International Journal of Smart Home
Vol.14, No.2 (2020), pp.39-46
http://dx.doi.org/10.21742/ijsh.2020.14.2.05

LSTM based Online Public Opinion Rumors Recognition Method

Liu Dan[1]

Heilongjiang Vocational College of Winter Sports, Heilongjiang, China
277416530@qq.com

Abstract

This paper tries to judge the true and fake information of a large number of public opinions in the network, keep the truth, remove the fake, and provide reference for the government public opinion workers to judge rumors. data such as rumor and non-rumor topics, replies and other data of microblog were collected as data sets, and then programmed with paddle fluid API, and the recurrent neural network model was configured. The data set was used for model training, and finally the model analysis and detection were carried out. Through LSTM model training and data analysis, rumor events in public opinion can be digitized, and fake information feature set in text can be mined, so as to make better rumor recognition and make public opinion workers better control rumors.

Keywords: *Network public opinion, LSTM, Rumor recognition*

1. Introduction

According to the development report of China Internet association in 2019 [1], by the end of 2018, the scale of Chinese Internet users will rise to 829 million, with 56.63 million new users, and the popularization rate will reach 59.6%. Compared with the end of 2017, it will increase by 3.8% and be 2.6% higher than the global average. With the continuous development of China’s Internet industry, the popularity of the Internet in people’s life is constantly improving, making the more convenient emerging industries are also developing rapidly. At the same time, in the era of information explosion, news communication becomes more and more important, and network information dissemination is more and more fast and convenient, which not only makes it easier for the public to release information, but also shortens the time for information dissemination, but also makes public opinion vulnerable to the influence of the public, producing rumors and fake information of different degrees. Due to various reasons, some incidents have just happened and have not been fully recognized by people. Some fake statements often affect the public attitude, which leads to the rapid fermentation of public opinion on emergencies in a short period of time, leading to social unrest [2]. In order to maintain the stability of the network, it is very important to maintain the stability of the network. At present, there are a variety of research directions on public opinion in China, including: Research on public opinion index system, evolution model, public opinion transmission path, public opinion identification analysis, public opinion data analysis, public opinion early warning, etc. in data analysis, it can be divided into several sub categories, including but not limited to: user sentiment analysis, user behavior analysis, rumor detection and recognition [3]. For public opinion workers, the importance of rumor detection

Article history:
Received (July 21, 2020), Review Result (August 27, 2020), Accepted (September 29, 2020)

Print ISSN: 1975-4094, eISSN: 2383-725X IJSH

and recognition is no less than public opinion identification and public opinion early warning. A perfect rumor recognition mechanism can greatly save public opinion work time and energy. According to the above situation, this paper analyzes the topic and reply data of rumors and non-rumors in microblog, and captures and analyzes the text information of fake data through the circular neural network, so as to have a more intuitive and clear understanding of the rumor feature set, and summarize it to find its shortcomings and maturity [4]. Scientific and effective screening of rumors can provide theoretical basis for government research and recognition, establishment of early warning mechanism and establishment of emergency plan [5]. Through the in-depth research on the recognition and detection of network and sentiment rumors, we can promote the development of real-time monitoring, correct guidance and scientific control of network public opinion, so as to reduce or avoid the social harm caused by negative emergency network public opinion.

2. Related works

At present, the three mainstream models in rumor recognition field are content-based modeling, such as knowledge base oriented model and social network based model. The following mainly introduces three rumor recognition methods: knowledge base oriented, content oriented style oriented and social network-based.

