Text classification on EDOS dataset

Ester Molinari[1]

[1]*Department of Computer Science, University of Bari Aldo Moro*

Abstract

Binary text classification can be performed by various technique, from Machine Learning approaces to Deep Learning ones. In this paper we will apply three different techniques in order to predict if a sentence can be labeled as sexist or not using the EDOS dataset from Task 10 of SemEval 2023 [1].

Keywords

Text classification, BERT Fine-tuning, Logistic Regression, Ngrams

1. Introduction and Motivations

Sexism can be difficult to detect from plain text and it is widely used in social media where phrasings can make detection even more difficult. Task 10 of SemEval 2023 proposed the **Explainable Detection of Online Sexism (EDOS)** dataset along with three main tasks to solve, named from A to C, in order to detect and explain sexism.

In this paper we will focus on Task A on binary classification of sexism in sentences addressing the difficulties presented by the dataset and some binary text classification approaches to solve the first task and some hints on the last one.

The poposed solutions can be tested on Kaggle on this link.

2. Related Work

From SemEval final report (Kirk et al.) emerges that, for all tasks, 90% of partecipants used a **transformer-based model** including RoBERTa, DeBERTa, BERT, BERTweet and DistilBERT while 8% of partecipants preferred to use traditional **machine learning methods** and the remaining **non-transformer deep neural networks**, which are often combined with other methods.

For example, partecipants have used two transformer-based models, BERT and RoBERTa, in order to perform all three tasks. It has been showed how fine-tuned RoBERTa performed better with respect to fine-tuned BERT for this dataset (Padmavathi).

Another transformer-based example can be found in another paper (Obeidat et al.) where they used RoBERTa to solve the text classification (task A) and they built an ensemble model based on BERT and RoBERTa to solve both task B and C proving that this technique performed better than fine-tuning assuming that ensemble approaches are more effective when there is variance in the data and avoid overfitting.

BERT model has also been used in another paper (Rifat et al.) for text classification using pre-trained GloVe to obtain embeddings from text. They also pointed out that the number of sexist data was comparatively low, so they decided to opt for an active learning approach.

The main inspiration for the approach that we are going to illustrate comes from this paper by Rodrguez et al.. Their proposed system incorporates information related to emotions, polarity, and

NLP Course 2023-2024

e.molinari3@studenti.uniba.it (E. Molinari)

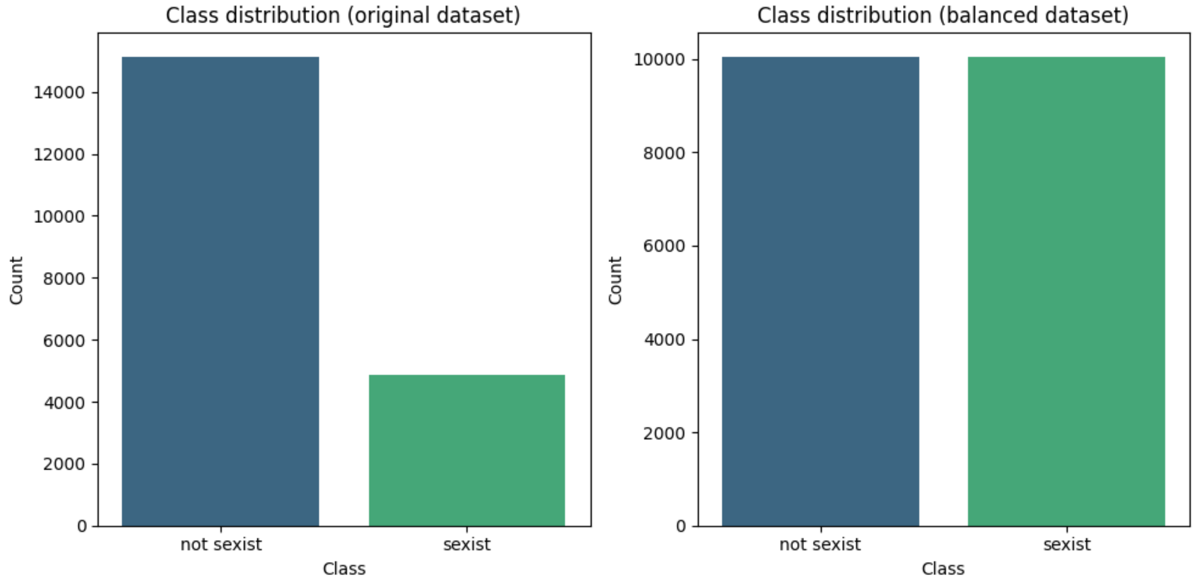

Figure 1: Class distribution in the original and balanced datasets

irony in the texts as metadata to get better results. With their experiments they have proved that, in general, the incorporation of **extra-linguistic information** helps the models to conduct the tasks.

3. Proposed Approach

3.1. Description of the solution and dataset

The proposed approach is performed on two datasets:

1. The **original** EDOS dataset, and
2. The **balanced** EDOS dataset (Rydelek et al.)

By performing some data analysis and exploration on both datasets we were able to show class distributions. The original dataset resulted very unbalanced, leading Rydelek et al. to the creation of a balanced version. For this reason, all the techniques used to solve the task will be performed on both datasets to better understand how much it influences the results. This is shown in 1.

Other interesting results can be drawn from **polarity** and **subjectivity** distribution in sentences for both datasets, shown in 2 and 3.

Lastly, we plotted the resized embeddings using PCA and we found out that both datasets are **not linearly separable**, as we can see in 4.

For solving the text classification task, **three approaches** are proposed:

1. Fine-tuning BERT for text classification,
2. Logistic regression on TF-IDF,
3. Text classification with ngrams embeddings.

