A Two-step Rumor Detection and Classification Method Using Machine Learning

Borui Pan

Lyle School of Engineering, Southern Methodist University, Dallas, Texas, USA
boruip777@gmail.com

Abstract

The spread of the Internet and mobile devices has made it easier, faster and more widely to disseminate information. But rumors also spread quickly through the Internet, which can have a big impact on people's lives and social stability, especially during the COVID-19 pandemic. Therefore, this paper presents a Two-step model for solving this problem. A novel feature selection method is first proposed to find the suitable features. Then a two-step model is offered to detect and classify COVID-19 rumors simultaneously. Finally, the analysis of the proposed method is described in detail as well.

Keywords

COVID-19, Rumor Detection and Classification, Multi-dimension Features, Two-step Model

1. Introduction

Today, social media has been an essential part of people's daily life. People use it to communicate with each other, obtain information and share information. Because of high level of internet accessing and popularity of personal computer and mobile devices, information becomes easier and faster to be obtained. However, rumors are also easier to spread by internet. Real information helps people to make better decision, but rumors can lead to incalculable consequences, especially when events happen.

The pandemic COVID-19 has been going on for 2 years and the pandemic's impact is reflected not just in health, but also in the economy, education and other aspects. The global economy has been stunted unprecedentedly by huge reduction of production and consumption, and tourism, hospitality and aviation which are important parts of the economy also have a huge impact on the economy [1]. In education, students in school and college can only learn through online classes instead of participating in classrooms [2]. On the other hand, as an important means of spreading the impact of the pandemic, rumors about COVID-19 have exerted a great negative influence on the whole society.

In China, the "Shuanghuanglian Incident" is a good example about how a rumor could have an impact on people's lives during the COVID-19 pandemic. The tag "Shuanghuanglian" had a total reading volume of around 3 billion and a total discussion volume around 1 million in a very short period on the Sina Weibo which is the one of the largest social media platforms in China [3]. The news with "Shuanghuanglian" tag stated that this product works on COVID-19 virus, and this has resulted in a large number of citizens queuing late at night to buy this product. The myth was believed because it played on people's anxiety about the spread of the COVID-19 virus in the circumstances.

As we can see, rumors could trigger social instability and people's anxiety and cause a series of social problems. Thus, in order to prevent the spread of infodemics and maintain social stability, it is important to detect rumors related to COVID-19. However, it is impossible to check every news or blog manually because of the large number. Researchers have done a lot of works about detecting rumors related to COVID-19 with natural language processing (NLP) [4,5]and machine learning method [6,7,8] but there is little research on classification of rumors related to COVID-19 which is also meaningful to prevent the spread of infodemics and maintain social stability. There could be a long duration for a rumor to spread before authorities or agencies receive, verify and respond this rumor. Nevertheless, if there is a model could detect and classify rumors, authorities and agencies could get the rumors from corresponding categories much easier and faster. With the help of such a model, authorities and agencies would have more time to verify the rumor and react to the rumor, so that there would be less time and smaller scale for rumors to spread.

In order to achieve this goal, this paper proposes a two-step model with a novel feature selection method from data of social media. Instead of only adopting profile data or text content, the features are combined with multi-dimension features, including user profile features, spread features, microblog features and temporal features which enables the analysis of news and speeches more stereoscopic. The COVID-19 rumors detection and classification model is consisted by XGBoost and Naive Bayes in 2 steps.

The following is the structure of this paper: Section 2 states some related work. Section 3 is the methodology and the feature selection and the 2 steps model will be illustrated. The discussion about the features and model is described in the section 4. Conclusions are presented in the final section.

2. Related work

Related work in this section is including COVID-19 rumor datasets, previous rumor detection methods and COVID-19 rumor detection methods.