2.1. Knowledge base

Knowledge base oriented, that is, according to the existing expert system to study and use, so as to realize rumor recognition. Facts confirm that [6] system is similar to rumor recognition system [7]. The system validates the viewpoints and objectives described in the paper. Similar to QA system, it is a relatively complex domain of NLP as knowledge expression and knowledge reasoning. Knowledge database data set has a centralized partition scheme. 1) Expert system [8]: knowledge database created by experts in various fields. Obviously, the efficiency and scalability of this method are very poor. However, for vertical categories (biology, historical records), we can try to use more objective facts for classification. 2) Collective intelligence [9]: a series of knowledge databases established from the feedback of users' collective knowledge. After 1 and 2 are available, similar retrieval methods can be used to evaluate the similarity of new content and make full use of the accumulated historical content features. 3) Algorithm classification [10]: use knowledge or case diagram to evaluate the reliability of content. At present, the most important open knowledge map is the dataset of DB pedia and Google relation extraction.

2.2. Content style

Content style oriented rumor recognition refers to the use of the writing style of the text itself to retain the syntactic structure of sentences, and to capture grammatical information through context free grammar or other deep-seated NLP models (such as RST rhetorical dependency theory). According to the description types of the recorded text information, the authors can be divided into two categories. This measures the extent of deception, as well as the extent of subjective and objective interpretation (more objective and fairer possibilities). The shocking body of the title party falls into this category. Among them, the features that can be used by deceptive news include conventional features and aggregate features.

General features, such as page, text, image, title, etc. Aggregation feature is a combination of several conventional features and supervised training of sub model problems. The output of these submodels can be used as aggregate functions in the spoofing message region. [Figure 1] shows the main feature set used, which is mainly classified and judged by the style dimension, text dimension [11], image dimension [12] and Title Dimension of graphic static content, and then refine and classify these four dimensional features to get the final recognition method.

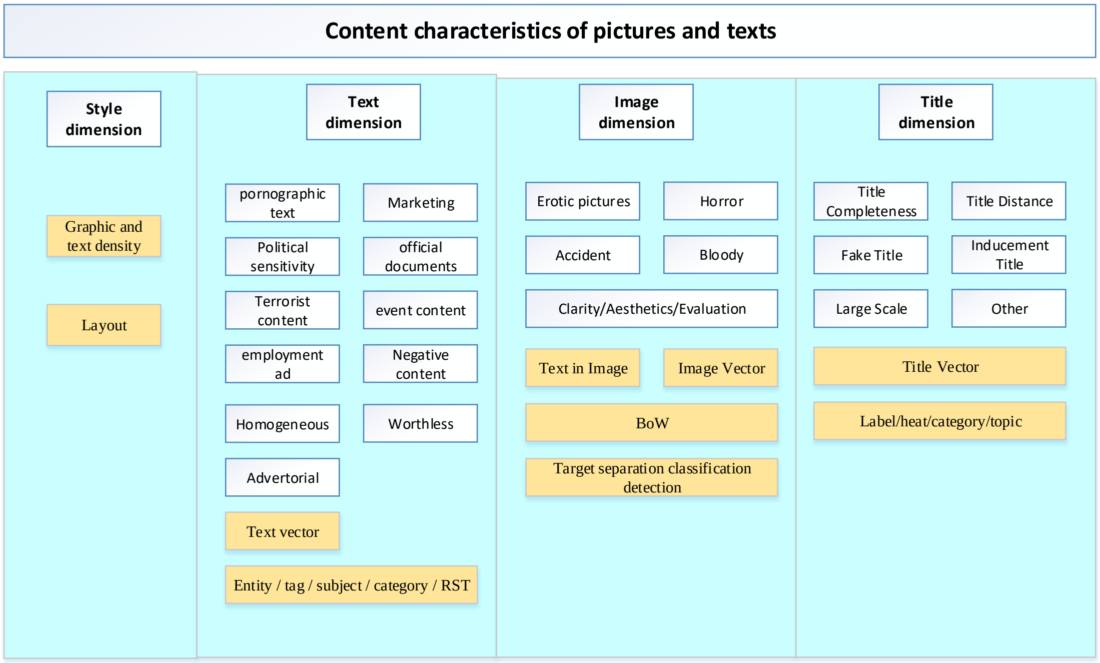

Figure 1. Content characteristics

2.3. Social network

Social network is a modeling method based on user behavior and rumor spreading trajectory. It can be divided into two types, based on position and based on communication behavior. The former is mainly based on the user's operation on the content (comments, likes, reports, etc.) to build a matrix or graph model. The object modeling based on propagation behavior is similar to the behavior transmission of PageRank [13]. (1) Tracking the propagation trajectory of fake news, and further investigation on specific fake news through graph model and evolution model. (2) Identifying the key communicators of fake news is very important to reduce the spread scope of social media.

