the number of tweets per day. All statistical analyses are performed using scipy.stats library

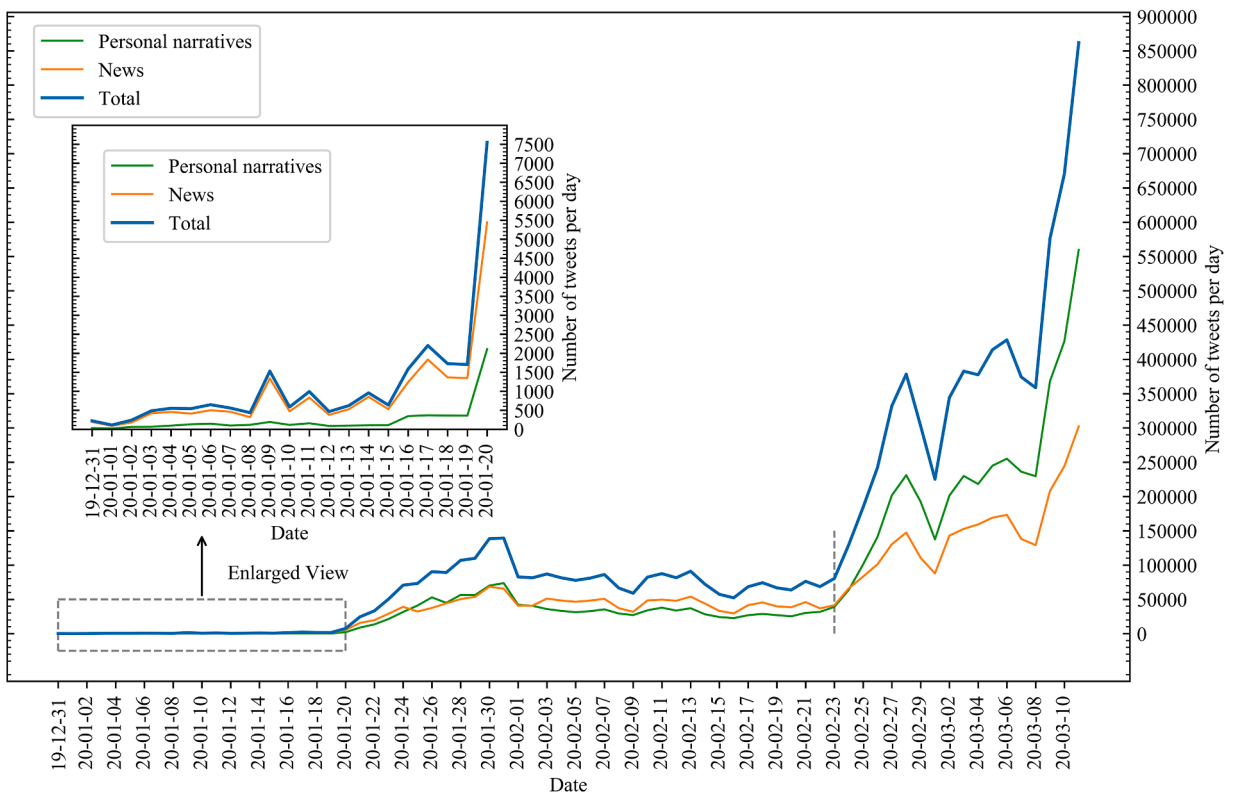


Fig. 5. Overview of the predicted classification results.

