Reddit sentiment analysis for natural language processing

Ang li

Hanyang University, Department of Computer Science, 15588, Seoul, South Korea

angli96116@gmail.com

Abstract. In the Internet age, social media has fully penetrated into people's lives. As one of the well-developed online platforms with a large user base, Reddit allows users to independently publish current news, life experiences, and interesting life stories. However, sometimes it sends a negative tone that affects the brand of a company or individual and destroys profits and it is necessary to prevent Twitter by identifying hate words. The biggest innovation of this post is that we use reddit data to compare various methods simultaneously. As we process more data, trying deep learning will yield good results. Compared to other machine learning classifiers, the transformer classifier achieves the best results.

Keywords: Transformer Classifier, NLP, VADER, Sentiment Analysis, API, Python, Polarity, Evaluation Metrics Introduction.

1. Introduction

Reddit is a social news site with tons of information. The categories of content on the site are called "subreddits," which include news, video games, movies, music, books, fitness, food, and photo sharing. In 2012, Reddit performed extremely well, with more than 30 million news items, 37 billion page views, and more than 40 million unique visitors[1]. Individuals and organizations have also increased their social media content from different fields. People have less need to rely on the opinions of their friends and family because users post numerous reviews on the internet about various products and services.

To protect the peace and stability of our society, we need a safe and secure online environment. As we can see now, most social media focuses on hate language. Because it creates a hostile and hostile atmosphere that sometimes leads to conflict and violence, making the modern world even worse. The online environment is full of articles attacking groups and individuals. We cannot limit people's liberal values, especially freedom of speech. Therefore, we must have an opinion and find a way to reduce the occurrence of the impact of hate speech that is the subject of this article.

After the Boston Marathon bombing in 2013, Reddit was criticized for mistaking several people for user suspects. One of those involved was mistaken for student Sunil Tripathi, who went missing before the blast. The body was found in the Rhode Island River on April 25, 2013. Authorities believe the cause of death did not include illegal acts. His family later confirmed to outsiders that his cause of death was suicide. Reddit general manager Eric Martin later apologized for the user's actions and criticized the site for "online witch hunts and dangerous speculation.

To avoid sending hate speech, it is necessary to detect hate so that it cannot be sent. In our research, we found that the accuracy of using deep learning was more experimental than using machine learning, so we continued to investigate hate speech detection in Reddit through deep learning.

2. Related work

2.1. SVM

Support Vector Machine (SVM) plays a pivotal role in machine learning and is a supervised classification learning algorithm[2]. Vapnik and others first proposed the relevant theory in 1922. Developed and updated over 30 years, widely used in many reseaches. In theory, SVM can solve any kind of binary classification problem.

2.2. Naïve Bayes Classifier

The general idea of Naive Bayes is based on probabilistic models, first introduced in 1960 by Maron and Kuhns [3]. A Bayesian classifier is a generative model that performs classification by calculating probabilities. It can be used to deal with multi-classification problems, and it also performs well for small-scale data prediction. Bayesian classifiers are suitable for multi-classification tasks and incremental training. For large-scale data, the computational complexity is low, and the algorithm principle is relatively simple and easy to understand.

2.3. Logistic Regression

Logistic Regression is not closely related to regression, but has repeatedly appeared in classification problems, it is a classification model that is often used for binary classification[4]. Logistic Regression is widely used in the industry because it is easy to understand, can be parallelized, and has strong interpretability.

2.4. Random Forest

Random forest is an easy-to-use and widely used machine learning algorithm that can achieve good results without hyperparameter tuning. Random Forest was proposed by Tin Kam Ho of Bell Labs in 1995 [5]. It is used for regression and classification tasks.

2.5. Transformer Classifier

Transformer (a network structure) was proposed by Google in 2017[6]. Before Transformer was proposed, the RNN-type sequence network structure was commonly used in the NLP field to process text data. In addition, the CNN network, due to its powerful ability to extract local information, also had a place in the NLP field. Transformer is a network structure different from RNN and CNN. It completely adopts a self-attention mechanism. There is no sequence form in Transformer, but a positional form is used to represent text data.

3. Data collection

Reddit is one of the most influential social interactive new media today. First, a lot of data involves user privacy, which authorizes developers to obtain part of the data on the platform in a specific way for research and analysis by developers. At present, there are two main ways to crawl Weibo comments: one is to capture through the API interface provided by the official Weibo, and the other is to obtain through the analysis of Weibo web pages. The second method can be divided into 3 basic steps: 1. Select a suitable browser, here Google is used, and use the website of the mobile terminal for Weibo-related access; 2. Decompose the obtained request list, and find the source url of the structured json. Come out; ③Analyze the url law and write the corresponding python code.

There are a total of 27,500 reddit comments crawled in this article, but the reddit comment text value density is low, and the repetition is manually filtered out manually. Give comments a positive, negative or neutral class label, treat negative class label comments as hate comments. At the same time, prepare stop word documents that do not contain emotional characteristics to prepare for subsequent emotional analysis.

