|  |
| --- |
| **Instructions for ACL 2023 Proceedings** |
|  |
| Atharv Badge Riley Halloran  Indiana University Indiana University  Institution 1 address Luddy SICE  [firstauthor@i1.org](mailto:firstauthor@i1.org) rifhall@iu.edu |
|  |
|  |
|  |
|  |
|  |

Abstract

This paper discusses the methodology and results for the sEXism Identification in Social neTwork (EXIST) 2023 challenge. This shared task contains three tasks; sexism identification, source intention, and sexism categorization. In this paper we tackle the 1st challenge and use the mixed Spanish/English sexism identification dataset. Our approach makes use of a roBERTa transformer model to classify the data. The model used was taken from the Hugging Face library. We created partitions in the data for both languages and also for the male and female annotators. We used this data to train various models and see how they performed on the EXIST shared task.

Introduction

Sexism is defined by in the Oxford Dictionary as, “prejudice, stereotyping, or discrimination, typically against women, on the basis of sex”. Sexism is a long running issue and in the modern era, social media has given a new platform for sexism to be shared. The EXIST shared task is a scientific challenge meant to further the progress in sexism detection and to catalog a year-on-year benchmark for progress. This task brings individuals and scientists around the world to develop models based on the provided data.

The EXIST task 1 is called Sexism Identification with the goal being thus. This challenge is a binary classifier task, meaning that our model will take in the provided data, and return a labeling for each tweet. Our submission will enter the database for EXIST and be compared with others around the globe. With the amount of content generated each day the need for automated detection algorithms has only risen. To that end EXIST and those participating are pushing the boundaries of what is possible in the space.

Our approach to this task was to use a roBERTa model to label our data. The model was pretrained with a masked language modeling (MLM) objective on a large corpus of English data. When trained this way, the model has 15% of its words masked over so the model can not take them in as an input. From there the data that is used gets better represented and the model learns a closer representation of the English language. This meant that we could set up the model very efficiently and have a working project up and running quickly. RoBERTa is a transformer-based model which uses self-attention to put the emphasis of understanding on the correct words. It is a model commonly used in natural language processing.

A large amount of emphasis was placed on tuning our model using the dev split of the data. We split the data based on annotator age and gender to examine how potential biases in the annotators can influence the efficacy of the model.

Background

Hate speech is a significant problem in today's society, particularly in online platforms where anonymity and lack of accountability often lead to its proliferation. Sexism is a type of hate speech and methods for detecting hate speech apply well towards sexism. In this literature review, we will explore the most significant works that we have referred to to finalize our approach for completing the EXIST 2023 Task 1.

The EXIST 2021 overview paper by Sanchez et. al [1] discusses the different classification approaches employed for Task 1 (sexism detection) — transformer-based models, other deep learning methods like LSTMs and traditional ML approaches of SVM, Logistic Regression, Random Forest. Transformer based model architecture is preferred by the majority of participating teams and all of the top-10 performing teams. The trend continues in the EXIST 2022 overview paper by Sanchez et. al [2] where with the exception of one team all other teams employ transformer-based models for both Task 1 and Task2 of sexism detection and classification. The traditional ML methods even with the extraction of additional features didn't produce comparable results to the transformer based models. We can see performance comparisons of different transformer models in the paper by Álvarez et al. [3], in which we see that the RoBERTa model is performing well on English datasets but lacks on the Spanish datasets. This is evident from [1] and [2] as well where for English datasets either RoBERTa or BERT based model are used, but for Spanish dataset only BERT is used by the top 10 teams across the tasks.

For exploring data preparation techniques we refer to Schutz et al. [4] where they compare the model performance on sexism detection between external data pre-training and data augmentation. The results showed that performance is enhanced by pre-training on external data but it deteriorates by the use of data augmentation. The decrease in model performance by using data augmentation is also shown in Bedmar et al. [5] where they compare model performance of BERT, RoBERTa, XLNet, DistilBERT with and without augmented data. For all the models the non-augmented data gives better results. The performance enhancement by pre-training on external data is also stated in [2] where the authors claim that transformer-based models benefit from training with data from the same source. One more unique data processing approach was explored in Paula et al. [6], where they first translate the foreign-language data into English and then train a single language model for sexism detection and classification. This framework of single language model yielded better results than the ensemble transformer models approach.