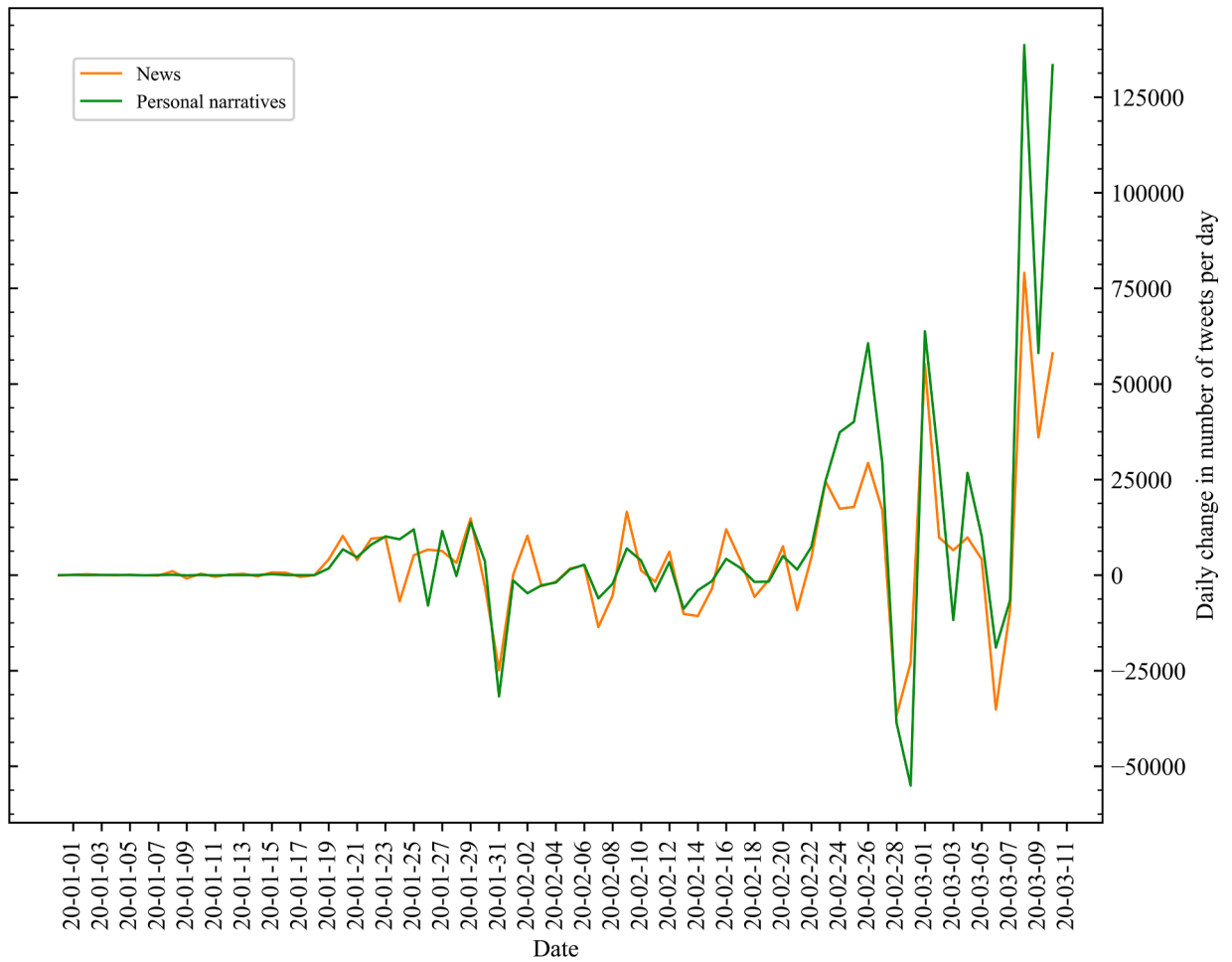


Fig. 6. Daily change in daily number of tweets: personal narratives and news reports.

in Python.

The daily change in the number of personal narratives per day has a strong positive correlation to the daily change in news coverage per day as revealed by the Spearman correlation coefficients (Spearman’s $r = 0.7874, p < 0.0001$). The correlation further increases when simply specifying on the daily number of tweets (Spearman’s $r = 0.9595, p < 0.0001$) instead of change. The results show that we may reject the null hypotheses and there is a significant and strong positive correlation between the number of personal narratives and news reports no matter at the daily or daily change level.

7. Discussion

7.1. Personal engagement and news coverage

The number of personal narratives reflects the public’s attentiveness and even their willingness to take precautions to prevent infection to some degree. As the correlation analysis shows, there is a strong positive correlation between the daily number of personal narratives and news reports. The dynamics of public engagement follow a similar pattern of news coverage but at a comparatively lower level in the first and middle periods, only exceeding that sharply till late February. The mirrored relationship between them has not been identified before but previous studies indicate tweets from major public health agencies could have a broad impact on people’s emotions online (Xi et al., 2022).

Government control is more effective if performed at the initial stages of an emergency (Li et al., 2020). Given the importance of government response on social media for public health management, early warnings of the outbreak should be considered. However, we have found two missed windows for government and media of early warnings and in-time responses for promoting health-related information. The low level of news attentiveness and public attention to the outbreak in early January and February indicates the absent awareness of the possible severity of the virus before any outburst beyond china. Thus, we recommend more extensive publicity and alerts on social networks proactively during the early days of emergency for increasing the public’s attentiveness and sensitivity,

informing precautionary measures, and controlling the spread before a soaring number of domestic cases emerges.