3. Model

3.1. Data collection

The data used in this practice is the Chinese rumor data collected from Sina Weibo fake information reporting platform. The data set includes 2458 rumors and 2206 non-rumors, all of which are in JSON format. As shown in the [Table 1], the text field is the original text.

Table 1. Partial captured data

Kids	UID	Parent	Text	Mid	Date

[]	Sunlmy	-	How could it happen? I've never heard of such a problem	y83bvg3	2018-04-25
[]	Chengchengcc	-	No expectation, no disappointment	8i3bvtg	2019-01-06
[]	1912456798	I8caei	I hope it can be solved properly	i8isbe4	2018-07-14
[]	ZhaNGiengh		It's not true. Ask for the truth	i8yeh3v	2018-12-18
["i8n8PI", "i8m8k()", "i8erpb"	Shmily397	I8nieya	It's obviously hyped	y34jeyb	2019-02-23

3.2. Data processing

The overall process of data processing and preparation is shown in [Figure 2].

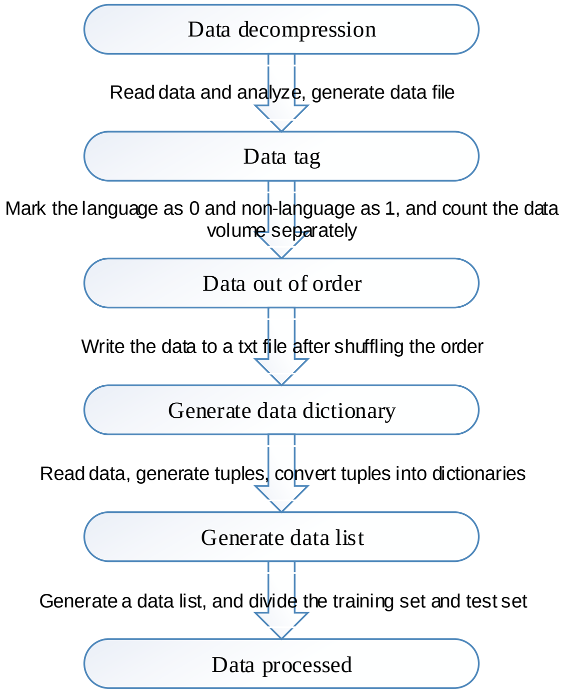
Data decompression
Read data and analyze, generate data file
Data tag
Mark the language as 0 and non-language as 1, and count the data volume separately
Data out of order
Write the data to a txt file after shuffling the order
Generate data dictionary
Read data, generate tuples, convert tuples into dictionaries
Generate data list
Generate a data list, and divide the training set and test set
Data processed

Figure 2. Flow chart of data preparation

(1) Decompress the data, read and parse the data, and generate all_ data.txt Documents. The packages to be imported are: zip file, OS, random, image in PIL, image enhance in PIL, JSON.

a. To decompress the original data, the data is decompressed into. TXT file format, and the Chinese field is extracted as data tuples.

b. Divide the rumor and non-rumor data, mark and count the rumor and non-rumor data, and put the rumor data, non-rumor data and all data into files. At the same time, the total amount of rumor and non-rumor data is analyzed by traversal data method, and the total number of rumor and non-rumor data is counted respectively. The final statistical results are as follows: the total amount of rumor data is 2458, and the total amount of non-rumor data is 2206.

c. After that, all the data will be disordered and written to all_ data.txt Medium.

