Classifying Misogynistic Tweets Using a Blended Model: The AMI Shared Task in IBEREVAL 2018

Elena Shushkevich, John Cardiff

Social Media Research Group
Institute of Techology Tallaght, Dublin, Ireland
e.shushkevich@yandex.ru, john.cardiff@it-tallaght.ie

Abstract. This article describes a possible solution for Automatic Misogyny Identification (AMI) Shared Task at IBEREVAL-2018. The proposed technique is based on combining several simpler classifiers into one more complex blended model, which classified the data taking into account the probabilities of belonging to classes calculated by simpler models. We used the Logistic Regression, Naive Bayes, and SVM classifiers. The experimental results show that blended model works better than simpler models for all three type of classification, for both binomial classification (Misogyny Identifivation, Target Classification) and multinomial classification (Misogynistic Behavior).

Keywords: Twitter, Tweets Classification, Machine Learning

1 Introduction

Nowaday, Twitter is one of the main public Internet platforms that allow people to learn news in real time, to communicate with people from around the world, to discuss the latest accidents and to express their their own opinions on any information event. The ever-increasing volume of data and some special features of the platform (for example, a limited number of characters in a message or the ability to respond to another user) are an attractive challenge for researchers in various scientific fields, including for solving some propbems connected with text mining area.

A negative aspect of the increased usage of platforms like Twitter is that incidents of aggression and related activities like harassment, misogynism, cyberbullying, etc. have increased significantly. The reach of the Internet has given such incidents unprecedented power and influence to affect the lives of billions of people. The societal problems, for women in particular, are volume and persistence. Thousands of misogynistic tweets can be made by different accounts in seconds, and once published, are almost impossible to remove without trace. It is paramount, from the perspective of social media users and platform providers, that these texts be identified and removed as quickly as possible.

The Automatic Misogyny Identification (AMI) shared Task at IBEREVAL-2018 is focussed on the classification of texts into pre-defined categories based on their degree of misogyny. The challenge is to build three different classifiers that allow the identification of misogynistic behavior. The training data is composed of 3251 tweets in English.

The first challenge involves building a binary classifier that determines whether a tweet is misogynistic or not. In the second challenge, it is necessary to make a more detailed classification of misogynistic tweets. Firstly, we need to determine the target of the message - whether the insult is active (insult a particular person) or passive (insult a group of people). Secondly, it is necessary to classify the type of misogynistic behavior. The following types are assumed: Stereotype & Objectification (widespread negative perception of women), Domination (emphasizing the gender superiority of men over women), Derailing (justification of male abuse of women), Sexual Harassment and Threads of Violence (as well as requests for sexual services) and Discredit (slurring without any other intentions).

This paper presents a possible approach to solve the above problems. The main purpose of the approach is to build a model that allows us to assess the belonging of any tweet to a particular group declared for classification with the best result.

The paper is organized as follows. Some useful researches in the area of texts classification are described in Section 2. Section 3 presents the methodology for building the desired model, and in Section 4 the results are described and analyzed. There is a summary of the work in Section 5.

2 Current work

Today it is difficult to say that the theme of misogyny and harassment is actively used for text processing by machine learning methods. However, there are a number of approaches in this area that deserve special attention. One of them is [7], in which NLP is used to analyze English-language misogynistic tweets to find out how often abusive words are used (a range of misogynistic words were highlighted for the study), who uses them most frequently and what drives traffic of using these words (who determines whether such words are used more often or less?).

In [8] and [9], modelling of classifiers was conducted for the evaluation of the tone of texts (from very negative to very positive). They created classifications for the different number of groups (3, 5 and 8 categories) and compared the accuracy of the results. Using additional tools for modeling and analysis (such as GMDH Shell and Semantic Orientation Calculator (SO-CAL)), the researchers were able to achieve a high accuracy for the constructed classifiers, which exceeded the accuracy obtained using the simple combinatorial algorithm. This conclusion indicates a high potential of using inductive modeling for text mining in future.

3 System

The system for determining whether an object (tweet) belongs to a class (with three different classifications in mind - misogyny or not, target classification and classification by types of misogyny) is described below. The system includes several sequential steps: preprocessing, building classifiers using different methods and assigning a final class to an object.

3.1 Preprocessing

Firstly, we prepared the data for the construction of the classifiers, so cleaned the data: we removed the string punctuation marks and brought the words to the lower case. Next, we used TF - IDF (we used TfidfVectorizer[1]) for vectorization. TF (term frequency) - IDF (inverse document frequency) is a method that increases the weight of frequently occurring in a given document words and reduces the weight of frequently occurring in many documents words.

