2022 International Conference on Electronics and Devices, Computational Science (ICEDCS) | 978-1-6654-5541-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICEDCS57360.2022.00065

Cyberbullying detection and classification with improved IG and BiLSTM

Mengtian Xin
Northeastern University
Shenyang, China
xmt2000@126.com

Jiayu Shen
Northeastern University
Shenyang, China
shenjiayu08@126.com

Peifeng Hao*
Northeastern University
Shenyang, China
haopeifeng@swc.neu.edu.cn

***Abstract*—Twitter is a microblogging and social networking service platform; users can post what they feel and think about to share with others. Although it facilitates users' social behaviour, a high degree of freedom of speech also leads to cyberbullying. The statement released by UNICEF showed that 36.5% of middle and high school students experienced cyberbullying, and 87% observed cyberbullying. Cyberbullying has greatly affected people's daily lives. We conduct this study to detect whether online comments contain cyberbullying behaviours and classify cyberbullying to alleviate this problem. This paper uses an improved information gain algorithm for feature selection, and the bidirectional LSTM neural network is used for classification. On the premise that the information gain threshold is limited to 0.0004, the precision on the test set can reach 95.15%.**

Keywords—cyberbullying, bidirectional LSTM, information gain

I. Introduction

With the development of information and communication technologies, interpersonal relationships have acquired a new medium for establishing communication. The vast majority of citizens rely on this basic medium for everyday communication, but the ubiquity of social media means that cyberbullying can effectively affect anyone, anytime, anywhere. As discussed in, cyberbullying is defined as "the use of digital technology to cause harm or bullying repeatedly", and cyberbullying can have a deeper and more lasting impact than physical bullying, so automated cyberbullying monitoring has far-reaching implications; this question supports our research.

Based on Twitter comments, this study first conducts cyberbullying detection and classifies cyberbullying types, such as religious discrimination, age discrimination, racism, and sexism. Our research uses the improved information gain based on mutual information for feature selection, which reduces the influence of the unbalanced number of samples between classes, extracts the most informative words, then uses bidirectional LSTM for learning, and finally completes cyberbullying detection and classification. In this study, improved information gain algorithm and bidirectional LSTM bring Precision, Recall and F1-score above 95%.

This study tested the effect of different data preprocessing methods on precision, recall and F1-score. At the same time, the information gain threshold setting and the train/validation/test set ratio are tested. It is concluded that when lemmatize is used, the information gain threshold is set to 0.0004. The ratio of the training set, test set and validation set is 8:1:1, and the overall precision of the model is the highest, reaching 95.15%. Finally, the superiority of our model is verified by comparing our model with Naive Bayes, Naive Bayes with information gain, and unidirectional LSTM with information gain.

II. Related Work

Sentiment analysis uses natural language processing (NLP), text analysis, and computational techniques to extract or classify sentiments from sentiment reviews automatically. Analysis of these sentiments and opinions has spread to many domains, such as social, consumer reviews, and so on. However, the first two cannot make full use of contextual information, so sentiment analysis based on deep learning is more popular. According to the different granularity of division, sentiment analysis tasks can be divided into the levels of words, phrases, sentences, chapters, entities, etc[1].

Yongfeng Zhang et al. pointed out that the current commonly used method for building contextual sentiment lexicons in phrase-level sentiment analysis assumes that the numerical star ratings of reviews represent the overall sentiment direction of the review text, which is not necessarily correct. Therefore, they used a novel constrained convex optimization framework to leverage the results given by review-level sentiment classification to improve phrase-level sentiment polarity labels in the contextual sentiment dictionary building task. The results showed that the framework improved the accuracy of sentiment polarity annotation by 5.6%[2].

Nikos Engonopoulos et al. introduced a new approach, ELS, for entity-level sentiment classification using sequence modelling of conditional random fields (CRF). Due to its sequential nature, CRF classifiers performed better than the common bag-of-words approaches[3].