From the above review we think it would be best to start with non-augmented data input using RoBERTa for English and BERT for Spanish datasets. We shall also pre-train the models using external data to enhance model performance and try to explore the single-language model approach.

The Data

The dataset for the EXIST 2023 shared task was created from a more than 8,000,000 English and Spanish tweets with two classes, Sexist and Not Sexist, as label. Data was split between a standard train, dev, test split with 6,920, 1,038, and 2,076 tweets for each respective category. The provided data contains a tweet scraped from twitter and the annotations. These annotations include a gold standard label, the language, and the gender and age breakdown of the annotators. The seeds for the data are seeded so that there is an even distribution of English and Spanish language tweets. Since the model chosen for this paper is a roBERTa model we chose to use no preprocessing for our data. BERT style models perform better when inputting a full string of text as opposed to bag of word tokens or stemmed text.

The Model

Hugging Face

As stated earlier our model takes advantage of the Hugging face library to build our model. Hugging Face is an AI group that maintains a standard open source library of machine learning models for general use. Their library is mainly made of transformer-based models that are tuned for specific tasks. These often come with pre-trained versions or ones that require a reference. They have many different types of roBERTa models, but we settled on the standard Roberta-large model. This model was pretrained using MLM and is designed to enable fine tuning.

* 1. RoBERTa Implementation

For our roBERTa model we load in the pretrained version. From there we map the features to the tweets and the targets to the task 1 labels. Our tokenizer features are set to the tweets and set the max length to 128. Our value tokenizer is set to encode the value of our features to the original tweets. We use an optimizer to get the most of our model. Specifically, we use the AdamW optimizer from Pytorch. Once all of the preparations have been set up, then the trainer is run and the roBERTa model gets trained using the optimizer.

In order to test the biases of the annotators and also see if they produce a better model, the model was trained on the full data set, then the English and the Spanish (figure 1). Then we created a category for male and female annotators and used them to train the roBERTa model.

Performance

On the full dataset, the roBERTa model reached an F1-score of 0.734. This baseline is larger than the Spanish trained model, and smaller than the solely English trained one. Most surprisingly the male female F1 score differences are greater than 0.02, the only female annotated dataset demonstrating a clear increase in results. A theory as to why this is the case is that since this is a task specifically targeted towards sexism, then the female annotators would be in unison in their labeling whereas the male annotators having not experienced sexism directed at them would be more likely to label not-sexist. All our models were tested and developed on the same split of data

Conclusion

In the pursuit of completing the EXIST 2023 shared task, our team created a roBERTa model for different data partitions and used them to train different models. The data showed that the English model performed better than the full dataset and the Spanish language dataset. In addition the model trained on female annotations performed the best out of any model that our team used. The data suggest a bias in female annotators to

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | Sr no | Data | F1-score | | 1 | Full | 0.734 | | 2 | English | 0.745 | | 3 | Spanish | 0.729 | | 4 | Male | 0.728 | | 5 | Female | 0.751 | |
| Figure 1: Data table with the result running the partitioned data**.** |

Acknowledgments

This document has been adapted by Jordan Boyd-Graber, Naoaki Okazaki, Anna Rogers from the template for earlier ACL, EMNLP and NAACL proceedings, including those for EACL 2023 by Isabelle Augenstein and Andreas Vlachos and EMNLP 2022 by Yue Zhang, Ryan Cotterell and Lea Frermann.

References

Rodríguez‐Sánchez, Francisco et al. “Overview of EXIST 2021: sEXism Identification in Social neTworks.” Proces. del Leng. Natural (2021).

Rodríguez‐Sánchez, Francisco et al. “Overview of EXIST 2022: sEXism Identification in Social neTworks.” Proces. del Leng. Natural (2022).

Álvarez, Victoria Pachón et al. “Automatic Sexism Identification Using an Ensemble of Pretrained Transformers” Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)

Schütz et al. “Automatic Sexism Detection with Multilingual Transformer Models”. https://arxiv.org/pdf/2106.04908.pdf (2021). Dan Gusfield. 1997. *Algorithms on Strings, Trees**and Sequences*. Cambridge University Press, Cambridge, UK.

Segura-Bedmar, Isabel et al. “HULAT at SemEval-2023 Task 10: Data augmentation for pre-trained transformers applied to the detection of sexism in social media”.10.48550/arXiv.2302.12840. (2023)

Magnossão de Paula, Angel Felipe et al .”Detecting and Classifying Sexism by Ensembling Transformers Models” Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)