3.2 Methods

Logistic regression (LR) is a method that works well with high dimension data and allows us to build exponential classifiers for text data analysis. Logistic regression involves the construction of a discriminant model, which calculates the probability from a function of a weighted set of observation features and assigns a class to each observation. The classifier based on logistic regression applies an exponential function to a linear combination of objects obtained from the input data [3, 4].

Naive Bayes classifier (NB) - classifiers built on the basis of this method are based on the use of Bayes theorem with the assumption of independence of events (features). The advantages of the Naive Bayesian classifier are the small amount of training data needed to estimate the parameters required for classification, as well as the speed of calculations [6] (in comparison with more complex methods).

Support Vector Machine classifier (SVM) - the basic idea of this method is to translate the source vectors into a higher dimension space and search for such a separating hyperplane so that the gap in this space is maximal. Two parallel hyperplanes are constructed on both sides of the hyperplane that separates the classes, and one hyperplane that will maximize the distance to two parallel ones is sought. This algorithm assumes that the error in the class definition will be decreasing as the distance between these parallel hyperplanes will be increasing. It is proved that the built on the basis of this method classifier works good with text analysis' cases[5].

Blended model: as it was shown in the work [2], the combination of methods allows us to achieve fairly high results, so we combined NB and SVM in a single model, and also used a construction that allows to obtain a final classification, in which the final class of object (tweet) was determined as follows: the probabilities of belonging to different classes in the different ways classification were summed and averaged. The class with the highest average probability was chosen as the final correct one.

4 Results

We used the F1 macro measure to evaluate the results of the obtained models and to compare them with the results shown by the final blended model. This measure is a good candidate for a formal quality metric of the classifier. It brings together two other fundamental metrics: recall and precision.

[1] http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Table 1 presents the results of the intermediate classifiers and the result of the final blended model for classification by misogyny, target and different types of misogyny. We can conclude that the blended model shows the best results for all three types of classification.

Task	Classifier	F1-score
Misogyny Identification	LR	0.78
	NB+SVM	0.73
	Blend	**0.79**
Target Classification	LR	0.72
	NB+SVM	0.76
	Blend	**0.76**
Misogynistic Behavior	LR	0.55
	NB+SVM	0.61
	Blend	**0.65**

Table 1. Performance on the validation set

Also note that the blended model shows the best results for Misogyny Identification and the worst results for classification by various categories of misogynistic behavior. This pattern can be explained by the fact that in the original training dataset we used all tweets for misogyny identification task, but there were fewer data to identify the target (i.e., about 1/2 for each type (agressive/passive) of tweet that is already recognized as misogynistic), and even fewer data to classification by types of misogynisticbehavior (about 1/5 for each type of tweet that is already recognized as misogynistic). Obviously, the constructed classifiers show better results in case when we have more input data for each type of classification.

5 Conclusion

This article demonstrates the method of constructing classifiers for solving the problems of IBEREVAL-2018 (Automatic Misogyny Identification). The technique includes input data preprocessing using TF-IDF and construction of a final model combining simpler models. The results show that the final blended model, as expected, gives the best results in the construction of all three classifiers.

References

1. Fersini, E., Anzovino, M., Rosso. P. Overview of the Task on Automatic Misogyny Identification at IberEval. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018). CEUR Workshop Proceedings. CEUR-WS.org, Seville, Spain, September 18, 2018

2. Wang, S., Manning, C.D.: Baselines and bigrams: simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, ACL 2012, vol. 2, pp. 90–94 (2012)

3. Raymond, E. Wright. Logistic regression. 1995.

4. Genkin, A., Lewis, D., Madigan, D. Large-scale bayesian logistic regression for text categorization. Technometrics, 49(3):291–304, 2007.

5. Joachims, T. Learning to classify text using support vector machines: Methods, theory and algorithms. Kluwer Academic Publishers, 2002.

6. Zhang, H. and Di Li. Na¨ıve bayes text classifier. In Granular Computing, 2007. GRC 2007. IEEE International Conference on, pages 708–708. IEEE, 2007.

7. Bartlett, J., Norrie R., Patel S., Rumpel R., Wibberley S. Misogyny on twitter, http://www.demos.co.uk/, 2014, 05

8. Alexandrov, M., Danilova, V., Koshulko, A., Tejada, J.: Models for opinion classication of blogs taken from Peruvian Facebook. In: Proceedings of 4th International Conference on Inductive Modeling (ICIM-2013), pp. 241–246

9. Kaurova, O., Alexandrov, M., Ponomareva, N.: The Study of Sentiment Word Granularity for Opinion Analysis (a Comparison with Maite Taboada Works). International Journal on Social Media. MMM: Monitoring, Measurement, and Mining 1(1), 45–57 (2010)