The Covid-19 Disinformation Detection on Social Media Using the NLP Approaches

Yi Quan

School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong, 510006, China

201930521058@mail.scut.edu.cn

Abstract. Artificial intelligence has emerged with big data technologies in natural language processing and been applied to creative solutions for overload information especially around the time of the COVID-19 epidemic. This paper provides a comprehensive review of research dedicated to applications of artificial intelligence in misinformation detection. This work organizes the necessary background material for COVID-19-related misinformation detection in NLP, concentrating on the transfer learning technique. Database, data preparation, and modeling make up the major body of information. In the part of modeling, it will merge the attributes of the pre-trained model with the specifical task scenario to explain and present pertinent comments on the future model's improvement under the task scenario. This research will benefit the decision-making and information screen for people's inability to distinguish truth from fiction during the COVID-19 pandemic.

Keywords: Natural Language Processing, COVID-19, Language Model, BERT, XLNet.

1. Introduction

As a global public health event, COVID-19 affects a wide range of people. This also indirectly leads to the huge number of pseudo information audiences. The entire amount of various COVID-19 falsehoods has almost increased multiple times during the epidemic, although the number of relevant fact-checkers is also growing, the volume of disinformation certainly grown even faster. These false or misleading information can lead to confusion and risk-taking behavior that can endanger health and even cause death in extreme cases. It also leads to distrust of health authorities and undermines the public health response. Thus, it is crucial for public health security to use efficient and intelligent approaches to fight misinformation in the time of COVID-19 [1]. One effective solution to this problem is the use of machine learning (ML) and natural language processing (NLP) algorithms to detect and flag fake news and misinformation, using pre-existing labelled datasets of true and false facts.

The bulk of earlier studies using COVID-19 misinformation detection were conducted under supervision, requiring a model to be trained over a long period of time using a dataset containing misinformation that has already been annotated. Obtaining a benchmark disinformation dataset for COVID-19 requires meticulous content inspection, which takes time and requires a lot of labor. It is necessary to verify further evidence as well, such as reliable reports, fact-checking websites, news articles, etc. Using a crowdsourcing strategy to gather annotations might reduce the need for expert review, but the quality of the annotations might decrease. Individual human employees may not possess the domain expertise to distinguish between true information and falsehood since misinformation is deliberately disseminated to deceive others.

The detection of false information is frequently determined as a supervised task learning from a labelled dataset. In contrast, real-world data typically lacks labels. As a result, it is advised to frame the problem as a clustering problem and apply semi-supervised and unsupervised approaches to build misinformation detection systems. When turning to process some emergent text, unsupervised models can also be more useful because it's simpler to get unlabelled datasets [2].

In this paper, the author will present a comprehensive review of the application of transfer learning based on pre-trained model in disinformation detection. Several relevant papers are presented the unique beneficent of transfer learning and it can bring improvement on text classification task. The structure of the present work is as follows: First and foremost, this article provides background information on the COVID-19 misinformation detection area and highlights the current state of the field's research through categorization. Second, the author described the database, pre-processing techniques, the most common text representation models for transfer learning, and how they are integrated with certain task situations in accordance with the modelling process. This research concludes with a list of potential issues and future directions for employing transfer learning for the identification of misinformation.

2. Data sourcing

Self-supervised learning requires a large amount of data for feature extraction and training, while supervised learning requires labelled data for prediction. Good datasets lead to good training results. The datasets below are all about posts on SNS, as shown in Table 1.

1) COVID-19 Twitter data: There are about 120 million tweets collected in the dataset and most of them (around 60%) are in English [3].
2) COVID-19 Category: This dataset is collected data covers the time from January 12 to February 24, 2020, is a portion of the information utilized to train CT-BERT.
3) Vaccine Sentiment: The dataset's contents were gathered between March 2, 2011, and October 9, 2016. It contains three classes: positive, negative, and neutral [4].
4) Twitter Sentiment SemEval: The dataset is available from SemEval-2016 Task 4: Sentiment Analysis in Twitter [5]. The data is labelled by researchers into there categories: negative, neutral, positive.
5) COVID-19-InstaPostIDs: This dataset is also open to the public. It was continuously collected from March 30, 2000 [6].
6) CONSTRAINT 2021: About 10 thousand data points were gathered for the dataset, which was created for COVID-19 false news detection, from multiple online social networks, including Twitter, Facebook, and Instagram. There are social media data points and their related labels, either true or fraudulent, in every dataset but the test set.

