A Simple Voting Mechanism for Online Sexist Content Identification

Chao Feng[1][0000−0002−0672−1090]

University of Zurich, Rämistrasse 71, CH-8006 Zürich, Switzerland
chao.feng2@uzh.ch

Abstract. This paper presents the participation of the MiniTrue team in the EXIST 2021 Challenge on the sexism detection in social media task for English and Spanish. Our approach combines the language models with a simple voting mechanism for the sexist label prediction. For this, three BERT based models and a voting function are used. Experimental results show that our final model with the voting function has achieved the best results among our four models, which means that this voting mechanism brings an extra benefit to our system. Nevertheless, we also observe that our system is robust to data sources and languages.

Keywords: Sexism detection · Social media · Contextual word embeddings.

1 Introduction

Equality, openness, and freedom, as the foundation and pillar of the Internet spirit, should have made our cyberworld a better place. However, inequality, prejudice and bias against females still deeply implants in online social networks [14]. Sexist contents in social networks affect females' life in both mental and physical sides, for instance the career equality, life quality, even the mental health [7, 10]. So that, an automatic sexism identification system is urgently required, which could help to build a better cyberworld with equality, openness, and freedom.

Currently, researches on the sexism detection are frequently related to misogyny detection. Misogyny and sexism are usually considered as synonyms, but there are some nuances between these two terms. The definition of misogyny emphasizes more on the enmity, hatred and contempt towards female [6]. On the other hand, sexism implies the expression of any manners of oppression or prejudice against women [14]. Thus, sexism is the hypernym of misogyny.

Traditionally, technologies of sexism detection are assembled into pipelined architectures. User and relationship based features (e.g. including followers, friends, number of tweets, interacting with other sexism users, etc.), as well

IberLEF 2021, September 2021, Málaga, Spain.

as the the text-based features (e.g. TF-IDF, Character n-grams, Token n-grams, etc.) are extracted from the social network. Then, a number of traditional machine learning algorithm could be applied for this problem [1, 4, 12, 15].

However, recent approaches based on end-to-end neural networks offer a promising alternative. Recurrent Neural Networks (RNN) and Long Short Term Memory networks are commonly used for Natural Language Processing tasks in earlier works. [2, 9] have applied Bi-LSTM models to sexism, misogyny detection. But these approaches are mainly used the context-independent word embeddings or character embeddings. Since the pretrained transformer-based language models have been introduced to the NLP area, context-dependent embeddings are heavily used in different tasks. Researches have adopted BERT model to sexism detection task, and achieved highly scores [8, 11, 14].

This work is done as part of the EXIST 2021 (Sexism Identification in Social Networks) shared task [10], which is a classification task for online social media data in English and Spanish. We are interested in the sub-task one called Sexism Identification. This sub-task is a binary classification problem, we aim to combine sentence level embeddings learned by deep-learning models like BERT [5] with a voting mechanism, to create a sexism identification system and help for fighting against online sexist. To reduce the noise of social media as well as the multilingual noise, we leverage various features of language models, and implement our system in three separate stages:

1. In order to solve the multilingual problem, we use three different pretrained language models in produce the sentence level embeddings, including the original English version BERT [5], the Multilingual BERT [13], and the Spanish version pretrained language model BETO [3].
2. Three different models are built upon these different embeddings and inference the output label independently.
3. We use a simple voting mechanism to get the final label.

The note is organized as follows. We present the description of our method in Section 2. Then we group the experimental results in Section 3 before discussing the perspectives and concluding in Section 4.

2 System Architecture

This section presents the architecture and the description of our system. Three basic models and a vote-inference system are described respectively in Subsection 2.1 and Subsection 2.2.

2.1 Basic Models

We develop three independent basic models for label prediction by combining language models with simple neural networks. As before mentioned, three language models are applied in our system, including BERT, Multilingual BERT, and BETO.

Basic Model One: This first basic model connects the language models with a feed forward network. In order to gain the multilingual representation for the input text data, our model uses two different pretrained language models, the BERT and the Multilingual BERT. After tokenization, the text data feeds into these two language models respectively, and generate two sentence embeddings. A concentration layer is applied to combine these two embeddings together. After that, embedded data flows to a feed forward network for the final inference I_1.

Basic Model Two: Research points out that the performance of Multilingual BERT model is not as good as the language specific language models [3]. On this count, we replace the Multilingual BERT by Spanish version language model, BETO. Similar with the first model, we also use a concentration layer to combine the English embedding with the Spanish embedding, and feed them into a inference network to get the final output I_2.