Orestes Appel et al. proposed a hybrid approach to the problem of sentiment analysis at the sentence level, which used NLP essential techniques, a sentiment lexicon enhanced with the assistance of SentiWordNet, and fuzzy sets to estimate the semantic orientation polarity and its intensity for sentences. The authors compared the obtained results with those obtained using Naive Bayes and maximum entropy techniques and obtained more accurate and precise results [4].

Automatic cyberbullying detection is a task of increasing interest, especially in the natural language processing and machine learning communities. It is challenging, but it is also a relevant need, given how social networking has become an important part of one's life and the dire consequences of cyberbullying, especially among teens. Through an in-depth analysis of 22 studies on automatic cyberbullying detection, H. Rosa et al. showed that cyberbullying was often misrepresented in the literature, resulting in inaccurate

978-1-6654-5541-1/22/$31.00 ©2022 IEEE
DOI 10.1109/ICEDCS57360.2022.00065

systems with little real-world application. In the article, the authors aimed to direct future research on the subject towards a more coherent viewpoint with the definition and representation of the phenomenon so that future systems can have practical and impactful applications. A series of recommendations were also made for future work[5].

Maral Dadvar et al. propose that recent research on cyberbullying detection has mainly focused on the content of conversations while largely ignoring the characteristics of actors involved in cyberbullying. Therefore, the authors used an SVM model to train a gender-specific text classifier. The results showed that considering gender-specific linguistic features can improve the discrimination capacity of the classifier to detect cyberbullying. In another paper, the authors showed that taking user context into account improves cyberbullying detection. The results showed that the detection accuracy based on user context has a 5% improvement over the baseline, reaching 77%.

Mohammed Ali Al-garage et al. also developed a supervised machine learning solution for cyberbullying, using the Twitter dataset for cyberbullying prediction, and the results showed that the area under the receiver operating characteristic curve was 0.943 and the f-measure was 0.936. The proposed model based on these features provided a feasible solution for cyberbullying detection in online communication environments.

Vinita Nahar et al. proposed that most current cyberbullying detection methods are static, and they cannot effectively handle noisy, imbalanced or streaming data. Therefore, they proposed a semi-supervised learning method to augment the training data samples and apply a fuzzy SVM algorithm. The augmented training technique automatically extracted and expanded the training set from unlabeled streaming text. Experimental results showed that the proposed enhancement method performed better than all other methods.

The dataset used in this paper is from the paper by Jason Wang et al. They established a framework for automatically generating balanced data. At the same time, the authors also proposed a Graph Convolutional Network (GCN) classifier. The authors compared GCN models on datasets of two sizes. The results showed that the GCN model matched or exceeded the performance of the baseline model.

Of course, there are many difficulties in sentiment analysis at this stage. Doaa Mohey El-Din Mohamed Hussein pointed out that there are also several challenges in sentiment analysis and evaluation, which become obstacles to analyzing the accurate meaning of emotions and discovering appropriate emotional polarity. The authors explored the importance and impact of the sentiment analysis challenge in sentiment assessment through pairwise comparisons of 47 papers.

III. Proposed Model

A. Preprocession

The first process in the workflow is data preprocessing. Since this dataset collects part of the comment data on Twitter, there may be frequent problems such as "@a user", wrong words, wrong punctuation and so on in the sentences. Therefore, given the above characteristics, the data preprocessing methods in this processing stage mainly include six types: masking user names, modifying cases, lemmatizing, deleting punctuation marks, correcting error words and removing stopwords. Among them, we use WordNetLemmatizer in the nltk Library to achieve lemmatization and, at the same time, build a vocabulary table and a stopword table for the latter two text preprocessing.