4. Sentiment analysis

The process of sentiment analysis includes five parts: feature extraction, feature dimensionality reduction, feature representation, division of training set and test set, and classification algorithm to build a classifier. The classification algorithm includes five types: Naive Bayes classifier, Random Forest, Logistic Regression, SVM, and Transformer. Using the idea of the control variable method, it is judged under which combination the classifier constructed has the highest accuracy.

4.1. Sentiment Analysis Process

Features are actually local features displayed by classified objects. important basis for classification. In practice, in order to obtain the best classification results, it is necessary to conduct multiple experiments on the selection method of features, the definition of feature weights, and the content of selected features, and compare and analyze them by controlling variables. The best experimental results are as follows.

But there will also be features that are completely unrelated to classification judgment, and some even mention for misleading, the classification results will be biased. Also, it's not easy to find all the features in the reviews. It is not possible to select all features objectively. Subjectively, only some features can be selected as features. This is a major feature of feature selection and requires human-supervised selection.

Dimensionality reduction refers to reducing the number of random variables under certain constraintsThe main variables are irrelevant, and feature dimensionality reduction has the following two meanings: (1) By reducing the number of features, the operation speed of each algorithm is accelerated, and the program efficiency is improved; (2) The information content is large, and high features are extracted to achieve the effect of noise reduction.

4.2. Divide data set

After characterizing the text, it is divided into training set and test set. In order to make the data processing more concise and clear, the first 500 texts are selected as the test set, and the rest are divided into the training set. The training set is used to train the algorithm to obtain a classifier; the test set is to use the trained classifier for classification, and label each piece of data with a class label, and then compare the manually labeled label with the classification prediction result, This gives the accuracy of the classifier. Therefore, the training set and the test set do not interfere with each other but are closely connected in the experiment. The training set does not directly affect the test set, but the use of the classifier of the training set on the test set is the key to the experiment.

4.3. Classification Algorithms Build Classifier

Build a classifier with different classification algorithms in machine learning in the training set, and use the pre-divided test set to test the accuracy of the classifier, so the test set is also called the prediction set, and the best classification algorithm is selected at the same time.

In this experiment, 5 methods including Naive Bayes, Logistic Regression and Support Vector Machine are used.

5. Results

Table 1 presents the recall and precision results obtained by the five versions of the domain-specific classifier. The study found that when the machine learning model considered all metrics (precision: 70.20%, recall: 70.20%, F1 score: 70.0%), the model trained with the random tree classifier showed better performance than other machine learning models. good feature. This is not far from the accuracy of several other machine learning models.

But when the deep learning Transformer classifier model was compared with them, all the indicators (precision: 84.54%, recall: 84.54%, F1 score: 84.35%) of the transformer classifier model were higher than those of the machine learning model, which greatly improved the performance.

Table 4. A slightly more complex table with a narrow caption.

	Accuracy	*F1 Score*	*Recall Score*
SVM	65.60%	65.60%	65.60%
Naïve Bayes	64.00%	63.60%	64.00%
Logistic Regression	69.10%	69.00%	69.10%
Random Forest	70.20%	70.00%	70.20%
Transformer	84.54%	84.54%	84.54%

6. Conclusion

In this experiment, NLP-based sentiment analysis was used to analyze and evaluate hate speech in Reddit data. In this model, the performance is greatly improved by using the Transformer classifier. Future plans are to evaluate the model on other social media datasets. An important aspect worth considering is that these findings will provide a direction for future research that may benefit aspect-based sentiment classification performance.

References

[1] Pang B, Lee L.Opinion Mining and Sentiment Analysis[j].Foundations and trends in Information Retrieval,2008,2(1-2):1-135Pang B, Lee L.Opinion Mining and Sentiment Analysis[j].Foundations and trends in Information Retrieval,2008,2(1-2):1-135

[2] Idicula-Thomas S,Kulkarni, A J,Kulkarni B D,et al.A Support Vector Machine-based Method for Predicting the Propensity of a Protein to be Soluble or to Form Inclusion Body on Overexpression in Escherichia Coli[J].Bioinformatics, 2006(22):278-284.

[3] Alharbi, A.S.M.; de Doncker, E. Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information. Cogn. Syst. Res. 2019, 54, 50–61. [CrossRef]

[4] Abid, F.; Alam, M.; Yasir, M.; Li, C.J. Sentiment analysis through recurrent variants latterly on convolutional neural network of Twitter. Future Gener. Comput. Syst. 2019, 95, 292–308. [CrossRef]

[5] Pouli, V.; Kafetzoglou, S.; Tsiropoulou, E.E.; Dimitriou, A.; Papavassiliou, S. Personalized multimedia content retrieval through relevance feedback techniques for enhanced user experience. In Proceedings of the 2015 13th International Conference on Telecommunications (ConTEL), Graz, Austria, 13–15 July 2015; pp. 1–8.

[6] Kraus, M.; Feuerriegel, S. Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees. Expert Syst. Appl. 2019, 118, 65–79. [CrossRef]