Short text filtering based on gradient boosting decision tree

Zikun Zhu\textsuperscript{1}, Xiaodan Yu\textsuperscript{2}

\textsuperscript{1}Beijing Jiaotong University, No.3 Shangyuancun, Haidian, Beijing, China
\textsuperscript{2}Shenyi Xin’anxian community, BaoAn Distrcet, Shenzhen, Guangdong, China
18120492@bjtu.edu.cn

Abstract. With the rapid development of the Internet, event detection from massive data has become a research hotspot. However, the existing event detection methods for social networks rarely consider the noise data in short text data in detail. Therefore, there is a lot of noise in the input of event detection, resulting in a large number of false-positive events in the event detection results, which affects the efficiency and accuracy of event detection. In this paper, three types of features are proposed to mine text features, and the gradient boosting decision tree algorithm is used. Experiments show that the algorithm has good filtering performance and interpretability for short texts in social networks.

1. Introduction

Web2.0 is an era in which users dominate the production of content, and huge user groups generate excessive amounts of information in social networks. In the current mainstream media, such as Twitter, Facebook, Weibo, users can record life, talk about current affairs, and share what they see and think anytime and anywhere. The massive amount of information has brought new problems to the data mining of social networks, and how to effectively mine it has become a new research hotspot. Event detection based on social networks is the key content. According to keywords, get related tweet collections through crawlers or APIs, clean and cluster data, mine hot events among them, understand real hot spots, and build a bridge between the network and reality. There is a lot of noise in the massive data, but the existing methods do not filter it in detail, resulting in low accuracy of event detection. Filtering noise data can be regarded as a text classification problem. In 1950, text classification used the method of expert rules to classify and then developed an expert system. The disadvantage of this system is that it consumes manpower and time, and has no learning ability and poor generalization ability. Later, a method based on machine learning appeared to classify data. First, label the data as training samples, then perform feature engineering processing on the text, and learn the machine learning algorithm to understand the relationship between features and text categories. Commonly used classification algorithms are Naive Bayes, Support Vector Machines, Hidden Markov Models, Gradient Boosted Trees, and Random Forests. In 2012, AlexNet, a model based on deep learning, gained a huge advantage in the ImageNet competition. Since then, deep learning models have been applied to a wide range of tasks in Computer Vision and Natural Language Processing to improve accuracy. These models try to learn feature representations and perform classification tasks in an end-to-end manner and can discover hidden patterns in the data, but lack interpretability. Compared with long texts, social network short text data has the following characteristics: a large amount of data but short; more noise and simple expression. Therefore, the traditional bag-of-words model that relies on text word frequency information is not applicable. Based on these, this paper considers the background characteristics of the text publisher, the statistical syntactic characteristics of the text, and the subsequent influence of the text on the social
network as features, to train the gradient boosting decision tree which has good performance and interpretability. The following parts are feature engineering, experiment and discussion, and conclusion.

2. Feature Engineering

Based on the short text data characteristics of social networks, user background characteristics, text syntax characteristics, and text propagation characteristics are selected as the feature input of the gradient boosting decision tree.

2.1. User background characteristics

User background features are mainly used to distinguish different users, such as ordinary public users and public figures or institutional personnel with different social purposes. Most users interact with friends and share their ideas. Some users rely on their own popularity or institutions rely on influence to post-oriented comments to drive topic development. Some users use pure marketing accounts to release relevant product information and user experience information.

User background characteristics specifically include: whether the user is authenticated, the number of users’ followers, the number of users’ friends, the number of users participating in public groups, and the number of users posting. According to whether the user is authenticated, the number of users followed, the user’s influence and credibility information can be obtained. The number of user followings, the number of users’ posts, and the number of participating public groups can be used to obtain information on the user’s activity. The user's credibility, influence, and activity can indirectly determine that the published text data is probably the effective information data.

2.2. Text syntactic features

The syntactic features of short texts on social networks mainly refer to the part-of-speech statistics, part-of-speech collocations, and the proportion of stop words in the text data. Posting news on platforms such as Weibo and Twitter has word limits, resulting in short texts and refined semantics. The main elements of an event are time, place, characters, and actions, which correspond to the part of speech of the sentence, namely nouns, verbs, and related collocations.

