JCT_DM at SemEval-2023 Task 10: Detection of Online Sexism: from Classical Models to Transformers

Luzzon Efrat Liebeskind Chaya
Department of computer science, Jerusalem College of Technology
21 Havaad Haleumi St., P.O.B 16031
9116991 Jerusalem, Israel
{e.luzzon6, liebchaya} @gmail.com

Abstract

This paper presents the experimentation of systems for detecting online sexism relying on classical models, deep learning models, and transformer-based models. The systems aim to provide a comprehensive approach to handling the intricacies of online language, including slang and neologisms. The dataset consists of labeled and unlabeled data from Gab and Reddit, which allows for the development of unsupervised or semi-supervised models. The system utilizes TF-IDF with classical models, bidirectional models with embedding, and pre-trained transformer models. The paper discusses the experimental setup and results, demonstrating the effectiveness of the system in detecting online sexism.

1 Introduction

Sexism is a pervasive problem online, particularly for women who are disproportionately targeted. It can cause harm, make online spaces unwelcoming, and reinforce social inequalities. Automated tools have been developed to identify sexist content at scale, but most current systems only provide generic, high-level classifications without further explanation. This lack of detail makes it difficult for users and moderators to understand and trust the decisions made by these systems.

SemEval 2023 - Task 10 - Explainable Detection of Online Sexism (EDOS) (Kirk et al., 2023) aimed at releasing a dataset to assist in the development of English-language models for sexism detection that are more accurate and explainable. This paper handles the first subtask A in task 10. We experimented with systems relying on classical models, deep-learning models, and transformer-based models to tackle the first task of binary sexism detection on Gab and Reddit. The goal of our system is to provide a comprehensive approach that can handle the intricacies of online language, which often involves slang, colloquialisms, and neologisms.

2 Related Work

Previous competition "sEXism Identification in Social neTworks (EXIST)" (Rodríguez-Sánchez et al., 2021) was conducted during the IberLEF 2021 international forum, which was exclusively focused on the field of Natural Language Processing. The competition aimed to classify multilingual posts into five categories, including ideological and inequality, stereotyping and dominance, objectification, sexual violence, misogyny, and non-sexual violence. The SemEval competition differs and involves the identification of 11 types of sexual posts, indicating a more complex and nuanced classification task.

In addition, Fortuna (Fortuna et al., 2022) addresses the challenges that arise when applying conventional NLP methodologies to hate speech detection. The authors reflect on the inadequacies of frequently used conventions for text classification with respect to hate speech detection and discuss the future of current methodologies for this task. Based on their analysis, they conclude that current models are incapable of detecting hate speech without causing harm to marginalized communities. Therefore, they call for the scientific community to adapt NLP methodologies and center on the impact that used methodologies may have on marginalized communities. The authors believe that by critically reflecting on the potential impacts of methodologies for hate speech detection, the scientific community can identify methodologies that result in more just futures.

3 Dataset

The dataset (Kirk et al., 2023) for the binary sexism detection task consists of three parts: a training set with 14,000 sentences, a development set with 2,000 sentences, and a test set with 2,000 sentences (Table 1). Within the training set, 75% of the sentences are labeled as non-sexist and the remaining

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 739–743

25% are labeled as sexist. The purpose of this dataset is to develop English-language models for automated sexism detection, which is an important and growing problem online.

In addition to the labeled data, there are two unlabeled corpora, one from Gab and one from Reddit. These corpora are provided by task organizers as well. Each contains 1,000,000 sentences. These corpora were included to provide additional unlabeled data for the development of unsupervised or semi-supervised models. The availability of this dataset provides a valuable resource for researchers and practitioners to develop more accurate and explainable models for detecting and combatting sexism online.

train	dev	test
14000	2000	4000

Table 1. The number of samples in each dataset.

4 System Overview

Our system incorporates a variety of models and methodologies, including classical models such as Random Forest (Ho, 1995) and Logistic Regression (Grimm and Yarnold, 1995) based on vectors generated by the TF-IDF statistical measure. We also explored deep learning models with bidirectional layers and word embedding created by Word2Vec (Mikolov et al., 2013). Additionally, we explored transformer-based models to further enhance our system's performance.

4.1 TF-IDF with Classical Models

For the classical models, we utilized the TF-IDF statistical measure, to convert sentences into vectors. We experimented with two approaches, converting sentences based on words and converting them based on characters with multiple values of n-grams. We used the resulting TF-IDF vectors as input for both Random Forest and logistic regression models. We performed this step twice, first by calculating the TF-IDF only on the training dataset, and second by calculating it on the unlabeled dataset.

