Profiling Irony and Stereotype Spreaders on Twitter with BERT

Yifan Xu[1], Hui Ning[1,*]

[1]*Harbin Engineering University, Harbin, China*

Abstract

This paper summarises the participation at the "Profiling Irony and Stereotype Spreaders on Twitter" shared task at PAN at CLEF 2022, and proposes a method which can detect irony and stereotype spreaders automatically. We detect whether a user is a irony and stereotype spreader instead of detecting a single content. In this paper, we use BERT embeddings and autogluon which can automates classic machine learning methods to train a classifier.We upload the forecast results to TIRA[1] Platform. Using our method, an accuracy of 94.3 % is achieved on the English training set. On the English test set, our system achieved an accuracy result of 94.4 %.

Keywords

Irony and stereotype, Twitter, Autogluon, BERT

1. Introduction

With the development of the Internet, social media has become an important medium which people use it to communicate. Information spreads widely and quickly on social media. However, with the popularity of social media, some problems have gradually emerged. Irony and stereotype spreaders on Twitter is one of them.

With irony, language is employed in a figurative and subtle way to mean the opposite to what is literally stated. In case of sarcasm, a more aggressive type of irony, the intent is to mock or scorn a victim without excluding the possibility to hurt. Stereotypes are often used, especially in discussions about controversial issues such as immigration or sexism and misogyny.

In this paper, we perform irony and stereotype spreaders identification from Twitter data[2], provided by the organizers of PAN'22. At PAN'22[3], we focus on profiling ironic authors in Twitter. Special emphasis will be given to those authors that employ irony to spread stereotypes, for instance, towards women or the LGBT community. The goal will be to classify authors as ironic or not depending on their number of tweets with ironic content.

In Section 2 we present some related work on profiling irony and stereotype spreaders. In Section 3 we describe the method proposed. In Section 4 we present the experimental results achieved. Finally, in Section 5, we present the conclusions and future work.

*Corresponding author

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

xvyifan@hrbeu.edu.cn (Y. Xu); ninghui@hrbeu.edu.cn (H. Ning)

2. Related Work

There are few researches on profiling irony and stereotype spreaders. But we can find many researches on hate speech detection. Both of them have a lot in common. Though a lot of sentences with irony don't contain rude words, they can hurt people deeply. Irony is another way to attack other people.

Detection of hate speech has been popular research in recent years. We can consider this problem as a text classification task. Researchers usually extract different types of features and exploit them in combination with the techniques of Machine Learning. There are various kinds of classifiers used for this: Naive Bayes in combination with a Bag-of-Words approach[4]; Support Vector Machines, again applied on Bag-of-Words features[5]; Logistic Regression, trained, for instance, on N-grams[6]. Besides these methods, Deep Learning techniques have also been used in this problem. In many studies, we can find Deep Learning Models such as Recurrent Neural Networks (RNN)[7] and Convolutional Neural Networks (CNN)[8].

These days, transformer[9] has been the most popular Deep Learning technique in tasks of Natural Language Processing (NLP)[10]. Using transformer such as BERT[11], the researchers have achieved good results in classification tasks such as profiling hate speech and fake news detection.

In the 2021 PAN[12] shared task, Profiling Hate Speech Spreaders on Twitter[13], there was a variety of methods used for classification, preprocessing, and feature selection, such as SVM[14], LSTM[15], Naive Bayes, BERT and RoBERTa[16].

With reference to the above research contents about hate speech detection, we can apply their methods to our experiments on profiling irony and stereotype spreaders.

3. Method

This section will introduce the datasets, data preprocessing and system for identifying irony and stereotype spreaders.

3.1. Datasets

The datasets for this task are given in English. There are 420 authors in the train dataset and 180 authors in the test dataset. Half of all the authors in both sets were labeled with "I" indicating that the author spreads irony and stereotype, while the other half in both sets was labeled with "NI" to indicate that the author does not spread irony and stereotypes. Data is provided by the organizers of PAN'22, which is collected on the Twitter. Each author has 200 unique tweet posts. There are a total 12000 tweet posts in the dataset.

3.2. Preprocessing

Data preprocessing is used to remove noise. Links, user mentions, hashtags, and retweets were removed. All punctuation and all numbers were removed. After that, we convert tweets to lower case. Stop-words were retained. We set labels of authors as the label of their each tweet.

Table 1
Data after Preprocessing

Text	Label
billion tshirt ngotta be some in the...	I
The simple answer is usd just like...	I
I honestly boggle at the very existence...	I
Why would it ath means nothing...	I

3.3. Training

We didn't finetune the BERT model on the data. Instead of that, we used the BERT embeddings from the BERT model. We used the embeddings extracted from the last hidden layer and the last four hidden layers to extract features of the data. In that way, we can find which layer can achieve a better result. After extracting tweets features, we average them to author features. Then, we send author features to autogluon[17]. AutoGluon automates machine learning tasks which can easily achieve strong predictive performance in applications. Using autogluon, we can find the best model for the classification. For the training, we used a 5-fold cross-validation.