(2) Generate data dictionary. The packages to be imported are OS and CPU in multiprocessing_ count、numpy、shutil、paddle、 paddle.fluid , image in PIL matplotlib.py -plot。 Generate data dictionary: read all the data and generate a tuple, then convert the tuple into a dictionary and save it locally. 3) The data list is generated, and the training set and verification set are divided. The training set and verification set are divided into two parts: create serialization data, divide training set and verification set according to proportion, and store them in eval_ list.txt And train_ list.txt 。

3.3. Model configuration and training

3.3.1. LTSM

Long Short-Term Memory is one of the deep learning algorithms. It is a kind of recurrent neural network which takes sequence data as input, recurses in the advancing direction of sequence, and all cyclic units are linked in chain. Among them, bidirectional recurrent neural network and long-term and short-term memory network are two kinds of common recurrent neural network. It is mainly used in speech recognition, language modeling, machine translation and other fields, and can also deal with computer vision problems including sequence input.

3.3.2. Model configuration

After the work of data preparation, we build a circular neural network, and extract the text features in it to complete the microblog rumor detection. Among them, dynamic in paddlepadde API_ The LSTM interface has implemented LSTM for us [14]. The loss function [15] and the function of accuracy are also defined. (1) Build a LTSM neural network. Firstly, the long-term and short-term memory network is defined. The IDs of the data is taken as the input, and the soft - Max is used as the output layer of the full connection. The size is 2, that is, the positive and negative sides. (2) Define the data type. Define input data, LOD_ Level does not specify that the input data is the sequence data (3) and defines the loss function and accuracy function.

a. After a loss function is defined, it is also averaged because the loss value of a batch is defined. The function of defining loss function is to measure the quality of model prediction.

b. We also define an accuracy function, which can output the classification accuracy when we train. (4) Training and evaluation of recurrent neural networks. At the end of each round of training, the verification set is used to verify the model, and the corresponding loss value cost and accuracy ACC are calculated, and the training curve and training results are displayed. After the above steps, the accuracy rate is obtained. Table 2 shows the loss value and accuracy rate.

Table 2. Cost rate and accuracy

Round	Training set			Testing set	
	Batch	Cost rate	Accuracy	Cost rate	Accuracy
1	0	0.69731	0.49221	0.64573	0.75049

2	0	0.63821	0.82804	0.62247	0.77609
3	0	0.60962	0.85945	0.60759	0.80348
4	0	0.59755	0.88279	0.58972	0.80778
5	0	0.56849	0.88283	0.57627	0.81597
6	0	0.55269	0.89066	0.56530	0.84028
7	0	0.53524	0.89842	0.55354	0.84027
8	0	0.52276	0.89851	0.54422	0.82769
9	0	0.50986	0.89838	0.53459	0.82781
10	0	0.49543	0.89851	0.52579	0.84041

4. Result and discussion

In this paper, the research from the knowledge base oriented, content-oriented style, based on social network modeling three aspects in theory that rumor detection can understand the information of text data from different methods, and obtain the feature set of fake information. The data sets of microblog rumors and non-rumor topics and replies are used as data sets. The paddle fluid API is used to program and the cyclic neural network is configured. The rumor and non-rumor data sets are trained by defining the network, defining the loss function, and defining the optimized scheme. Finally, the information is screened through the evaluation of the model. The results show that the trained model can identify rumors and non-rumors through the feature set of fake data to a certain extent. At the same time, in order to judge rumors better, the following feature sets can be selected to judge rumors:

(1) sensational news of non-governmental newspapers. Big news on the Internet, such as: a cancer has been conquered and a satellite will hit the earth. It is necessary to determine the source of such news. If it is a small media or personal source, the official will release it. This is a rumor.

