Kalpa Publications in Computing

Volume XXX, 2022, Pages 157–160

Proceedings of V XoveTIC Conference. XoveTIC 2022

Some Experiments on the Use of Natural Language Processing for Sexism Detection and Classification in Social Media

Roi Santos-Ríos, Jesús Vilares, and Miguel A. Alonso

Universidade da Coruña and CITIC. LyS Research Group, Department of Computer Science & Information Technologies, Campus de Elviña, 15071 A Coruña, Spain
{roi.santos.rios, jesus.vilares, miguel.alonso}@udc.es

Abstract

As the world's digital population grows, so does the reach and usage of social media: in 2021, 56% of the global population were social media users [1]. Social networks are now a part of our everyday life and continue to transform the way we interact with others on a global scale The downside is that negative behaviors in social interactions are also increasing their presence. For example, between March 1 and April 30, the OBERAXE (Spanish Observatory of Racism and Xenophobia) has detected a 27% increase in hate speech on social networks with respect to the previous two-month period [2]. In this paper we target the detection and classification of sexist content in social media texts. Two tasks are considered: (i) a binary classification task to decide whether a given text is sexist or not; and (ii) a multiclass classification task according to the type of sexism present in it.

1 System Architecture

Three proposals were developed to tackle those tasks. We started with a Multinomial Naive Bayes classifier (MNB)[3] (Fig. 1) as a baseline model for both classification tasks. Punctuation marks (except "?"), stopwords, mentions and links were removed during preprocessing. We optimized the model by tunning its hyperparameter alpha in order to maximize the AUC score in the first task and the f1-score for the second task. At this point we must point that, as Sect. 2 explains, due to the limited size of the dataset, divided evenly in Spanish and English tweets, we translated each language to the other to effectively doubling its size.

For our second approach, we implemented a FastText [4] classifier for both binary and multiclass classification. The preprocessing was the same as with MNB. Hyperparamenters were also optimized using the corresponding FastText function.

Finally, we implemented a BERT-based [5] classifier (Fig. 2) employing different pre-trained models, as shown in Table 1. The preprocessing removed mentions and links, as well as properly tokenizing the sentences according to each BERT model. We implemented a hyperparemeter optimization loop using Ray Tune library, but even so it is not an exhaustive operation, as we don't have the means to train the models on TPUs and really evaluate each possible configuration of hyperparameters. Moreover, in order to tackle the dataset imbalance for the multiclass

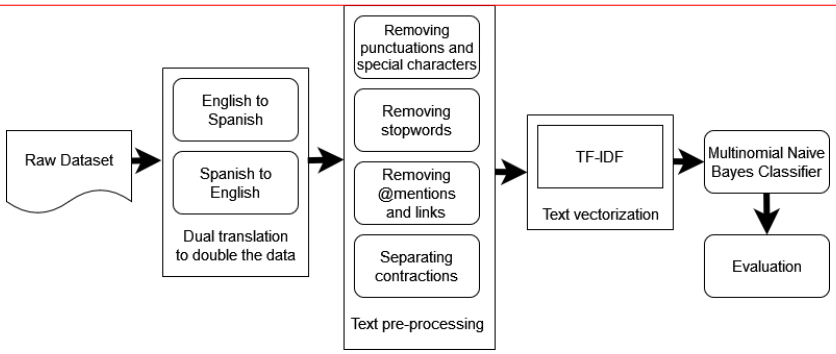

Figure 1: Architecture of the Multinomial Naive Bayes classifiers.

classification, we used synonym replacement to duplicate the amount of samples from each minority class. We settled for that method as it is not computationally demanding.

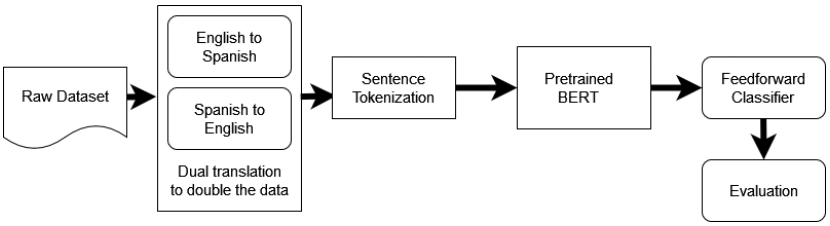

Figure 2: Architecture of the BERT-based classifiers.