2.1. COVID-19 rumor datasets

Social media is the main way for people to learn information and news reports, which is also the main way for rumor to spread. Therefore, social media could be a good choice to collect data of COVID-19-related events. Chen et al. (2020) collected the first Twitter dataset related to COVID-19 and keep tracking from January 2020 [9]. They have published over 123 million tweets and most of these tweets are in English. Patwa et al. (2021) release dataset with 10,700 fake and real COVID-19-related news [10]. The dataset is collected from different social media and fact checking websites with verifying the veracity manually. Cheng et al. (2021) collected dataset from Twitter and news website, which not only contains the news and tweets but also contains some meta data of COVID-19-related rumors [11]. There are some COVID-19 datasets collected from Chinese social media, and most of them are collected from Sina Weibo, which is the largest social media in China. Hu et al. (2020) collect a large dataset Weibo-COV containing 40 million microblogs from Sina Weibo during December 2019 and April 2020 [12]. Yang et al. (2021) release CHECKED dataset which contains 2,104 microblogs during December 2019 and August 2020 [13]. These microblogs are all fact-checked from Weibo.

2.2. Rumor detection

Rumor detection is not a new topic and there has been a lot of research on it. Jabir et al. (2020) propose new features based on user behavior, propagation features and temporal features from a Twitter dataset [14]. The dataset is called PHEME containing rumors from 5 events: Ferguson unrest, Ottawa shooting, Sydney siege, Charlie Hebdo shooting, and German wings plane crash [15]. And they used a ensemble voting classifier with K Nearest Neighbor (KNN), Naive Bayes (NB) and Support Vector Machine (SVM) to lead to a great obtain in performance. Lotfi et al. (2021) used PHEME as the dataset as well [8]. They identify a new structural features of rumor conversations from Twitter by using reply tree and user graph to make it a computational model. And the classifiers they used are balanced random forest, easy ensemble and XGBoost. Hamidian and Diab (2019) compared a one-step classifier with two-step classifier and they suggested a new meta linguistic and pragmatic features on the PHEME dataset [7].

2.3. COVID-19 rumor detection

Serrano et al. (2020) conducted research about detecting misinformation or rumor in the COVID-19-related videos on YouTube [5]. The way they used to detect is that extracting the comments of every video and use the percentage of conspiracy comments as feature. If the percentage of conspiracy comments is high, the video could contain some misinformation claims. They used simple classifiers on the raw content of these comments and god 89.4% accuracy. Wang et al. built a rumor reversal model during the COVID-19 pandemic [3]. They constructed a new model based on

the traditional SIR model called G-SCNDR model and chose “Shuanghuanglian” as the case for simulation, whose data is collected from Weibo. In the result G-SCNDR model preforms better than the traditional SIR model. Lu et al. (2021) conducted a study on few-shot COVID-19 rumor detection, and the dataset is collected from Weibo [16]. There are only a small number of labeled data. They introduced a few-shot learning-based model called COMFUSE which includes pre-trained BERT model, multilayer Bi-GRUs, multi-modality feature fusion module and meta-learning. Shi et al. (2020) identified COVID-19-related rumor using a new ensemble learning method with user profile, information dissemination and text message form Weibo dataset [6]. Firstly, they chose 4 types of features: emotion-based feature, interaction-based feature, text characteristics and user-related feature and there are 16 features in these 4 dimensions. Then they built a new rumor detection model by applying the XGBoost ensemble learning algorithm and achieved 91% accuracy which is higher than each feature separately.

3. Methodology

This paper provides a new two-step model to detect and classify rumor related to COVID-19. The model is presented in Figure 1, which includes XGBoost and Naive Bayes classifier.