The turning point of covid-19-related tweets around January 20 corresponds to the internet attention change on the Chinese social media (weibo) (Cui & Kertész, 2021; Yu et al., 2021; Zhu et al., 2020). Previous studies on agencies, stakeholders and public health authorities also support our findings largely. Wang et al. (2021) have found that 67 agencies and stakeholders in the U.S. communicated at a very low level with less than ten messages per day over the first two weeks of January and the communication frequencies only increased significantly from late February to middle March. And findings from Li et al. (2021) suggest there was a decrease in the volume of tweets from 21 official accounts at the beginning of February and an increase in late February around February 23.

7.2. Applications of ENDA

At first, we conduct experiments with ESCD dataset with different sizes and augmentation numbers and ENDA attains superior performance than AEDA. The visualization results of the label validity of the augmented text indicate that compared with AEDA, ENDA keeps the original information and conserves the label of original text better. To demonstrate the efficacy of the proposed technique for generality further, the AEDA and ENDA methods are applied to two benchmark datasets: sst-2 and TREC. The results indicate that both AEDA and ENDA could improve the performance of deep model on all sizes. But the superiority of ENDA is not as significant as it is in Section 4. We assume that the different linguistic characteristics of text in these datasets may play a role in influencing augmentation performance. The classification types of early-stage covid-19 dataset are news and personal narratives, which differ greatly in their writing styles as personal narratives are more casual, and emotional, along with plenty of punctuations like question marks and exclamation marks while the news is kind of rigid, precise, and calm. The statistical results in Table 2 and non-parametric test in Section 4.1 have already shown that there is a strong difference on the occurrences of punctuations among different text groups, and news category has more punctuations compared with personal narratives.

The ablation studies on augmentation number, dataset size, and percentage parameter indicate that the former two impact the improvements greatly while the percentage parameter demonstrates irregular fluctuations on model performance. It is worth noticing that the specific number of insertions for each sample is not determined directly by the percentage parameter but selected randomly from the scope with a boundary set by the percentage parameter. The indirect relationship may weaken the influence of percentage parameter change and result in a nonlinear influence.

7.3. Limitations and future research

In this paper, we collect English tweets and study the trends of personal narratives tweets and news coverages on Twitter during the early outbreak stage. However, due to the limited web scraper capability and language barrier, we filter English tweets only but the discourses on other popular languages like Spanish, French, German, and Japanese are also worth exploring. Future studies could investigate the situation in these language communities to see if there is any difference. It is a pity that we only include text data for analysis without considering multimodal data like pictures or videos. Besides that, explorations on demographical and gender differences are also missed in this study as we did not include the geo or gender data of tweets while web scraping.

We choose two other benchmark datasets for the general application of ENDA: sst-2, a movie review sentiment dataset and TREC, a question type dataset. But they present different tendencies within varied augmentation numbers and sample sizes to some extent. More datasets with diverse types could be analyzed to test the efficacy of the proposed technique under different real application circumstances, like author identification.

7.4. Theoretical and practical implications

Despite the limitations, this research proposes a novel but easy data augmentation technique, pictures the growing tendency of daily personal narratives and news reports in the initial stages of the outbreak, and identifies their associations, which delivers insights into the participatory dynamics of public attention and government response on social media during crises. This study has 2-fold implications, theoretically and practically.

From the theoretical point of view, this paper has several implications. In terms of covid-19 social media analysis, despite such an abundance of studies, our research fills the blank of quantitative social media analysis of government response and public attention during the outbreak stage of covid-19 in a broad view with over 9 million covid-19-related tweets. Unlike previous studies which focus on specific Twitter accounts from the government leader or actors, official health authorities, and experts, we provide a more general and comprehensive scope, reducing the bias brought by small samples. Also, our findings underline strong positive correlations between government response and public attention at both the daily and daily change levels.

Secondly, for the data augmentation field, the proposed ENDA approach contributes to an alternative method for creating new samples in an easy but effective way. By adding noises into the original data to create modified copies, the numeric augmentation utilizes the objective of data augmentation and improves classification model performance and robustness with increased amount of training data. The label validity of the augmented text has been examined by extracting the high-dimensional layer output vectors from pre-trained models and visualizing their distributions by t-sne approach. Briefly, we reckon that ENDA could create new samples for training while preserving the original label of text after augmentation. Our evaluation using the annotated ESCD dataset demonstrates that ENDA is superior to AEDA for the classification of personal narratives and news reports. Also, the results on other benchmark datasets: sst-2 and TREC suggest the efficacy of ENDA that it could improve deep model performance for more general text classification tasks.