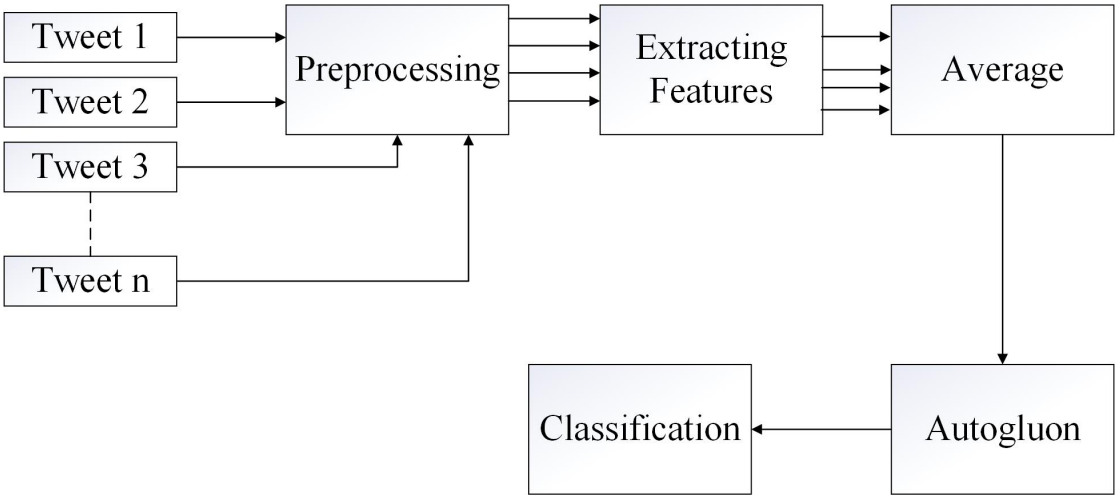

Figure 1: Irony and stereotype classification

4. Results

According to the requirements of PAN'22, we use accuracy as the evaluation. From Table 2, we can get the results. For the training set, the accuracy is 94.3 % using embeddings from the last hidden layer of BERT. We achieved a accuracy of 94.0 % by the last four hidden layers. For the testing set, the result is 93.3 % and 94.4 %.

Table 2
Results

Embeddings	Accuracy (Train Set)	Accuracy (Test Set)
The last	94.3 %	93.3 %
The last four	94.0 %	94.4 %

5. Conclusion

The emergence of social networking sites has epoch-making significance for the whole Internet industry. Because it brings the real world into the virtual network world, profoundly changes the way where people interact, and greatly improves the efficiency of spreading information. But the social media also triggers some adverse issues such as irony and stereotype spreading. In this paper, we propose a method which can detect irony and stereotype spreaders automatically. Instead of analyzing the single content, the aim is to detect users who tend to publish posts that fall into the category of "irony and stereotype". We used BERT embeddings and autogluon which can automate classic machine learning methods used to classify irony and stereotype spreaders on Twitter. The accuracy we achieved is 94.2% on the training set and 94.4% on the testing set.

References

[1] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA Integrated Research Architecture, in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World, The Information Retrieval Series, Springer, Berlin Heidelberg New York, 2019. doi:10.1007/978-3-030-22948-1_5.

[2] O.-B. Reynier, C. Berta, R. Francisco, R. Paolo, F. Elisabetta, Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO) at PAN 2022, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.

[3] J. Bevendorff, B. Chulvi, E. Fersini, A. Heini, M. Kestemont, K. Kredens, M. Mayerl, R. Ortega-Bueno, P. Pezik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of PAN 2022: Authorship Verification, Profiling Irony and Stereotype Spreaders, and Style Change Detection, in: M. D. E. F. S. C. M. G. P. A. H. M. P. G. F. N. F. Alberto Barron-Cedeno, Giovanni Da San Martino (Ed.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), volume 13390 of *Lecture Notes in Computer Science*, Springer, 2022.

[4] I. Kwok, Y. Wang, Locate the hate: Detecting tweets against blacks, in: National Conference on Artificial Intelligence, 2013.

[5] E. Greevy, A. F. Smeaton, Classifying racist texts using a support vector machine, ACM (2004).

[6] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, SRW@HLT-NAACL (2016) 88–93.

[7] D. F. Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, M. Tesconi, Hate me, hate me not: Hate speech detection on facebook, ITASEC (2017) 86–95.
[8] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, WWW (Companion Volume) (2017).
[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, N. A. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems 30 (NIPS 2017) (2017) 5998–6008.
[10] D. C. Manning, H. Schütze, Foundations of statistical natural language processing, Foundations of statistical natural language processing (2003) 37–38.
[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, north american chapter of the association for computational linguistics (2019).
[12] J. Bevendorff, B. Chulvi, L. D. l. P. G. Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of pan 2021 - authorship verification, profiling hate speech spreaders on twitter, and style change detection, CLEF (2021) 419–431.
[13] F. Rangel, L. D. l. P. G. Sarracén, B. Chulvi, E. Fersini, P. Rosso, Profiling hate speech spreaders on twitter task at pan 2021, CLEF (2021) 1772–1789.
[14] I. Vogel, M. Meghana, Profiling Hate Speech Spreaders on Twitter: SVM vs. Bi-LSTM—Notebook for PAN at CLEF 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2936/paper-196.pdf.
[15] M. Uzan, Y. HaCohen-Kerner, Detecting Hate Speech Spreaders on Twitter using LSTM and BERT in English and Spanish—Notebook for PAN at CLEF 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2936/paper-194.pdf.
[16] T. Anwar, Identify Hate Speech Spreaders on Twitter using Transformer Embeddings Features and AutoML Classifiers—Notebook for PAN at CLEF 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2936/paper-153.pdf.
[17] N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, A. Smola, Autogluon-tabular: Robust and accurate automl for structured data, arXiv preprint arXiv:2003.06505 (2020).