2 Evaluation and Conclusions

For evaluation purposes, we will use the resources provided by the sEXism Identification in Social neTworks (EXIST) shared task at the IberLEF 2021 workshop [6]. In this event, participants were asked to perform two subtasks: (i) a binary classification task to decide whether a given text is sexist or not; and (ii) a multiclass classification task to categorize such texts according to the type of sexism present (non-sexist, stereotyping-dominance, ideological-inequality objectification, misogyny-non-sexual-violence and sexual-violence). EXIST 2001 dataset consisted of 6 977 tweets, divided evenly in Spanish and English. As explained before, in our experiments we translated each language to the other, thus effectively doubling the size of the dataset. The results (F1-scores) obtained are shown in Table 1.

The MNB classifier is not a bad choice for a more lightweight model that performs decently, but it is really outclassed by FastText. It served its purpose as an introduction to the bag-of-words model, and thanks to the Scikit-learn library it was fairly quick to set up.

In the other hand, FastText is really easy to set up and to fine tune, as the python library already provides functions for each task. Performance-wise, it does a pretty good job too, even rivaling BERT in the multiclass classification. Overall it's a really versatile model that can be quickly prototyped and executed to produce really decent results, compared to the state-of-the-art BERT models. With the augmented dataset it performed incredibly well, even though this may be ea result of overfitting the classifier.

Model	Binary English	Binary Spanish	Multi English	Multi Spanish
MNB	68.45	68.68	41.07	40.94
FastText	73.05	72.51	59.83	58.32
FastText (*ext*)	–	–	81.56	88.33
BERT base	79.94	73.70	68.50	58.44
BERT base (*mult*)	77.32	77.32	67.48	67.48
BERT base (*ext*)	–	–	72.59	65.77
BERT base (*ext*, *mult*)	–	–	74.32	74.32
SpanBERTa	–	77.76	–	64.46
BERTweet	83.72	–	72.17	–
RoBERTuito	–	76.65	–	73.87

Table 1: F1-scores obtained for our models. *ext* stands for extended dataset with synonym replacement; *mult* stands for multilingual model, trained and evaluated with both languages at the same time.

Finally we have the various BERT pre-trained models that we tested, including BERTweet and RoBERTuito, that were pre-trained with Twitter data. BERT base multilingual (as well as the extended dataset version) was trained and evaluated with both English and Spanish tweets at the same time, while the others have been trained and evaluated for each language separately, hence the different F1-score between languages. As expected, they produced the best results, but they haven't reached their fullest potential yet. After adequately preprocessing the data and fine-tunning the models, our next best choice consists on increasing the entries in the dataset. Hyperparameter tunning and model optimizations can surely improve the results of the models, but they reach a point where it's only a matter of petty decimals. However, expanding the dataset by acquiring more sexist/non-sexist tweets and properly labeling them is a time-consuming process, and it has to be done by human experts, so that's out of the question. But surely, employing advanced over-sampling techniques such as text generation could prove to be the best way to improve these classifiers.

Finally, even though the project is aimed at this specific context, it is just an example of a field where such Text Mining and Natural Language Processing (NLP) techniques can be applied to. Moreover, our proposals are easily adaptable to other specific contexts involving sentiment analysis.

References

[1] Statista, "Number of social media users worldwide from 2018 to 2022, with forecasts from 2023 to 2027," Aug 2022. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (accessed on August 27, 2022).

[2] OBERAXE, "Boletín de monitorización del discurso de odio en redes sociales: 2022 monitorización marzo-abril," May 2022. https://www.inclusion.gob.es/oberaxe/ficheros/ejes/discursoodio/Boletin_marzo-abril2022_monitorizacion.pdf (accessed on August 27, 2022).

[3] A. R. Susanti, T. Djatna, and W. A. Kusuma, "Twitter's sentiment analysis on GSM services using multinomial Naïve Bayes," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 15, no. 3, pp. 1354–1361, 2017.

[4] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Volume 1 (Long and Short Papers), pp. 4171–4186, Association for Computational Linguistics, 2019.

[6] F. J. Rodríguez-Sanchez, J. Carrillo-de-Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, and T. Donoso, "Overview of EXIST 2021: sEXism Identification in Social neTworks," *Procesamiento del Lenguaje Natural*, vol. 67, pp. 195–207, 2021.