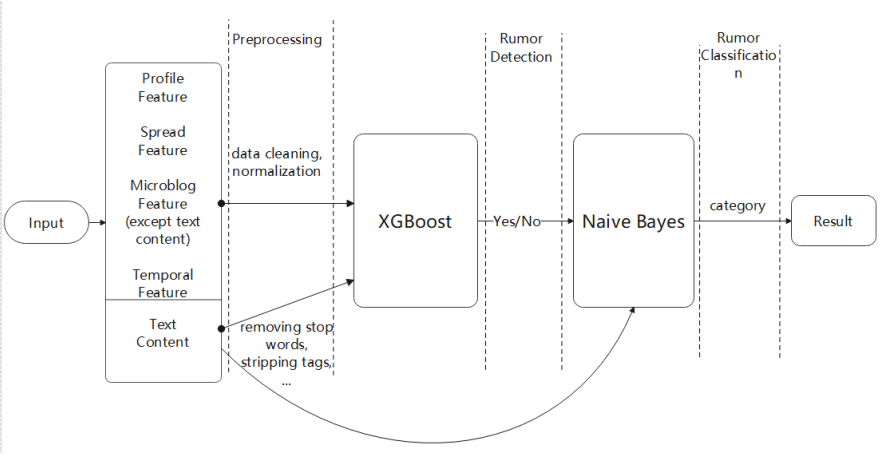

Figure 1: 2-step rumor detection and classification mode

3.1. Features

As Table 1 shows, there are 4 feature sets in this method: profile features, spread features, microblog features and temporal features.

Table 1
Feature set for COVID-19 rumor detection and classification

Feature set	Features
Profile feature	Username, Activity, Post, Following, Followers, Likability, Reputation
Spread feature	Reply, Forwarding, Spread_Speed
Microblog feature	Length, Image, Hashtag, URL, URL_Value, Content, News_C
Temporal feature	Weekday

3.1.1. Profile features:

This feature set is to provide some basic information of the user account.

a) *Username:* String. Represent the username of the account.
b) *Activity:* Integer. Represent the total number of user posts for the last 6 months.

c) *Post:* Integer. Represent the total number of user posts.
d) *Following:* Integer. Represent the total number of people the user follows.
e) *Followers:* Integer. Represent the total number of people follow the user.
f) *Likability:* Float. The definition of Likability was provided by Gupta et al. (2013) as follow [17]:

$$\text{Likability} = \frac{\text{number of favorited}}{\text{number of posts}} \quad (1)$$

g) *Reputation:* Float. The concept of reputation was defined by Shi et al. (2020) as follow [6]:

$$\text{Reputation} = \frac{\text{number of follower}}{\text{number of following} + \text{number of follower}} \quad (2)$$

3.1.2. Spread features:

This feature set shows the scale of the spread of the news or post.

a) *Reply:* Integer. Represent the total number of reply to the news or post.
b) *Forwarding:* Integer. Represent the total number of forwarding of the news or post, like retweeting in Twitter.
c) *Spread Speed:* Float. This paper defines spread speed as following:

$$\text{Spread speed} = \frac{\text{Reply} + \text{Forwarding}}{\text{time (1 day)}} \quad (3)$$

3.1.3. Microblog features:

The feature set describes the content and related information.

a) *Length:* Integer. Represent the total number of words in the news or post.
b) *Image: Binary.* It describes whether the content includes image information. 1 represents the content includes image information, and 0 represents it does not include.
c) *Hashtag:* Binary. It describes whether the content includes hashtag. 1 represents the content includes hashtag and 0 represents it does not include.
d) *URL: Binary.* It describes whether the content includes URL. 1 represents the content includes URL and 0 represents it dose not.
e) *URL_Value:* String. If URL exists, this is the value of URL, else the value is "NA".
f) *Content: String.* This is the content of the news or post.
g) *News_C: Category.* This is the category of the news or post.

3.1.4. Temporal features

This feature shows whether the time of the post is weekday. 1 represents it is weekday and 0 represents it is holiday.

3.2. Preprocessing

After collecting the COVID-19 related dataset, the first step to do some preprocessing to the dataset. For the numeric features, data cleaning and normalization are needed, since normalization could speed up the gradient descent processing to find the optimal solution. For the category feature, News_C need to be one-hot encoded. The text data in the content also needs a cleaning with changing text to lowercase, removing stop words, stripping tags, punctuation, multiple white spaces and numbers and stemming text.