B. Feature Selection

The next process in the workflow is feature selection. In general, if a term appears repeatedly in all categories, the term is not considered to affect text classification. Conversely, a term is useful in text classification if it only appears in one category and rarely in other categories. The more concentrated the distribution of feature items among categories are, the greater the value obtained. The more uniform the distribution of feature items among the categories, the smaller the value obtained. The calculation formula of the information gain $\mathrm{IG(t)}$for the term t is:

$$\mathrm{IG(t)} = \mathrm{H(C,t)} - \mathrm{H(C|t)}$$

$$= -\sum_{\mathrm{i=1}}^{|\mathrm{c}|} \mathrm{P(c_i)log_2P(c_i)}$$

$$+ \mathrm{P(t)}\sum_{\mathrm{i=1}}^{|\mathrm{c}|} \mathrm{P(c_i|t)log_2P(c_i|t)}$$

$$+ \mathrm{P(\bar{t})}\sum_{\mathrm{i=1}}^{|\mathrm{c}|} \mathrm{P(c_i|\bar{t})log_2P(c_i|\bar{t})}$$

where $\mathrm{c_i}$ represents category i cyberbullying; $\mathrm{P(c_i)}$ is the proportion of data items with category i in the training set; $\mathrm{P(t)}$ is the proportion of data items in the corpus where the term t appears; $\mathrm{P(c_i|t)}$ is the proportion of data items in the corpus where term t appears and belongs to category i cyberbullying; $\mathrm{P(\bar{t})}$ is the proportion of data items in the corpus where the term t does not appear, and satisfies $\mathrm{P(\bar{t})} = 1 - \mathrm{P(t)}$; $\mathrm{P(c_j|\bar{t})}$ is the proportion of data items in the corpus where term t does not appear and belongs to category i cyberbullying.

However, there are certain problems in the calculation of such information gain. When the distribution of samples in different categories of the corpus is unbalanced, the information gain value obtained from a small number of sample sets will be very small. At this time, information gain has a negative impact on feature selection. Therefore, based on the original information gain model, we use information theory. An improved information entropy algorithm is used to measure the degree of concentration of terms between classes. The specific calculation steps are as follows:

① Calculate $\mathrm{H(C,t)}$

$$H(C,t) = -\sum_{i=1}^{|c|}\left(\frac{f_{c_i}(t)}{A_i} log_2 \frac{f_{c_i}(t)}{A_i}\right)$$

where $\mathrm{f_{c_i}(t)}$ is the frequency of term t in category i cyberbullying corpus; $\mathrm{A_i}$ is the total number of terms in category i cyberbullying corpus.

② Standardization of $\mathrm{H(C,t)}$

The value range calculated above is $[0, \mathrm{log_2|c|}]$, and we standardize it:

$$a = \frac{H(C,t)}{\log_2|c|}$$

③ Factor inversion

After normalization, we can find that the more concentrated the distribution of term t among classes, the smaller the value of term t(closer to 0); The more evenly the term t is distributed among classes, the greater the value of the term t(closer to 1). This is negatively correlated with the desired result, so we perform a factor inversion operation:

$$b = 1 - a$$

④ Calculate IG(t)

Combined with the above, we can get the information gain calculation formula of term t:

$$IG(t) = b[p(t)\sum_{i=1}^{|c|} p(c_j|t)\log_2\frac{p(c_j|t)}{p(c_j)} + p(\bar{t})\sum_{i=1}^{|c|} p(c_j|\bar{t})\log_2\frac{p(c_j|\bar{t})}{p(c_j)}]$$

IV. Experimental Results

In this section, we compare the results of 4 scenarios. We use precision, recall, and F1-score as outcome measures. In the first scenario, we explore text preprocessing methods to determine the impact of different preprocessing methods on the experimental results. The second scenario explores the effect of the information gain threshold on feature selection and the final result. The third scenario selects the appropriate training set, validation set, and test set ratio. The last scenario compares Naive Bayes without Information Gain, Naive Bayes with Information Gain, Unidirectional LSTM with Information Gain, and Bidirectional LSTM with Information Gain to verify the superiority of our model.