Basic Model Three: To leverage the sequential information of the input text, we use a Bi-directional LSTM network to obtain the consecutive features after the language model. In this time, we just use a single multilingual BERT, and get the sequential embeddings for each input token. After the Bi-directional LSTM, we use a Sigmoid function to get the final inference I_3.

2.2 Voting Mechanism

As before mentioned, I_1, I_2, I_3 are three inference values predicted by our basic models, and a simple voting mechanism is applied to count the final output. If two or more models draw the same inference label, this label will be our final prediction. The final output is predicted as:

$$f(x) := \begin{cases} 0 \text{ if } I_1{+}I_2{+}I_3 < 2 \\ 1 \text{ if } I_1{+}I_2{+}I_3 >= 2 \end{cases} \tag{1}$$

The architecture of our model is shown in Figure 1. As we can see, the input text date feeds into each basic model, and then flows to each inference network respectively. Finally, all of this information is combined in the voting function for the prediction.

3 Experiments and Results

In this section, we present the description of datasets, the setup of our experiments, and the final results of our experiments. Experimental validation is conducted on the EXIST 2021 training and test corpus. The datasets are described in subsetion 3.1 followed by the analysis of our final results in subsetion 3.2.

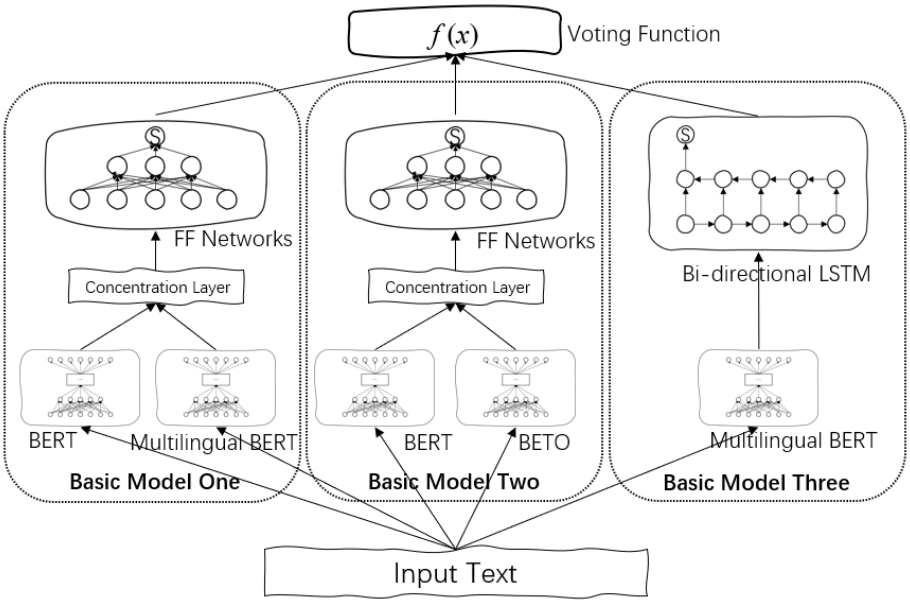

Fig. 1. The final model combines three basic models with a voting mechanism

3.1 Data Description

There are two sub-tasks in this shared task, the first one is Sexism Identification, which is a binary classification task, and the system needs to decide whether a given social media text is or is not sexist. The second sub-task is to categorize these sexist texts into different predefined sexism types. This paper focuses on the first sub-task.

Both of the training and test datasets consist in English and Spanish text data, which are collected from the social media such as Twitter and Gab. Table 1 briefly describes the corpus. As we can see, there are 6977 tweets/gabs in the training set, and 4368 tweets/gabs in the test set, and labels are balanced in this sub-task.

3.2 Results and Discussion

The evaluation of this sexism identification sub-task is mainly based on Accuracy score. Our results are gathered in Table 2, which contains the Accuracy and F-Measure scores for the basic models and the final systems. Our first question is whether the voting function brings additional benefits to our system. As we can see, the our final system (with a simple voting mechanism) performs the best Accuracy and F-measure metrics among these four systems, which provides a 3% extra performance regarding basic models.

Unlike the training data set, the test set contains the text data from different sources, which are the Twitter and Gab. It brings the second question, which is

Table 1. Sentence numbers for each data set for sub-task one.

Date sets	Language	Label	Number of tweets/gabs
Training	en	non-sexist	1800
	en	sexist	1636
	es	non-sexist	1800
	es	sexist	1741
Test	en	non-sexist	1050
	en	sexist	1158
	es	non-sexist	1037
	es	sexist	1123

Table 2. Evaluation Results: Accuracy and F-Measure

Date set	Model	Accuracy	F1
Test	Basic Model One	0,7370	0,7360
	Basic Model Two	0,7280	0,7275
	Basic Model Three	0,7287	0,7287
	Final Model (with voting mechanism)	**0,7553**	**0,7551**

whether our final model is data source sensitive. So we list the results in different data sources for each model. As shown in Table 3, even though the training data comes from Twitter, the results in Gab are better than those in Twitter in the test set, which shows a robust ability of the language models.