Generally speaking, several major practical implications for public health policymaking can be yielded from our main findings. Our study not only contributes to the field of emergency response analysis during crises, but also provides insights into public attention in emergencies. The government response online is of great importance for public health management as twitter has evolved into a major tool for information consumption with the rise of social media (Tsao et al., 2021). Related news coverage allows the public to determine the evolution of emergencies and facilitates the government to guide the online sentiment, which would be more effective if occurred in the early stages, for responding to emergencies to the public in real-time (Li et al., 2020). Our findings provide a better comprehension of public social media activity and discourse during crises. The finding on the lack of alerts before January 20 and February 23 demonstrates the existing flaws of timely action on emergency response. For the practice of government response during health or other crises, this finding indicates the future orientation on increasing warnings and alerts in the early stages before the outburst. The finding on public attention level reveals people's awareness and attentiveness towards covid-19 and could have implications for the acceptance and adoption of precautionary and control measures as well.

8. Conclusion

The proposed ENDA approach contributes to an alternative method for creating new samples in an easy but effective way, which keeps label validity of augmented data greatly. Results of ENDA on the early-stage COVID-19 dataset indicate significant improvements for deep models and its superiority compared with the AEDA method. In addition, the efficacy of ENDA has been verified on benchmark datasets through ablation studies. Prediction results on the ESCD dataset present the turning and crucial points for public attention and news coverage with the advent of momentous events in the timeline. We have also found evidence that the number of personal narratives and news reports are greatly correlated at both daily and daily change levels. By dividing the early stage into three sub-periods, we summarize the dynamics of tweets over time, suggesting the missed window for COVID-19 early warnings proactively in early January and February. Thus, we recommend more publicity and timely alerts at the very beginning to increase the public's attentiveness, and to deliver precautionary measures for the control and interruption of the spread before the outburst.

Our study answers important questions about the social media activity from the public and officials during the initial phase of the outbreak and proposes a novel but easy technique for text data augmentation for future research. This quantitative research manages to discover the public discourses on twitter during the covid-19 pandemic, illustrating the dynamic patterns of public engagement and news coverage in the early outbreak stages and providing future orientation on the government response in the networked sphere during global health crises.

CRedit authorship contribution statement

Yuan Chen: Conceptualization, Data curation, Methodology, Formal analysis, Software, Validation, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Zhisheng Zhang:** Supervision, Project administration, Funding acquisition.

Acknowledgments

The work presented in this paper has been support by the National Natural Science Foundation of China under Grant No. 51775108.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:[10.1016/j.ipm.2022.103073](https://doi.org/10.1016/j.ipm.2022.103073).

References

- Agley, J., & Xiao, Y. (2021). Misinformation about COVID-19: Evidence for differential latent profiles and a strong association with trust in science. *BMC Public Health*, 21(1), 1–12. <https://doi.org/10.1186/S12889-020-10103-X>. 2021 21:1.
- Ahmed, W., Vidal-Alaball, J., Downing, J., & Seguí, F. L. (2020). COVID-19 and the 5G conspiracy theory: Social network analysis of twitter data. *Journal of Medical Internet Research*, 22(5), e19458. <https://doi.org/10.2196/19458>
- Alaa Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z. (2020). Top concerns of tweeters during the COVID-19 pandemic: Inveillance study. *Journal of Medical Internet Research*, 22(4), e19016. <https://doi.org/10.2196/19016>
- Alhassan, F. M., & AlDossary, S. A. (2021). The Saudi Ministry of Health's Twitter Communication strategies and public engagement during the COVID-19 pandemic: Content analysis study. *JMIR Public Health and Surveillance*, 7(7), e27942. <https://doi.org/10.2196/27942>
- Alkhodair, S. A., Ding, S. H. H., Fung, B. C. M., & Liu, J. (2020). Detecting breaking news rumors of emerging topics in social media. *Information Processing & Management*, 57(2), Article 102018. <https://doi.org/10.1016/J.IJPM.2019.02.016>
- Aygun, I., Kaya, B., & Kaya, M. (2021). Aspect Based twitter sentiment analysis on vaccination and vaccine types in COVID-19 pandemic with deep learning. *IEEE Journal of Biomedical and Health Informatics*. <https://doi.org/10.1109/JBHI.2021.3133103>
- Ayoub, J., Yang, X. J., & Zhou, F. (2021). Combat COVID-19 infodemic using explainable natural language processing models. *Information Processing & Management*, 58(4), Article 102569. <https://doi.org/10.1016/J.IJPM.2021.102569>
- Basiri, M. E., Nemati, S., Abdar, M., Asadi, S., & Acharyya, U. R. (2021). A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets. *Knowledge-Based Systems*, 228, Article 107242. <https://doi.org/10.1016/J.KNOSYS.2021.107242>
- Behl, S., Rao, A., Aggarwal, S., Chadha, S., & Pannu, H. S. (2021). Twitter for disaster relief through sentiment analysis for COVID-19 and natural hazard crises. *International Journal of Disaster Risk Reduction*, 55, Article 102101. <https://doi.org/10.1016/J.IJDRR.2021.102101>