A. The Experiment On The Effect Of Using Data Preprocessing

In this scenario, we designed two experiments to explore the impact of different text preprocessing methods on precision, recall and F1-score and find the optimal text preprocessing method for subsequent experiments. First, we compared the results without removing expressions, abbreviations, stopwords, and hashtags when no morphological normalization methods were used. The result is shown in Table I. Next, using the above four processing methods, we compare the results of using lemmatization, using stemming and the case where neither is used. The result is shown in Table II. In this scenario, the ratio of the training set, validation set, and test set of 8:1:1 is used, the information gain threshold is 0.0001, and the bidirectional LSTM is used for experiments.

TABLE I. Experiment results on the effect using data preprocessing

Data Preprocessing I	Precision	Recall	F1-score
Without removing emoji	0.9270	0.9275	0.9272
Without removing stopwords	0.8936	0.8930	0.8932
Without removing hashtag	0.9256	0.9253	0.9254
Without decontracting	0.9255	0.9252	0.9251
Removing emoji, stopwords, hashtag and decontracting	0.9299	0.9292	0.9294

TABLE II. Experiment results on the effect using morphological normalization

Data Preprocessing II	Precision	Recall	F1-score
without morphological normalization	0.9299	0.9292	0.9294
using lemmatize	0.9454	0.9450	0.9452
using stemming	0.9391	0.9389	0.9389

As can be seen from Table I, without morphological normalization, whether to remove stopwords has the greatest impact on the Precision, Recall and F1 score of the experimental results because these words can hardly help to understand the whole sentence; this low-level information is removed from the text so that the model can focus on more important information. Less influential than the former is whether to remove abbreviations and whether to remove hashtags related to the characteristics of the Twitter dataset. Twitter uses hashtag tags to tag social media content, allowing social media users to find content related to common topics or interests, but Hashtags do not provide enough information for our classification of cyberbullying.

Table II shows that using lemmatize works best for this problem. Using lemmatize as the morphological normalization method, combined with the four preprocessing methods, achieved higher Precision, recall and F1-score, 94.54%, 94.50%, and 94.52%, respectively.

B. The Experiment On Using Different IG Threshold

In the second scenario, different information gain thresholds are set to determine the impact of the threshold on feature selection and the final result. In this scenario, lemmatize and 4 other data preprocessing methods are used; the ratio of the training set, validation set, and test set of 8:1:1, and bidirectional LSTM is used for experiments. The results are shown in Table III.

TABLE III. Experiment results on the effect using different IG threshold

Threshold	Precision	Recall	F1-score
0	0.9354	0.9341	0.9346
0.0001	0.9454	0.9450	0.9452
0.0002	0.9433	0.9430	0.9430
0.0003	0.9463	0.9461	0.9459
0.0004	0.9515	0.9514	0.9514
0.0005	0.9486	0.9477	0.9479

The effect of the information gain threshold at the feature selection on six experiments ranging from 0 to 0.0005 can be seen in Table III.

Precision, Recall, and F1-score generally increase as the information gain threshold increases. This is because the higher the feature selection threshold, the remaining features will contain more useful information for classification. When the threshold is set higher than 0.0004, the overall accuracy will decrease since the threshold is too high, so that feature items left for classification are too few. The best results were obtained using an information gain threshold of 0.0004, with Precision, recall, and F1-score of 95.15%, 95.14%, and 95.14%, respectively.

C. The Experiment On Using Different Train/Validation/Test Set Ratio

In the third scenario, we tested different ratios of training set, validation set, and test set, in order to determine the combination that will yield the highest precision, recall and F1-score. In this experiment, lemmatize and 4 other data preprocessing methods are used, the information gain threshold is set to 0.0001, and the bidirectional LSTM is used for experiments. The results are shown in Table IV.