On this basis, this paper puts forward the syntactic features of the text: the number of verbs, the number of named entities, the number of prepositions, the number of nouns, the number of pronouns, the number of adjectives, the proportion of stop words and punctuation marks in sentences. Through statistical analysis of the frequency of nouns and verbs, we can find whether there is a subject in a sentence; through statistical analysis of the frequency of part of speech, we can find that if stop words and punctuation occupy too much, it may be an unrelated sentence. If there are too many pronouns, such as the first person, it may be an opinion sentence.

2.3. Characteristics of spreading influence

For the social network short text impact characteristics, it mainly refers to the follow-up discussion on the popularity, communication strength, and subsequent impact of the social network short text data after the release of the social network short text data. An unexpected event generally has the characteristics of high discussion enthusiasm, high attention, and widespread range. These events allow more users to participate in the discussion of the event, and for a small matter in daily life In general, it will not have a high degree of attention, will not cause extensive discussion in the society, and the scope of influence will be relatively small.

The characteristics of text influence include the number of likes, the number of retweets, and the number of comments. By analyzing the influence characteristics of text communication, we can distinguish the trivial things in ordinary life from the hot events in current affairs. The number of favorites represents the popularity of the event, while the number of retweets and comments shows the influence range of the event.
3. Experiment and discussion

3.1. Data introduction
Use the "Hong Kong" keyword to get relevant tweets in the August 13, 2020, Twitter stream. After tagging 1000 tweets, 643 noise sentences and 357 event-related sentences are obtained.

3.2. Gradient boosting decision tree
Input: Data set \(\{(x_i, y_i)\}_{i=1}^n\), and a differentiable loss function \(L(y_i, F(x_i))\)

Step 1: \(F(x_0) = \arg \min \left(\sum_{i=1}^n L(y_i, F(x_i))\right)\)

Step 2: for \(m = 1\) to \(M\):
   a) For \(i = 1, ..., n\), compute \(\gamma_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x) = F_{m-1}(x)}\)
   b) Fit a regression tree to the \(\gamma_{im}\) values and create terminal regions \(R_{jm}, j=1, ..., J_m\)
   c) For \(j = 1, ..., J_m\) compute \(\gamma_{im} = \arg \min \sum_{x \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)\)
   d) Update \(F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} j_{jm} I(x \in R_{jm})\)

Step 3: Output \(F_M(x)\)

3.3. Experimental steps
Step 1: Data preprocessing: word segmentation, part-of-speech tagging, named entity recognition
Step 2: Training: train GBDT, Random Forest, Support Vector Machine and adjust parameters.
Step 3: Result: get the accuracy of verification set by five fold cross validation.
Step 4: Analysis: comparative analysis of different classifiers and features.

3.4. Experimental result
From the table 1, various indicators of GBDT are better than SVM and RF classifiers. Because GBDT is a strong classifier formed by a variety of weak classifiers. Its effect is better than a single classifier. The classification effect of GBDT classifier is better than RF. Because the output of GBDT is the accumulation of all results, and RF uses the principle of majority voting determines the final result.

| Evaluating indicator | SVM     | Random Forest | GBDT   |
|----------------------|---------|---------------|--------|
| Accuracy             | 0.701   | 0.760         | 0.805  |
| Precision            | 0.681   | 0.705         | 0.805  |
| Recall               | 0.601   | 0.622         | 0.812  |
| F1                   | 0.638   | 0.661         | 0.807  |
| AUC                  | 0.802   | 0.829         | 0.896  |

3.5. Parameter adjustment
This section introduces the adjustment of GBDT parameters, such as: \(n\) estimators, learning_rate, max_depth, min_samples_leaf, min_samples_split. First adjust the \(n\) estimators and learning_rate in the promotion framework, choose a smaller learning rate, and adjust the \(n\) estimators grid parameters. Next, adjust the parameters of the decision tree, first adjust the max_depth and min_samples_split together, determine the max_depth according to the optimal value of the output, and then adjust the minimum number of divided samples. Then adjust the parameters of min_samples_split and min_samples_leaf together. Confirm whether the optimal value of the two is on the boundary, if it is on the boundary, further change the parameter range and grid adjustment. Finally passed, reduce the learning rate, increase the number of iterations, increase the generalization ability, and prevent over-fitting. Keep the product of the two basically unchanged, but if the step size is set too small, the fitting effect will become
worse, so the learning rate should be appropriately reduced. Taking AUC as the performance index, Table 2 shows the final optimal parameter.