- Blanco, G., & Lourenço, A. (2022). Optimism and pessimism analysis using deep learning on COVID-19 related twitter conversations. *Information Processing & Management*, 59(3), Article 102918. <https://doi.org/10.1016/J.IJPM.2022.102918>
- Burel, G., Farrell, T., & Alani, H. (2021). Demographics and topics impact on the co-spread of COVID-19 misinformation and fact-checks on Twitter. *Information Processing & Management*, 58(6), Article 102732. <https://doi.org/10.1016/J.IJPM.2021.102732>
- Burel, G., Farrell, T., Mensio, M., Khare, P., & Alani, H. (2020). Co-spread of misinformation and fact-checking content during the COVID-19 pandemic. *Lecture Notes in Computer Science*, 28–42. https://doi.org/10.1007/978-3-030-60975-7_3/FIGURES/4 (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12467 LNCS.
- Chen, Y., & Zhang, Z. (2022). Exploring public perceptions on alternative meat in China from social media data using transfer learning method. *Food Quality and Preference*, 98, Article 104530. <https://doi.org/10.1016/J.FOODQUAL.2022.104530>
- Cui, H., & Kertész, J. (2021). Attention dynamics on the Chinese social media Sina Weibo during the COVID-19 pandemic. *EPJ Data Science*, 10(1), 1–16. <https://doi.org/10.1140/EPJDS/S13688-021-00263-0>. 2021 10:1.
- Cui, X., Goel, V., & Kingsbury, B. (2015). Data augmentation for deep neural network acoustic modeling. *IEEE Transactions on Audio, Speech and Language Processing*, 23(9), 1469–1477. <https://doi.org/10.1109/TASLP.2015.2438544>
- De Rosa, M., Fenza, G., Gallo, A., Gallo, M., & Loia, V. (2021). Pharmacovigilance in the era of social media: Discovering adverse drug events cross-relating Twitter and PubMed. *Future Generation Computer Systems*, 114, 394–402. <https://doi.org/10.1016/J.FUTURE.2020.08.020>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAAACL-HLT* (pp. 4171–4186). <https://doi.org/10.18653/v1/N19-1423>
- Drews, S., Savin, I., van den Bergh, J. C. J. M., & Villamayor-Tomás, S. (2022). Climate concern and policy acceptance before and after COVID-19. *Ecological Economics*, 199, Article 107507. <https://doi.org/10.1016/J.ECOLECON.2022.107507>
- Freifeld, C. C., Brownstein, J. S., Menone, C. M., Bao, W., Filice, R., Kass-Hout, T., & Dasgupta, N. (2014). Digital drug safety surveillance: Monitoring pharmaceutical products in twitter. *Drug Safety*, 37(5), 343–350. <https://doi.org/10.1007/S40264-014-0155-X>
- Griffith, J., Marani, H., & Monkman, H. (2021). COVID-19 vaccine hesitancy in Canada: Content analysis of tweets using the theoretical domains framework. *Journal of Medical Internet Research*, 23(4), e26874. <https://doi.org/10.2196/26874>
- Guntuku, S. C., Sherman, G., Stokes, D. C., Agarwal, A. K., Seltzer, E., Merchant, R. M., & Ungar, L. H. (2020). Tracking mental health and symptom mentions on twitter during COVID-19. *Journal of General Internal Medicine*, 35(9), 2798–2800. <https://doi.org/10.1007/S11606-020-05988-8>. 2020 35:9.
- Haman, M. (2021). Twitter followers of canadian political and health authorities during the COVID-19 pandemic: What are their activity and interests? *Canadian Journal of Political Science/Revue Canadienne de Science Politique*, 54(1), 134–149. <https://doi.org/10.1017/S0008423921000020>
- Han, Xu, Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., Tang, J., Wen, J.-R., & Zhu, J. (2021). Pre-trained models: Past, present and future. *AI Open*, 2, 225–250. <https://doi.org/10.1016/j.aiopen.2021.08.002>
- Han, X., Wang, J., Zhang, M., & Wang, X. (2020). Using social media to mine and analyze public opinion related to COVID-19 in China. *International Journal of Environmental Research and Public Health*, (8), 17. <https://doi.org/10.3390/IJERPH17082788>
- Hanyin Li, Y., Hutch, M., Naidech, A., & Luo, Y. (2021). Using tweets to understand how COVID-19-related health beliefs are affected in the age of social media: Twitter data analysis study. *Journal of Medical Internet Research*, 23(2), e26302. <https://doi.org/10.2196/26302>
- Huang, W., Cao, B., Yang, G., Luo, N., & Chao, N. (2021). Turn to the internet first? Using online medical behavioral data to forecast COVID-19 epidemic trend. *Information Processing & Management*, 58(3), Article 102486. <https://doi.org/10.1016/J.IJPM.2020.102486>
- Imran, M., Ofii, F., Caragea, D., & Torralba, A. (2020). Using AI and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions. *Information Processing & Management*, 57(5), Article 102261. <https://doi.org/10.1016/J.IJPM.2020.102261>
- Jamali, M., Nejat, A., Ghosh, S., Jin, F., & Cao, G. (2019). Social media data and post-disaster recovery. *International Journal of Information Management*, 44, 25–37. <https://doi.org/10.1016/J.IJINFORMMGT.2018.09.005>