Table 1. An overview of the COVID-19 disinformation detection related corpus.

Corpus	Size	Source	Label(s)
COVID-19 Twitter dataset [3]	123M	Twitter	{-}
COVID-19 Category	-	Twitter	{personal, news}
Vaccine Sentiment [4]	-	Twitter	{negative, neutral, positive}
Twitter Sentiment SemEval[5]	20.6K	Twitter	{negative, neutral, positive}
COVID-19-InstaPostIDs [6]	18.5K	Instagram	{-}
CONSTRAINT 2021	10.7K	Twitter, Facebook, Instagram	{real, fake.}

3. Processing of input

Since the raw text cannot be entered directly into a machine learning system, most of the functions are in the preparation stage. There are several procedures that the input data must go through. The procedures will prepare and clean the data for feature extraction and, eventually, model training.

Examples of these procedures include tokenization, stop-word removal, stemming and lemmatization, and part-of-speech tagging.

1) Tokenization. The text data is divided into tokens, which are smaller chunks devoid of any punctuation, during the tokenization process. The tokenization method may be used to convert texts into lowercase or uppercase.
2) Stop-word removal. Stop words, which provide minimal context and retain little important information, are the most common terms in a language. These words aid in the formation of sentences. The most frequent stop-word include articles, prepositions, conjunctions, and a few pronouns like an, are, as of, on, or that, the, these, this, too, was, what, when, where, who, would, etc.
3) Stemming. To express a word as a single phrase, stemming strips a word down to its grammatical roots. For instance, the words "happy," "happiness," and "happily" can all be substituted with the word "happy." Stemming is a method for enhancing and accelerating categorisation. This is accomplished by lowering the input dimension, which raises the possibility of obtaining higher accuracy.

Numerous different types of text sources have been subjected to NLP techniques, but the text found in social media platforms have dramatically different characteristics. Personal thoughts, statuses, or feelings are reflected in online content like tweets. Compared to published literature or news stories, this form of text source is very colloquial and has far greater rates of spelling and grammatical mistakes.

The preparation procedures will enable the ML algorithm to concentrate on the most crucial aspects of the text and avoid being tricked by other, less significant components. Although various NLP pre-processing libraries may carry out each function in a different way, the fundamental ideas are the same across the board.

4. Modelling

Text classification is a common task in the field of NLP. Classification problems require mapping sequences to categories. In traditional machine learning methods, the features on which the machine classification depends need to be informed in advance. To perform classification efficiently, researchers not only rely on traditional machine learning methods, but also dig deep into the field of deep learning and develop many neural network models. This also allows the features to be picked out by the neural network without human's intervention. However, the price of this performance improvement is that the training of the neural network needs to be based on a much higher volume of data, using the existing corpus text to learn to classify the text. If the application field is universal, there is still a huge source of corpus available in this field, but since many natural language classification problems to be solved are under the emerging field, this also leads to a very limited corpus.

A machine learning technique that is now most popular is transfer learning which means one model that has been trained on a specific job is re-proposed on another task. Transfer learning deletes the original output layer of the pre-trained model, and the remaining part of the transfer is used as the input of the new model which is called the base model. The parameters of the base model must be adjusted very precisely to fit with new task. Thus, when initializing the base model, specific dataset will be used to initialize the weights of the neural network. Then after adding a new output layer, updating all the way afterwards and training on the new dataset, the model is formulated and adapted to this domain of specific task throughout this kind of process called fine-tuning [7].

There are a lot of low-level features such as positive or negative objects. In this instance, learning from a huge dataset could enable learning algorithms detect misinformation more accurately. So, in the task of misinformation detection, researchers often migrate the trained model and apply to some sub-domain tasks. In disinformation detection, there are some models usually applied to finish new task.

4.1. Classifier based on BERT model

The BERT model is a pre-training model developed by Google AI lab in 2018, the layout can be shown in Figure 1. It aims to represent unlabeled text by combining the left and right text environments, which means it can generate contextualized embeddings. The BERT model can be modified without substantial structural modification. To adapt to multi-domain natural language processing tasks, it is determined that it is very suitable for use as a base model for transfer learning [8].