4.2 Bidirectional Model with Embedding

To obtain vectors with better semantic context, we incorporated pre-trained embedding models such as Google's Word2Vec, which was trained on the Google News dataset, and GloVe-Twitter (Pennington et al., 2014). The GloVe model was trained on a smaller dataset, but its corpus is more similar to ours. Additionally, we implemented another word2vec model trained on the unlabeled dataset to obtain a more specific context for our mission.

The embedding matrix size is determined by the number of unique words in the training corpus and the vector's size. We used the resulting matrix as input to the embedding layers in our bidirectional models. Each row in the matrix represents a unique word in the training dataset. In case the word exists in the embedding model, we initialize the row with its corresponding vector, otherwise, we initialize it with a random vector from a normal distribution (mean and standard deviation similar to those of the embedding vectors).

Our bidirectional model architecture is based on a corresponding problem that involves classifying the emotional sentiment of COVID-19-related text as positive or negative (Arbane et al., 2023). The paper demonstrates the main advantage of replacing a regular LSTM (Hochreiter and Schmidhuber, 1997) with a Bi-LSTM (bidirectional LSTM) (Wang et al., 2017), which allows the system to learn from both directions in the text, unlike systems that use a regular LSTM, which can only learn from left to right.

4.3 Transformed-Based Models

We explored two different pre-trained transformers approaches. The first transformer used a general corpus with a higher volume of data. A smaller corpus, but more closely aligned with our problem, was used in the second approach.

For the first approach, we utilized the RoBERTa (Liu et al., 2019) pre-trained transformer model with its corresponding tokenizer. The RoBERTa model was trained on a large dataset, making it an effective choice for a wide range of tasks. For the second approach, we used the BERTweet (Nguyen et al., 2020) model, which was trained on a smaller dataset of tweets. The model and its corresponding tokenizer were specifically designed for sentiment analysis on Twitter data.

5 Experimental Setup

Our implementation uses Sklearn library (Pedregosa et al., 2011) for the classical models (TF-IDF, logistic regression and Random Forest), Gensim library (Rehurek and Sojka, 2011) for the embedding models, Keras library (Chollet et al., 2015) for the Bi-LSTM model and the transformers li-

brary by HuggingFace (Wolf et al., 2020) for the pre-trained transformer models and corresponding tokenizers.

5.1 Pre-processing

The text data of social media is known for its informal nature, with users frequently using slang, abbreviations, and nonstandard grammar. The system incorporates three pre-processing techniques: handling emojis, capital words, and lemmatization. Emojis are graphical symbols used to express emotions and reactions, which can have a substantial impact on the interpretation and meaning of a sentence. Capitalization of words can alter the meaning of a sentence by emphasizing specific words. Finally, lemmatization is employed to identify the base form of a word, which can help to disambiguate the meaning of words with multiple senses or connotations.

During pre-processing, we used the Spacy library (Honnibal and Montani, 2017) and specifically the pre-trained en_core_web_sm model for lemmatization. We also utilized the emoji and emot library for converting emojis to words.

5.2 Classical Models

We experimented with the TF-IDF statistical measure by varying n-gram values for both word and character models. Specifically, we tested the word statistical measure with n-gram ranges of [1, 1], [1, 2], [1, 3], and the character statistical measure with n-gram ranges of [1, 1], [2, 2], [3, 3]. Logistic regression and Random Forest models were trained on the TF-IDF vectors. The logistic regression model was trained with hyperparameters of penalty L2, lbfgs solver, regularization equal to 10, and class weight set to False. The Random Forest model was trained with hyperparameters of 50 n_estimator and 2 max depth. The experiment was conducted twice, once with the TF-IDF being fitted only on the training dataset, and once with the TF-IDF being fitted on the unlabeled dataset.

5.3 Embedding Models

For the Bi-LSTM model, we used GloVe-Twitter-200 pre-trained vectors, trained on a corpus of Twitter posts, and thus more similar to our dataset. Additionally, we examined the Word2Vec-Google-News-300 pre-trained vectors, which were trained on a larger dataset. We also experimented with training vectors on the unlabelled dataset to obtain vectors that were from the same corpus and potentially more tuned to our task. The hyperparameters for the pre-trained vectors were a vector dimensionality size of 30, a minimum word count of 7, and a window context size of 5. An architecture for a Bi-LSTM model was developed based on the same layers in the Social Media-based COVID-19 (Arbane et al., 2023) model, with the softmax loss. Additionally, we trained the model twice, once using Focal Loss and once using CrossEntropy.

5.4 Transformer Models

To investigate the performance of state-of-the-art transformer models, we used RoBERTa (RoBERTa-base) and BERTweet (vinai/bertweet-base) along with their corresponding tokenizers. We fine-tuned the models using the adamw optimizer, a learning rate of 2e-5, and a batch size of 16. The training was performed on a GPU T4x2 machine for 1.5 epochs. We use the softmax loss function.