- Karimi, A., Rossi, L., & Prati, A. (2021). AEDA: An easier data augmentation technique for text classification. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2748–2754. <https://doi.org/10.18653/V1/2021.FINDINGS-EMNLP.234>
- Kaufhold, M. A., Bayer, M., & Reuter, C. (2020). Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning. *Information Processing & Management*, 57(1), Article 102132. <https://doi.org/10.1016/J.IJPM.2019.102132>
- Knox, E., & Hara, N. (2021). Public engagement with science via social media: A case of communicating the pandemic on Twitter. In , 58. *Proceedings of the association for information science and technology. Association for information science and technology* (pp. 759–761). <https://doi.org/10.1002/PRA2.553>
- Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Proceedings of the 16th annual conference of the international speech communication association*. <https://doi.org/10.21437/INTERSPEECH.2015-711>
- Kouzy, R., Jaoude, J. A., Kraitem, A., Alam, M. B. E., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E. W., & Baddour, K. (2020). Coronavirus goes viral: Quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus*, 12(3). <https://doi.org/10.7759/CUREUS.7255>
- Kozłowski, D., Lannelongue, E., Saudemont, F., Benamara, F., Mari, A., Moriceau, V., & Boumadane, A. (2020). A three-level classification of French tweets in ecological crises. *Information Processing & Management*, 57(5), Article 102284. <https://doi.org/10.1016/J.IJPM.2020.102284>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90. <https://doi.org/10.1145/3065386>
- Kumari, R., Ashok, N., Ghosal, T., & Ekbal, A. (2021). Misinformation detection using multitask learning with mutual learning for novelty detection and emotion recognition. *Information Processing & Management*, 58(5), Article 102631. <https://doi.org/10.1016/J.IJPM.2021.102631>
- Laguna, L., Fiszman, S., Puerta, P., Chaya, C., & Tárrega, A. (2020). The impact of COVID-19 lockdown on food priorities. Results from a preliminary study using social media and an online survey with Spanish consumers. *Food Quality and Preference*, 86, Article 104028. <https://doi.org/10.1016/J.FOODQUAL.2020.104028>
- Lazarus, J. V., Ratzan, S. C., Palayew, A., Gostin, L. O., Larson, H. J., Rabin, K., Kimball, S., & El-Mohandes, A. (2021). A global survey of potential acceptance of a COVID-19 vaccine. *Nature Medicine*, 27(2), 225–228. <https://doi.org/10.1038/S41591-020-1124-9>
- Li, B., Hou, Y., & Che, W. (2022). Data augmentation approaches in natural language processing: A survey. *AI Open*. <https://doi.org/10.1016/j.aiopen.2022.03.001>
- Li, L., Aldosery, A., Vitiugin, F., Nathan, N., Novillo-Ortiz, D., Castillo, C., & Kostkova, P. (2021). The response of governments and public health agencies to COVID-19 pandemics on social media: A multi-country analysis of twitter discourse. *Frontiers in Public Health*, 9. <https://doi.org/10.3389/FPUBH.2021.716333>
- Li, L., Wang, Z., Zhang, Q., & Wen, H. (2020). Effect of anger, anxiety, and sadness on the propagation scale of social media posts after natural disasters. *Information Processing & Management*, 57(6), Article 102313. <https://doi.org/10.1016/J.IJPM.2020.102313>
- Li, S., Liu, Z., & Li, Y. (2020). Temporal and spatial evolution of online public sentiment on emergencies. *Information Processing & Management*, 57(2), Article 102177. <https://doi.org/10.1016/J.IJPM.2019.102177>
- Li, X., & Roth, D. (2002). Learning question classifiers. In *Proceedings of the COLING 2002: The 19th international conference on computational linguistics* (pp. 1–7). <https://doi.org/10.3115/1072228.1072378>
- Li, Z., Hu, H., Wang, H., Cai, L., Zhang, H., & Zhang, K. (2022). Why does the president tweet this? Discovering reasons and contexts for politicians' tweets from news articles. *Information Processing & Management*, 59(3), Article 102892. <https://doi.org/10.1016/J.IJPM.2022.102892>
- Luo, X., He, M., & Yu, Z. (2022). An ideological analysis of the former president Donald Trump's Tweets during COVID-19. *Corpus Pragmatics*, 6(1), 23–38. <https://doi.org/10.1007/S41701-021-00113-4/FIGURES/1>
- Madhani, N., & Dorr, B. J. (2010). Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3), 341–387. https://doi.org/10.1162/COLI_A_00002
- Merkley, E., Bridgman, A., Loewen, P. J., Owen, T., Ruths, D., & Zhilin, O. (2020). A rare moment of cross-partisan consensus: Elite and public response to the COVID-19 pandemic in Canada. *Canadian Journal of Political Science/Revue Canadienne de Science Politique*, 53(2), 311–318. <https://doi.org/10.1017/S0008423920000311>
- Mohammad, S. M., Zhu, X., Kiritchenko, S., & Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4), 480–499. <https://doi.org/10.1016/J.IJPM.2014.09.003>