(2) The threat of non consensus. People are instinctively afraid of the unknown. What kind of danger will endanger life is basically the focus of mainstream experts, and there will be a consensus of the whole society. For example, we all have a common understanding of the dangers of heart disease, car accidents, cancer and so on. However, faced with the threat of unknown fields, the public chose to “believe in its harm, not in its harmlessness”, and did not have the energy and time to study the truth of these problems. So, as long as you exaggerate the facts, intimidate others, and mainstream science and the media do not explicitly conclude that it is harmful, it is rumor. (3) Information of unknown origin. All information depends on the identity of the publisher. Pretending to be an expert is often a rumor. Experts have insights in their field, but they can make mistakes in other areas. Even famous journalists, if their information is not from authoritative experts in professional fields, it is not worth believing. For example, Zhang Jie, as a singer, is not a physics expert at all, then his insight in the field of physics is not credible. Even for different categories of the same discipline, we can not be vague. For example, the opinions of respiratory department on orthopedics are basically separated by different lines, so there is no credibility.

References

[1] CNNIC, “The 43rd “Statistical report on internet development in China,” http://www.cac.gov.cn/2019-02/28/c_1124175677.htm, **(2019)**

[2] Zhang Yuliang, “Network opinion risk evaluation index system based on the cycle of emergency,” Information Science, vol.30, no.7, pp.1034-1037, **(2012)**

[3] Jiang Yanchuan, Xiao Tieyan, and Ling Xiaoming, “Study on the situation of public opinions from network on campus in the new media environment and the guiding strategy,” Journal of Chongqing University (Social Science Edition), vol.18, no.1, pp.142-148, **(2012)**

[4] Zhao Jianhua, Wan Kewen, “Research on the communication dynamics model of social network public opinion based on the SIS model,” Information Science, vol.35, no.12, pp.34-38, **(2017)**

[5] Tang Tao, “Research into monitoring net-mediated public sentiment based on methods of information science”, Information Science, vol.32, no.1, pp.124-127, **(2014)**

[6] Xiong Yan, “Refuting and paraphrasing rumors can eliminate the illusion of fact”, Modern Communication (Journal of Communication University of China), vol.2018, no.3, pp.74-79, **(2018)**

[7] He Gang, Lv Xueqiang, and Li Zhuo, “Automatic rumor identification on microblog”, Library and Information Service, vol.57, no.23, pp.114-120, **(2013)**

[8] Liu Hanbo, “Rumors of WeChat as a risk culture – information mutual reward and role playing under “ignorance and shame”,”, National Arts, vol.2017, no.5, pp.36-41, **(2017)**

[9] Zhan Xin, Xia Zhijie, Luo Mengying, and He Yin, “Research on the factors affecting the collective intelligence to suppress social media rumors spreading,” Library, vol.2018, no.8, pp.85-90**, (2018)**

[10] Lin Rongrong, “Research on microblog rumor recognition based on sensitive word database,” Zhongnan University of Economics and Law, M.S. thesis, **(2018)**

[11] Jiang Ying, Zhang Jing, Zhu Lingxuan, and Qu Chang, “Analysis of online rumor text syntactical structure features and the monitoring system” Electronic Design Engineering, vol.25, no.23, pp.7-10, **(2017)**

[12] Deng Shengli, and Fu Shaoxiong, “An analysis of the influence of attached information on trusting and sharing health-related rumors in social media,” Information Science, vol.36, no.3, pp.51-57, **(2018)**

[13] Zhiwei Jin, Juan Cao, Yongdong Zhang, “News verification by exploiting conflicting social viewpoints in microblogs,” hirtieth Aaai Conference on Artificial Intelligence. AAAI Press, **(2016)**

[14] Chen Fan, “Research on weibo rumor recognition method based on LSTM sentiment analysis model,” Chongqing University, M.S. thesis, **(2018)**

[15] Yang Guiyuan, Tang Xiaowo, “A new prediction and evaluation method - loss function method,” Prediction, vol.17, no.3, pp.38-40, **(1998)**

This page is empty by intention.