Table 3. Evaluation Results by Data Source: Accuracy

Date set	Data Source	Model	Accuracy
Test	Twitter	Basic Model One	0,7271
		Basic Model Two	0,7312
		Basic Model Three	0,7312
		Final Model (with voting mechanism)	**0,7504**
	Gab	Basic Model One	0,7709
		Basic Model Two	0,7169
		Basic Model Three	0,7200
		Final Model (with voting mechanism)	**0,7719**

The last question about our system is whether it is language sensitive. Table 4 lists the evaluation metrics of our systems by different languages. Results show that the performance of our models in different languages is similar to each other,

and it seems that to introduce the language specific language models, i.e. the BETO has not brought extra benefits to our final system.

Table 4. Evaluation Results by Language: Accuracy

Date set	Language	Model	Accuracy
Test	en	Basic Model One	0,7197
		Basic Model Two	0,7351
		Basic Model Three	0,7486
		Final Model (with voting mechanism)	**0,7559**
	es	Basic Model One	**0,7546**
		Basic Model Two	0,7208
		Basic Model Three	0,7083
		Final Model (with voting mechanism)	**0,7546**

4 Conclusions

In this paper, we implement a classification system that combines the language models with a simple voting mechanism. Our system has been evaluated on the EXIST 2021 sub-task one. The evaluation results have shown that our system is datasource and language robust, and the voting function indeed brings in extra benefits to our final system. Our experiments prove that the pre-trained language model is also highly adaptable to social media texts, and a simple voting mechanism can highly leverage the predictive ability of the multi-model system.

References

1. Anzovino, M., Fersini, E., & Rosso, P. (2018). Automatic Identification and Classification of Misogynistic Language on Twitter. *NLDB*.
2. Buscaldi, D. (2018). Tweetaneuse@ AMI EVALITA2018: Character-based Models for the Automatic Misogyny Identification Task. *EVALITA Evaluation of NLP and Speech Tools for Italian, 12*, 214.
3. Canete, J., Chaperon, G., Fuentes, R., & Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. *PML4DC at ICLR*, 2020.
4. Canós, J. S. (2018). Misogyny Identification Through SVM at IberEval 2018. In *IberEval@ SEPLN* (pp. 229-233).
5. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
6. Fersini, E., Rosso, P., Anzovino, M. (2018). Overview of the Task on Automatic Misogyny Identification at IberEval 2018. *IberEval@ SEPLN, 2150*, 214-228.

7. Francisco Rodríguez-Sánchez, Jorge Carrillo-de-Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, Trinidad Donoso. Overview of EXIST 2021: sEXism Identification in Social neTworks. *Procesamiento del Lenguaje Natural, vol 67*, septiembre 2021.
8. Fortuna, P., Soler-Company, J., Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?. *Information Processing Management, 58*(3), 102524.
9. Goenaga, I., Atutxa, A., Gojenola, K., Casillas, A., de Ilarraza, A. D., Ezeiza, N., ... Perez-de-Viñaspre, O. (2018, September). Automatic Misogyny Identification Using Neural Networks. In *IberEval@ SEPLN* (pp. 249-254).
10. Manuel Montes, Paolo Rosso, Julio Gonzalo, Ezra Aragón, Rodrigo Agerri, Miguel Ángel Álvarez-Carmona, Elena Álvarez Mellado, Jorge Carrillo-de-Albornoz, Luis Chiruzzo, Larissa Freitas, Helena Gómez Adorno, Yoan Gutiérrez, Salud María Jiménez Zafra, Salvador Lima, Flor Miriam Plaza-de-Arco and Mariona Taulé (eds.): Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), *CEUR Workshop Proceedings*, 2021.
11. Pamungkas, E. W., Basile, V., Patti, V. (2020). Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing Management, 57*(6), 102360.
12. Pamungkas, E. W., Cignarella, A. T., Basile, V., Patti, V. (2018). 14-ExLab@ UniTo for AMI at IberEval2018: Exploiting lexical knowledge for detecting misogyny in English and Spanish tweets. In 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2018 (Vol. 2150, pp. 234-241). CEUR-WS.
13. Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual bert?. *arXiv preprint arXiv:1906.01502.*
14. Rodríguez-Sánchez, F., Carrillo-de-Albornoz, J., & Plaza, L. (2020). Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data. *IEEE Access, 8*, 219563-219576.
15. Waseem, Z. (2016). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science* (pp. 138-142).