- Muqsiht, M. A., Kuswanti, A., Pratomo, R. R., & Muzykant, V. L. (2021). Trump's Twitter propaganda during COVID-19. *Jurnal The Messenger*, 13(3), 223–237. <https://doi.org/10.26623/THEMESSENGER.V13I3.3991>
- Obadimu, A., Khaund, T., Mead, E., Marcoux, T., & Agarwal, N. (2021). Developing a socio-computational approach to examine toxicity propagation and regulation in COVID-19 discourse on YouTube. *Information Processing & Management*, 58(5), Article 102660. <https://doi.org/10.1016/j.ipm.2021.102660>
- Raamkumar, A. S., Tan, S. G., & Wee, H. L. (2020). Measuring the outreach efforts of public health authorities and the public response on Facebook during the COVID-19 pandemic in early 2020: Cross-country comparison. *Journal of Medical Internet Research*, 22(5), e19334. <https://doi.org/10.2196/19334>
- Rodrigues, H., Valentin, D., Franco-Luesma, E., Ramarosan Rakotosamimanana, V., Gomez-Corona, C., Saldaña, E., & Sáenz-Navajas, M. P. (2022). How has COVID-19, lockdown and social distancing changed alcohol drinking patterns? A cross-cultural perspective between britons and spaniards. *Food Quality and Preference*, 95, Article 104344. <https://doi.org/10.1016/j.foodqual.2021.104344>
- Rufai, S. R., & Bunce, C. (2020). World leaders' usage of Twitter in response to the COVID-19 pandemic: A content analysis. *Journal of Public Health*, 42(3), 510–516. <https://doi.org/10.1093/PUBMED/FDAA049> (Oxford, England).
- Savin, I., Drews, S., van den Bergh, J., & Villamayor-Tomas, S. (2022). Public expectations about the impact of COVID-19 on climate action by citizens and government. *Plos One*, 17(6), Article e0266979. <https://doi.org/10.1371/JOURNAL.PONE.0266979>
- Seki, K., Ikuta, Y., & Matsubayashi, Y. (2022). News-based business sentiment and its properties as an economic index. *Information Processing & Management*, 59(2), Article 102795. <https://doi.org/10.1016/j.ipm.2021.102795>
- Şerban, O., Thapen, N., Maginnis, B., Hankin, C., & Foot, V. (2019). Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. *Information Processing and Management*, 56(3). <https://doi.org/10.1016/j.ipm.2018.04.011>
- Shen, C., Chen, A., Luo, C., Zhang, J., Feng, B., & Liao, W. (2020). Using reports of symptoms and diagnoses on social media to predict COVID-19 case counts in Mainland China: Observational infoveillance study. *Journal of Medical Internet Research*, 22(5), e19421. <https://doi.org/10.2196/19421>
- Shleifer, S. (2019). *Low resource text classification with ULMFiT and backtranslation*. arXiv preprint arXiv:1903.09244.
- Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Deep learning applications for COVID-19. *Journal of Big Data*, 8(1), 1–54. <https://doi.org/10.1186/S40537-020-00392-9>. 2021 8:1.
- Socher, R., Bauer, J., Manning, C. D., & Ng, A. Y. (2013). Parsing with compositional vector grammars. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (pp. 455–465) (Volume 1: Long Papers).
- Stamatelatos, G., Gyftopoulos, S., Drosatos, G., & Efraimidis, P. S. (2020). Revealing the political affinity of online entities through their Twitter followers. *Information Processing & Management*, 57(2), Article 102172. <https://doi.org/10.1016/j.ipm.2019.102172>
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).