BERT is based on a transformer architecture. Figure1. To correctly represent a word, it uses attention mechanisms to obtain information about the pertinent context of a particular word and then encodes that context in a rich vector. The network employs an encoder-decoder architecture much like neural nets. The BERT model comes in two variations: BERT-base and BERT-large. BERT-base is equivalent in size to the OpenAI Transformer; BERT-big is a tremendously enormous model that produced cutting-edge findings. Both have a significant number of encoder layers; the difference is the quality of encoder. The comparison between BERT-base and BERT-large can be shown in Table 2.

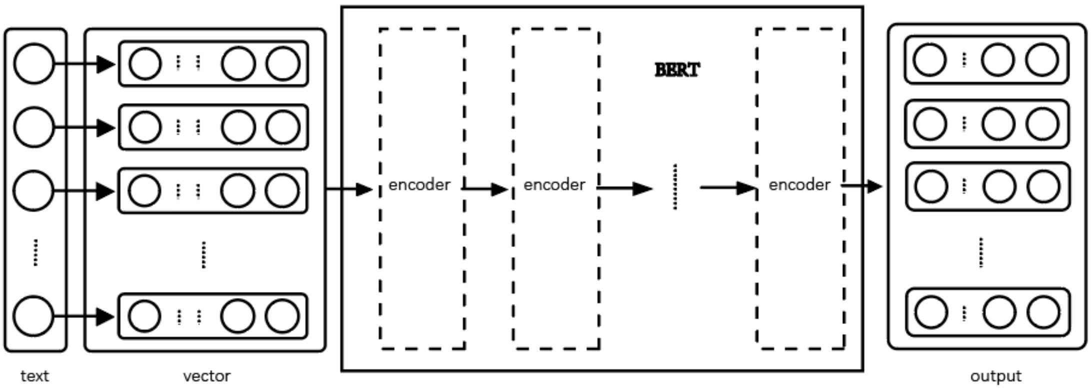

Figure 1. The layout of BERT model.

Table 2. The comparison between BERT-base and BERT-large.

	BERT-base	BERT-large
encoder	12	24
node	768	1024
attention heads	12	16

The BERT model mainly uses training techniques including mask language modeling (MLM), next sentence prediction (NSP), and sentence order prediction (SOP). The MLM makes the model bidirectional in nature because the model must fill the blank in sentence learning from the text before and after. All the techniques above lead BERT to be a suitable model to learn from scratch without expertise about sub-domain. There are currently two methods for using pre-trained language representations in downstream tasks. One is based on the features, the other on is fine-tuning. For dataset with lower volume, feature-based strategy is preferred. The simplest approach to apply BERT for sentence categorization is to use it to categorize a single message. Take the sentence as input, BERT as the first few layers of the model, and add a classifier to the output layer. To train such model, the main process is training the classifier with minimal changes happening to BERT model during the training phase. Based on the above reasons, the BERT model has been widely used in task-specific fields to solve related text representation problems.

The COVID-Twitter-BERT model uses BERT-LARGE as its own base model and achieves significant performance improvements in the target domain by training with text libraries of different classifications. After optimization, it can be particularly used for classification on COVID-19 content, from social media [9]. The detection of COVID-19 misinformation videos model employs transfer learning for the classification by fine-tuning pre-trained models including BERT-base model. They manually select and examine the features of video, such as the percentage of conspiracy comments per video and the percentage for the agreement label and use the final feature combination to check the video's conspiratorial contents [10]. This broadens the field of NLP methods to detect misinformation from text to video though it essentially still relying on video-related text information.

4.2. Classifier based on XLNet model

XLNet model is also a bidirectional transformer model trained by huge volume of unlabeled text corpuses. It learns unsupervised representations of text sequences like BERT. But they differ in pre-training model construction methods. As previously stated, the BERT model uses MLM (Masked Language Models) training approaches, which makes it dependent on masking the input, ignores dependencies between the masked locations, and suffers from a pretrain-finetune discrepancy. Taking the pros and cons into consideration, XLNet was proposed based on a generalized autoregressive pretraining method with the ideals from Transformer-XL. Therefore, XLNet outperforms BERT on plenty of tasks including sentiment analysis and classification [11]. Sunil Gundapu el. proposed an automatic COVID-19 Fake News Detection System based on transformer. The ensemble of three transformer models serves as the model's foundation which includes XLNet to detecting fake news and the system obtained high f1-score on performance [12].