3.3. XGBoost based Rumor Detection

Chen and Guestrin (2016) introduced the XGBoost model which is considered as ensemble learning method [18]. XGBoost keeps iterating in the learning process, and new tree would be

generated based on the previous tree during the process. CARTs regression model is the tree model applied in XGBoost [19]. The formula is as following:

$$\widehat{y_i} = \sum_{t=1}^{n} f_t(x_i) f_t \in F \quad (4)$$

In the formula, n represents the number of trees, F is the set of possible CARTs, $f_t$is function in F, $x_i$is the input and $\widehat{y_I}$ is the prediction. XGBoost has become a popular model because of its accuracy and scalability in most scenarios, and it has advantages over other ensemble learning methods in preventing overfitting and refining speed of computing. Moreover, XGBoost could effectively solve the problem of missing data.

3.4. Naive Bayes based Rumor Classification

Naive Bayes is a traditional machine learning method, which have been widely applied in the field of text classification. The formula of Naive Bayes is as following:

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(B|A)*P(A)}{P(B)} \quad (5)$$

Where this formula represents the probability of A event occurring when B event has already occurred [20]. In the rumor classification task, A event is the category of this news and B event is features of the news, so that the goal is to find the biggest probability of category which is also the result of the classification. Naive Bayes has advantages in speed of computing and the good performance in the text classification [21].

The detailed implementation process is depicted as follows. The preprocessed data would be trained by XGBoost classification. Finally, the preprocessed content data of instance which is detected as rumor would be trained by Naive Bayes classification. The formula here is as following:

$$P(News_C|prep_{contend}) = \frac{P(prep_{contend}|News_C)*P(News_C)}{P(pre_{contend})} \quad (6)$$

4. Discussion

In this section, this paper will discuss why these features and classifiers are adopted.

4.1. Feature selection

There are four dimensions of features are used to detect COVID-19 rumors on social media and each of them is meaningful for the detection.

4.1.1. Profile features:

User with "News" in its username are more likely to post real information. The features activity, post and likability could be considered together. If an inactive user suddenly posts a microblog and has a high likability, it is suspicious. Following, followers and reputation have strong correlation, and the higher the user's reputation, the less likely the user would post a rumor.

4.1.2. Spread features:

Using the number of reply, number of forwarding and spread speed to describe a spread trend of the post.

4.1.3. Microblog features:

The feature length could help to discover the relationship between number of words and whether it is a rumor content with image, hashtag and URL would be more reliable in people's common sense.The content itself is fundamental to our ability to detect rumors.

News_C has two roles in this 2-steps rumor detection and classification model. The first is to help predict as a feature in the XGBoost classifier. As the first role, it can help to discover which category of news or post is more likely to be rumor. The second role is as a prediction in the Naive Bayes classifier.

4.1.4. The temporal features:

It discovers the relationship between rumor publishing and whether it is working days.

4.2. Classifier selection

4.2.1. XGBoost:

Firstly, the number of data might not be large because the dataset with features mentioned in this paper is relatively expensive to collect. However, the XGBoost does a good job in preventing overfitting of small number of data. Secondly, because of the same reason, there might be some missing values in the dataset, nevertheless, XGBoost minimizes the impact of missing values on the prediction results. Moreover, as mentioned in the introduction, this paper tries to propose a method that could help to reduce the time that authorities and agencies get the rumors alerts. XGBoost have a great ability in the computing speed, which means it can achieve the goal of reducing time. Furthermore, Wang et al. (2019) have proved that XGBoost has better performance than Logistic Regression, Support Vector Machine (SVM), and random forest (RF) [22].

4.2.2. Naive Bayes:

To begin with, Naive Bayes is also not sensitive with missing value. In another hand, if the model is deployed online, Naive Bayes is a simple model for online deployment and its good for incremental training. The performance of Naive Bayes has been proved to be good by Miao et al. (2018) in News_Classification and it achieved 92% accuracy on the Fudan University News_Corpus dataset [20].