- Tang, L., Fujimoto, K., Amith, M., Cunningham, R., Costantini, R. A., York, F., Xiong, G., Boom, J. A., & Tao, C. (2021). Down the rabbit hole" of vaccine misinformation on YouTube: Network exposure study. *Journal of Medical Internet Research*, 23(1), e23262. <https://doi.org/10.2196/23262>
- Tang, L., Liu, W., Thomas, B., Tran, H. T. N., Zou, W., Zhang, X., & Zhi, D. (2021). Texas public agencies' tweets and public engagement during the COVID-19 pandemic: Natural language processing approach. *JMIR Public Health Surveill*, 7(4), E26720. <https://doi.org/10.2196/26720>. 2021 <https://PublicHealth.Jmir.org/2021/4/E26720>. 7(4), E26720.
- Tsao, S. F., Chen, H., Tisseverasinghe, T., Yang, Y., Li, L., & Butt, Z. A. (2021). What social media told us in the time of COVID-19: A scoping review. *The Lancet Digital Health*, 3(3), e175–e194. [https://doi.org/10.1016/S2589-7500\(20\)30315-0](https://doi.org/10.1016/S2589-7500(20)30315-0)
- Wang, S., Yang, Y., Wu, Z., Qian, Y., & Yu, K. (2020). Data augmentation using deep generative models for embedding based speaker recognition. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 28, 2598–2609. <https://doi.org/10.1109/TASLP.2020.3016498>
- Wang, Y., Hao, H., & Platt, L. S. (2021). Examining risk and crisis communications of government agencies and stakeholders during early-stages of COVID-19 on Twitter. *Computers in Human Behavior*, 114, Article 106568. <https://doi.org/10.1016/j.chb.2020.106568>
- Wang, Y., Croucher, S. M., & Pearson, E. (2021). National leaders' usage of twitter in response to COVID-19: A sentiment analysis. *Frontiers in Communication*, 6, 183. <https://doi.org/10.3389/fcomm.2021.732399>
- Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 6382–6388). <https://doi.org/10.18653/v1/D19-1670>
- Wilson, S. L., & Wiysonge, C. (2020). Social media and vaccine hesitancy. *BMJ Global Health*, 5(10), Article e004206. <https://doi.org/10.1136/BMJGH-2020-004206>
- Xi, H., Zhang, C., Zhao, Y., & He, S. (2022). Public emotional diffusion over COVID-19 related tweets posted by major public health agencies in the United States. *Data Intelligence*, 4(1), 66–87. https://doi.org/10.1162/DINT_A_00101
- Yu, A. W., Dohan, D., Luong, M. T., Zhao, R., Chen, K., Norouzi, M., & Le, Q. V. (2018). QANet: Combining local convolution with global self-attention for reading comprehension. In *Proceedings of the 6th international conference on learning representations, ICLR 2018 - conference track*. <https://arxiv.org/abs/1804.09541v1>.
- Yu, S., Eisenman, D., & Han, Z. (2021). Temporal dynamics of public emotions during the COVID-19 pandemic at the epicenter of the outbreak: Sentiment analysis of Weibo Posts From Wuhan. *Journal of Medical Internet Research*, 23(3), e27078. <https://doi.org/10.2196/27078>
- Zhang, X., Zhao, J., & Lecun, Y. (2015). Character-level convolutional networks for text classification. *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems*, 1, 649–657.
- Zhu, Y., Fu, K. W., Grépin, K. A., Liang, H., & Fung, I. C. H. (2020). Limited early warnings and public attention to coronavirus disease 2019 in China, January–February, 2020: A longitudinal cohort of randomly sampled Weibo users. *Disaster Medicine and Public Health Preparedness*, 14(5), e24–e27. <https://doi.org/10.1017/DMP.2020.68>