| Table 2. Best parameter of GBDT |
|----------------------------------|
| Best parameter of GBDT          | Value |
| n_estimators                    | 220    |
| learning_rate                   | 0.1    |
| max_depth                       | 11     |
| min_samples_leaf                | 2      |
| min_samples_split               | 2      |

3.6. Feature importance analysis

According to the split characteristics of the gradient boosting decision tree, a feature importance score map is made. In Figure 1, the proportion of stop words and the proportion of punctuation in the grammatical features are the two most critical features for short text filtering in social networks. From the box plot in Figure 1, it can be seen that most of the event sentence text data is short, concise, and accurate. The non-event sentences are not precise enough or a little verbose.

Figure 1. Feature importance score map (left) and the proportion of stop words and punctuation in noise sentences and event sentences (right)

Figure 2. The number of followers, friends and statues in noise sentences and event sentences

The number of followers, statues, and friends in the user's background characteristics are the more critical features, which can determine the credibility of the text data publisher, the audience range, and the user's influence on the social network. It can be seen from the Figure 2 box diagram that the background characteristics of the users who posted the event sentence: the number of users posting more, the number of fans is more, and the number of followers is slightly less.

4. Conclusion

This paper proposes a short text filtering model for social networks. From the perspective of user background, text syntactic features, and text influence, the extracted features can provide comprehensive
and specific prior knowledge for the generation, dissemination, and influence of short text data in social networks. This information plays an important role in the classification and has a good ability to distinguish useless or useful information of social network short text data. Based on the extracted features and small sample training, the GBDT algorithm is selected for classification, which achieves a good classification effect and provides effective input data for event detection.

References
[1] LU T Q, SHI B, YAN Z M, et al. Semi-supervised learning algorithm applied to microblog rumors detection [J]. Application Research of Computers, 2016, 33(3):744-748.
[2] XIE B L, JIANG S Y, ZHOU Y M, et al. Misinformation detection based on gatekeepers' behaviors in microblog [J]. Chinese Journal of Computers, 2016, 39(4):730-744.
[3] DUAN D G, WANG C S, HAN Z M, et al. A rumor detection model based on Weibo' reviews [J]. Computer Simulation, 2016, 33(1) :386-390.
[4] JIANG S Y, CHEN D Y, PANG G S, et al. Research review of information credibility analysis on microblog [J]. Library and Information Service, 2013, 57(12): 136 - 142.
[5] DAI W, WANG Y, JIN Q, et al. An integrated incentive framework for mobile crowdsourced sensing [J]. Tsinghua Science and Technology, 2016, 21(2):146 -156.
[6] ZHANG X H, LI G L, FENG J H. User topic aware task assignment in large data group computing [J]. Journal of Computer Research and Development, 2015, 52 (2) : 309-317.
[7] CHEN Z, FU R, ZHAO Z, et al. gMission: a general spatial crowdsourcing platform [J]. Proceedings of the Very Large Data Base Endowment, 2014, 7(13): 1629 - 1632.
[8] Friedman JH. Greedy function approximation: A gradient boosting machine [J]. Ann Stat, 2001, 29(5): 1189-1232.
[9] LONG C, WONG R C-W, YU P S, et al. On optimal worst-case matching [C] / / SIGMOD '13: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2013: 845 — 856.
[10] ZHANG X H, LI G L, FENG J H. User topic aware task assignment in large data group computing [J]. Journal of Computer Research and Development, 2015 ,52 (2): 309 -317.
[11] Alibaba Group Holding Limited; Patent Issued for Method and System for Text Filtering (USPTO 9600570) [J] Journal of Engineering, 2017
[12] LIN Wei. An improved feature selection method in Chinese microblog public opinion classification [J]. Journal of People' s Public Security University of China (Nature science edition) 2017, 23(2): 72-75.
[13] ZENG Hui, TANG Jiali, XIONG Liyan, et al. Personal social relation extraction in Chinese based on feature selection of CHI, verb and noun [J]. Application research of computers, 2017, 34(6): 1631-1635.