5. Conclusion

This paper introduces a novel feature selection method and proposes a two-step rumor detection and classification model for the COVID-19-related social media data. The methodology is first described with selected features and two classifiers, including XGBoost and Naive Bayes model. The implementation issue is then stated in the following. Moreover, the paper discusses the underlying logic for selecting these features and shows that 18 basic features from 4 dimensions construct a more stereoscopic, hierarchical and comprehensive feature set to enable a effective COVID-19-related rumor detection and classification model, compared to traditional methods. Finally, the feasibility and practicability of the two machine learning based classifiers and the whole rumor detection and classification model is stated in this paper.

6. References

[1] Seetharaman, G. "How different sectors of the economy are bearing the brunt of the coronavirus outbreak." Retrieved from economictimes. com: https://economictimes. indiatimes. com/news/economy/policy/how-different-sectors-of-the-economy-arebearing-the-brunt-of-the-C orona Virus-outbreak/articleshow/74630297. cms (2020).

[2] Debata, Byomakesh, Pooja Patnaik, and Abhisek Mishra. "COVID-19 pandemic! It's impact on people, economy, and environment." *Journal of Public Affairs* 20.4 (2020): e2372.

[3] Wang, Xiwei, et al. "A rumor reversal model of online health information during the Covid-19 epidemic." *Information Processing & Management* 58.6 (2021): 102731.

[4] Li, Zongmin, et al. "Social media rumor refuter feature analysis and crowd identification based on XGBoost and NLP." *Applied Sciences* 10.14 (2020): 4711.

[5] 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*. 2020.
[6] Shi, Anqi, et al. "Rumor detection of COVID-19 pandemic on online social networks." *2020 IEEE/ACM Symposium on Edge Computing (SEC)*. IEEE, 2020.
[7] Hamidian, Sardar, and Mona T. Diab. "Rumor detection and classification for twitter data." *arXiv preprint arXiv:1912.08926* (2019).
[8] Lotfi, Serveh, et al. "Rumor conversations detection in twitter through extraction of structural features." *Information Technology and Management* 22.4 (2021): 265-279.
[9] Chen, Emily, Kristina Lerman, and Emilio Ferrara. "Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set." *JMIR public health and surveillance* 6.2 (2020): e19273.
[10] Patwa, Parth, et al. "Fighting an infodemic: Covid-19 fake news dataset." *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*. Springer, Cham, 2021.
[11] Cheng, Mingxi, et al. "A COVID-19 rumor dataset." *Frontiers in Psychology* 12 (2021).
[12] Hu, Yong, et al. "Weibo-COV: A large-scale COVID-19 social media dataset from Weibo." *arXiv preprint arXiv:2005.09174* (2020).
[13] Yang, Chen, Xinyi Zhou, and Reza Zafarani. "CHECKED: Chinese COVID-19 fake news dataset." *Social Network Analysis and Mining* 11.1 (2021): 1-8.
[14] Jabir, Hussam Mohammed, Mohammed Abdullah Naser, and Safaa O. Al-mamory. "Rumor detection on twitter using features extraction method." *2020 1st. Information Technology To Enhance e-learning and Other Application (IT-ELA*. IEEE, 2020.
[15] Zubiaga, Arkaitz, Maria Liakata, and Rob Procter. "Learning reporting dynamics during breaking news for rumour detection in social media." *arXiv preprint arXiv:1610.07363* (2016).
[16] Lu, Heng-yang, et al. "A novel few-shot learning based multi-modality fusion model for COVID-19 rumor detection from online social media." *PeerJ Computer Science* 7 (2021): e688.
[17] Gupta, Aditi, et al. "Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy." *Proceedings of the 22nd international conference on World Wide Web*. 2013.
[18] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.
[19] Trendowicz, Adam, and Ross Jeffery. "Classification and regression trees." *Software project effort estimation*. Springer, Cham, 2014. 295-304.
[20] Miao, Fang, et al. "Chinese news text classification based on machine learning algorithm." *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*. Vol. 2. IEEE, 2018.
[21] Katari, Rohan, and Madhu Bala Myneni. "A survey on News_Classification techniques." *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*. IEEE, 2020.
[22] Wang, Shihang, et al. "Machine learning methods to predict social media disaster rumor refuters." *International journal of environmental research and public health* 16.8 (2019): 1452.