5. Discussion

The potential of transfer learning is enormous, and it frequently improves on already effective learning algorithms. However, further study and investigation are still needed in several transfer learning-related areas. Ideally, transfer learning can lead to improved model performance. However, transfer learning occasionally might result in a decline in performance. Negative transfer may occur for several reasons, including that the source job and the destination task are not adequately connected and that the transfer mechanism is unable to effectively utilize this relationship. Negative transfer must be avoided, which necessitates rigorous research [13].

Current research reveals that traditional transfer learning can only be completed when two fields are sufficiently similar, and that when two fields are not similar, transfer learning can use several fields between these two fields to transfer knowledge transfer-style completion transfer. This information is relevant to the question of how to improve negative transfer [14]. As a result, choosing a suitable base model and enhancing the model's prediction performance from the standpoint of increasing the correlation should be done going forward when employing the transfer learning approach for misinformation detection.

6. Conclusion

With the normalization of the epidemic, relevant information is also flooding all aspects of social life. The processing method of NLP provides an effective and convenient way to detect false information. This paper mainly discusses the disinformation detection model based on transfer learning. The self-supervised learning model using big data has a wide range of applicability, but the training data is relatively static and cannot be well adapted to the scene requirements of false information detection. Retraining not only requires massive data, but also consumes a lot of time and computing power. However, classifiers that use labelled data for supervised learning have the problems of few training samples and difficult data acquisition in emerging scenarios. Transfer learning can transfer pre-trained models such as BERT and XLNet based on a large amount of data as the input layer of the new model. The new model can train only the output layer or all layers, which better balances efficiency and cost. Some disinformation detection models also obtain better inspection results by approaching based on an ensemble of several transformer models.

References

[1] WHO The COVID-19 infodemic . (2022).

[2] Ullah, A. R., Das, A., Das, A., Kabir, M. A., and Shu, K. "A Survey of COVID-19 Misinformation: Datasets, Detection Techniques and Open Issues." arXiv preprint arXiv:2110.00737 (2021).

[3] Emily Chen, Kristina Lerman, and Emilio Ferrara. Covid-19: The first public coronavirus twitter dataset. arXiv preprint arXiv:2003.07372. (2020).

[4] A Demetri Pananos, Thomas M Bury, Clara Wang, Justin Schonfeld, Sharada P Mohanty, Brendan Nyhan, Marcel Salathé, and Chris T Bauch. Critical dynamics in population

vaccinating behavior. Proceedings of the National Academy of Sciences, 114(52):13762–13767, (2017).

[5] Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. Semeval-2016 task 4: Sentiment analysis in twitter. arXiv preprint arXiv:1912.01973, (2019).

[6] Zarei, Koosha, Reza Farahbakhsh, Noel Crespi, and Gareth Tyson. "A first instagram dataset on covid-19." arXiv preprint arXiv:2004.12226 (2020).

[7] Ayoub, Jackie, X. Jessie Yang, and Feng Zhou. "Combat COVID-19 infodemic using explainable natural language processing models." Information Processing & Management 58.4 (2021).

[8] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[9] Müller, Martin, Marcel Salathé, and Per E. Kummervold. "Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter." arXiv preprint arXiv:2005.07503 (2020).

[10] Serrano, Juan Carlos Medina, Orestis Papakyriakopoulos, and Simon Hegelich. "NLP-based feature extraction for the detection of COVID-19 misinformation videos on YouTube." Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020.

[11] Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. "Xlnet: Generalized autoregressive pretraining for language understanding." Advances in neural information processing systems 32 (2019).

[12] Gundapu, Sunil, and Radhika Mamidi. "Transformer based automatic COVID-19 fake news detection system." arXiv preprint arXiv:2101.00180 (2021).

[13] Wang, Zirui, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. "Characterizing and avoiding negative transfer." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[14] Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." IEEE Transactions on knowledge and data engineering 22.10 (2009): 1345-1359.