The measure for evaluation is the Macro-F1 Score as defined by task organizers. The classical models were trained on Intel(R) Xeon(R) CPU, the embedding model on GPU Tesla T4, and the transformer models on GPU T4x2.

6 Results

The results of various models on the dev dataset, evaluated using the F1 metric score, are presented in Table 2. Our analysis focuses on the classical models, as well as two variants of a Bi-LSTM model trained with CrossEntropy and Focal Loss, and two transformer models, RoBERTa and BERTweet.

Among the classical models, the Random Forest model failed to learn the underlying patterns, resulting in a poor F1 score of 43%. In contrast, the Logistic Regression model achieved a better F1 score of 73.63% in the char approach and 73.65% in the words approach. Using an unlabeled corpus for training gives a small improvement in the Logistic Regression model's performance, compared to using only the training corpus. In In general, the TF-IDF model trained on the unlabeled corpus with a word-based approach produced the highest score.

The Bi-LSTM models with GloVe-Twitter-CrossEntropy and the Word2Vec model trained on the unlabelled corpus with CrossEntropy achieved a minor improvement, compared to their corresponding models with Focal Loss. In contrast, the

pre-trained Word2vec model achieved a minor improvement in Focal loss compared to the CrossEntropy model. In contrast to the TF-IDF, the leveraging of the unlabeled corpus did not improve performance. The results decreased by 0.636, and 0.429 compared to using the most effective Bi-LSTM models (GloVe-Twitter embedding vectors with and without focal loss, respectively). However, the Bi-LSTM model's overall performance was lower than that of the Logistic Regression model.

The transformer models, RoBERTa and BERTweet, delivered the best performance among all models. Specifically, BERTweet outperformed the RoBERTa model, resulting in a higher performance improvement of 12.02% compared to the best TF-IDF model and 16.09% compared to the Bi-LSTM model. Hyper-tuning the BERTweet model by limiting the pre-processing to lemmatization resulted in the best score on the dev test with 84.07%, which was submitted and ranked 40th with 83.62% performance on the test set.

Moving forward, potential avenues for future research include the development of the procedure of pre-training BERT on Gab and Reddit as done on Twitter in BERTweet to improve performance. Furthermore, An ensemble of models of all the models developed in this study could lead to even better performance. By leveraging the strengths of each individual model and mitigating their weaknesses, An ensemble of models has the potential to produce more reliable results. As such, we believe that our work contributes to the broader goal of addressing online sexism. It provides a foundation on future studies to build upon.

Model Name	Macro-F1
TF-IDF, based chars	0.7363
TF-IDF, based words	0.7365
TF-IDF, based chars unlabelled corpus	0.7379
TF-IDF, based word unlabelled corpus	0.7314
Bi-LSTM, pre-trained Word2Vec	0.6656
Bi-LSTM, Glove tweet	0.7120
Bi-LSTM, trained Word2Vec	0.6690
Bi-LSTM, pre-trained Word2Vec (focal loss)	0.6779
Bi-LSTM, Glove tweet (focal loss)	0.6897
Bi-LSTM, trained Word2Vec (focal loss)	0.6260
RoBERTa	0.8065
BERTweet	0.8266

Table 2. The f1 score of the various models on the dev set. The TF-IDF represents the score with the Logistic Regression.

7 Conclusion

In this study, we proposed an approach for the Explainable Detection of Online Sexism (EDOS) task, which involved the development of various systems such as classical models, deep learning models, and transformer-based models to address the binary sexism detection challenge. Through our experiments, we found that the use of similar corpora such as Twitter led to improved performance across most models. The vast volume of data on Twitter, coupled with its diverse range of topics, makes it an ideal source for training machine learning models.

References

Mohamed Arbane, Rachid Benlamri, Youcef Brik, and Ayman Diyab Alahmar. 2023. Social media-based covid-19 sentiment classification model using bi-lstm. *Expert Systems with Applications*, 212:118710.

Francois Chollet et al. 2015. Keras.

Paula Fortuna, Mónica Domínguez, Leo Wanner, and Zeerak Talat. 2022. Directions for nlp practices applied to online hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11794–11805.

Laurence G Grimm and Paul R Yarnold. 1995. *Reading and understanding multivariate statistics.* American Psychological Association.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. github.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learn-ing in python, journal of machine learning research, 12.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2):2.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67:195–207.

Jingyuan Wang, Fei Hu, and Li Li. 2017. Deep bi-directional long short-term memory model for short-term traffic flow prediction. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, Proceedings, Part V 24*, pages 306–316. Springer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.