TABLE IV. EXPERIMENT RESULTS ON THE EFFECT USING DIFFERENT RATIOS

Ratio	Precision	Recall	F1-score
8:1:1	0.9454	0.9450	0.9452
7:2:1	0.9409	0.9402	0.9404
6:2:2	0.9413	0.9406	0.9407
7:1:2	0.9417	0.9413	0.9413

Four experiments were conducted with different proportions of training set, validation set, and test set respectively, and the results are shown in Table IV. Using the 8:1:1 ratio yielded the best Precision, Recall, and F1-score results of 94.54%, 94.50%, and 94.52%, respectively.

D. The Experiment On Using Different Models

In the fourth scenario, we compare the bidirectional LSTM model with the naive Bayes model, the naive Bayes model using information gain for feature selection, and the unidirectional LSTM model to determine the optimal model. In this experiment, lemmatize and 4 other data preprocessing methods are used, the information gain threshold is set to 0.0004, and the ratio of the training set, validation set, and test set of 8:1:1 is used. The results are shown in Table V.

TABLE V. EXPERIMENT RESULTS ON THE EFFECT USING DIFFERENT MODELS

Model	Precision	Recall	F1-score
Naive Bayes(without information gain)	0.8481	0.8392	0.8211
Naive Bayes(with information gain)	0.8658	0.8647	0.8534
Unidirectional LSTM(with information gain)	0.9416	0.9418	0.9415
Bidirectional LSTM(with information gain)	0.9515	0.9514	0.9514

From the results in Table V, it can be seen that even the naive Bayes model using information gain as feature selection, its Precision, Recall and F1-score only achieved 86.58%, 86.47%, 85.34% results, while our proposed bidirectional LSTM with information gain achieved the best results, with Precision, Recall and F1-score of 95.15%, 95.14%, and 95.14%, respectively. The reason is that although the Naive Bayes model has strong interpretability and fast calculation speed, as a probabilistic classifier, it is highly dependent on prior knowledge. If the data is not representative or has an imbalance, it will lead to poor text reasoning ability. Furthermore, it assumes that features are independent of each other, which means that the contribution of lexical features in all sentences is equal, regardless of their relative position in the text, while our bidirectional LSTM model can learn the semantics of words in context and maintain contextual memory associations.

V. CONCLUSION AND FUTURE WORK

Our paper aims to establish an effective fine-grained network bullying classifier to identify whether there is network bullying and the category of network bullying. According to the experimental results, it can be concluded that the data preprocessing process, information gain threshold and train/validation/test set ratio have a great impact on the optimal results of the classification process. From the experimental results, using lemmatize combined with four preprocessing methods, the information gain threshold is set to 0.0004, the train/validation/test set ratio is set to 8:1:1, and the optimal results can be obtained, with precision, recall and F1-score reaching 95.15%, 95.14% and 95.14% respectively.

At this stage, our research focuses mainly on the content of Twitter comments, ignoring the characteristics of cyberbullying participants. In the future, we can mine other aspects of user information, such as frequently viewed content, preferences, etc., for more accurate cyberbullying detection.

REFERENCES

[1] Habimana, O., Li, Y., Li, R., Gu, X., & Yu, G. (2020). Sentiment analysis using deep learning approaches: an overview. Science China Information Sciences, 63(1), 1-36.

[2] Zhang, Y., Zhang, H., Zhang, M., Liu, Y., & Ma, S. (2014, July). Do users rate or review? Boost phrase-level sentiment labeling with review-level sentiment classification. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval (pp. 1027-1030).

[3] Engonopoulos, N., Lazaridou, A., Paliouras, G., & Chandrinos, K. (2011, May). ELS: a word-level method for entity-level sentiment analysis. In proceedings of the international conference on web intelligence, mining and semantics (pp. 1-9).

[4] Appel, O., Chiclana, F., Carter, J., & Fujita, H. (2016). A hybrid approach to the sentiment analysis problem at the sentence level. Knowledge-Based Systems, 108, 110-124.

[5] Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., ... & Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. Computers in Human Behavior, 93, 333-345.