In the next sections we will explore every detail of these techniques.

3.1.1. Data preprocessing

We have applied some standard preprocessing procedures on the sentences such as removing punctuation and english stopwords, removing the words [USER] and [URL]. For sentence tokenization,

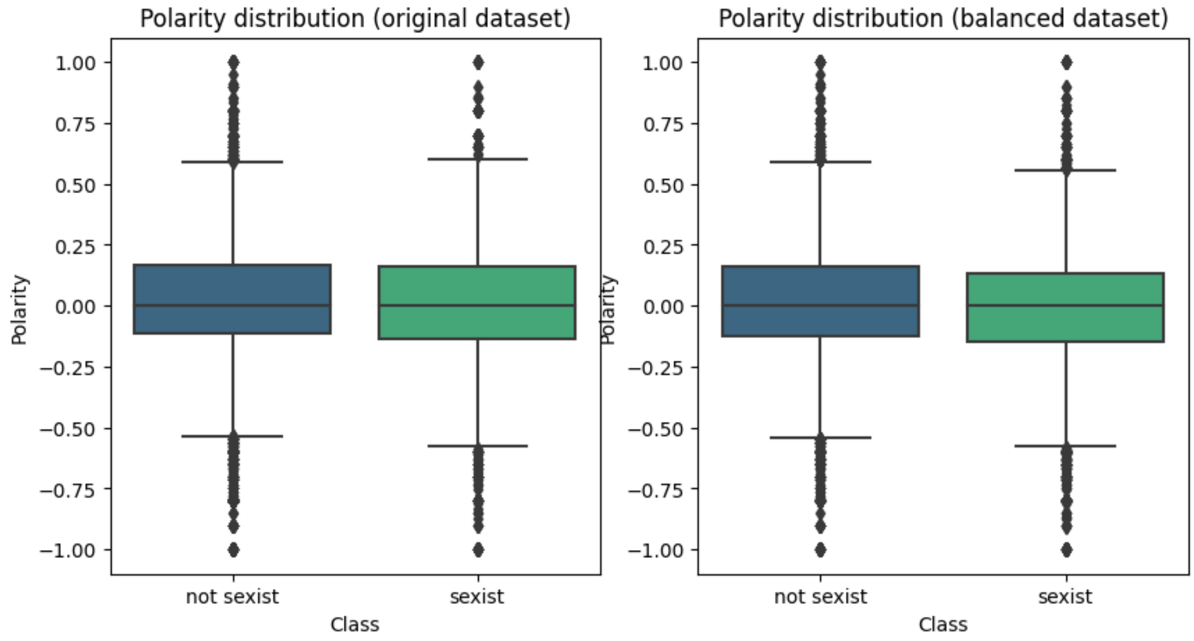

Figure 2: Polarity distribution in the original and balanced datasets

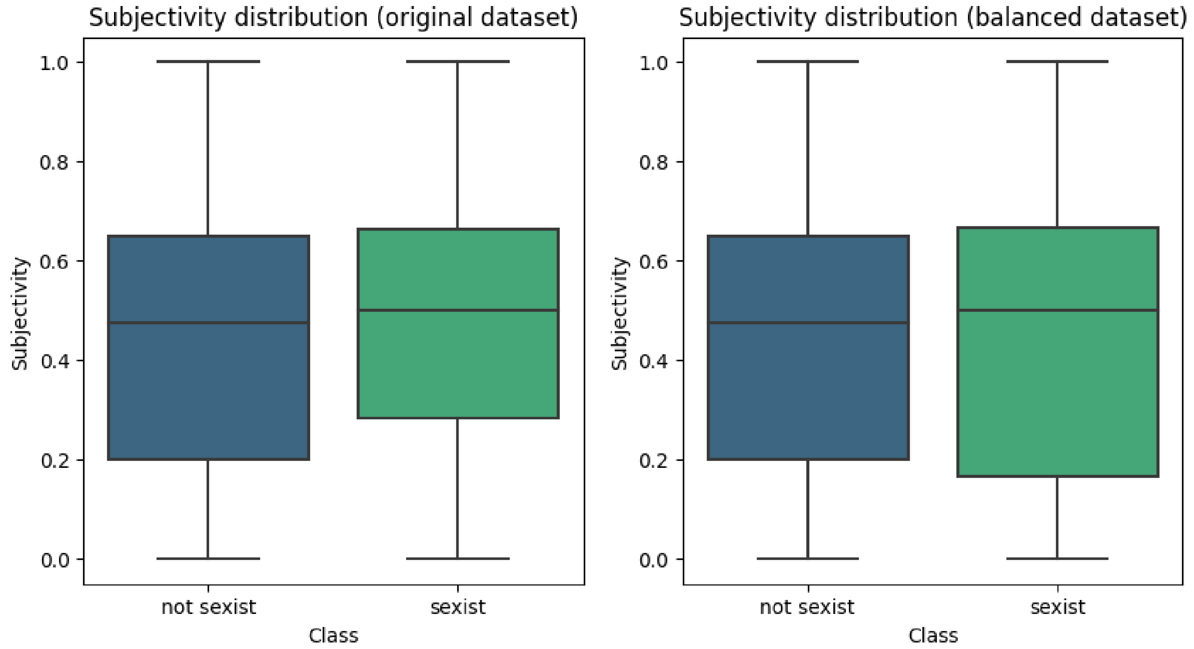

Figure 3: Subjectivity distribution in the original and balanced datasets

we used BERT tokenizer, if needed.

We also used spaCy to extract trigrams from every sentence for further experiments.

3.1.2. Fine-tuning BERT for text classification

The first approach is a **transformer-based** one using a BERT model and fine-tuning it on sentences provided by both datasets.

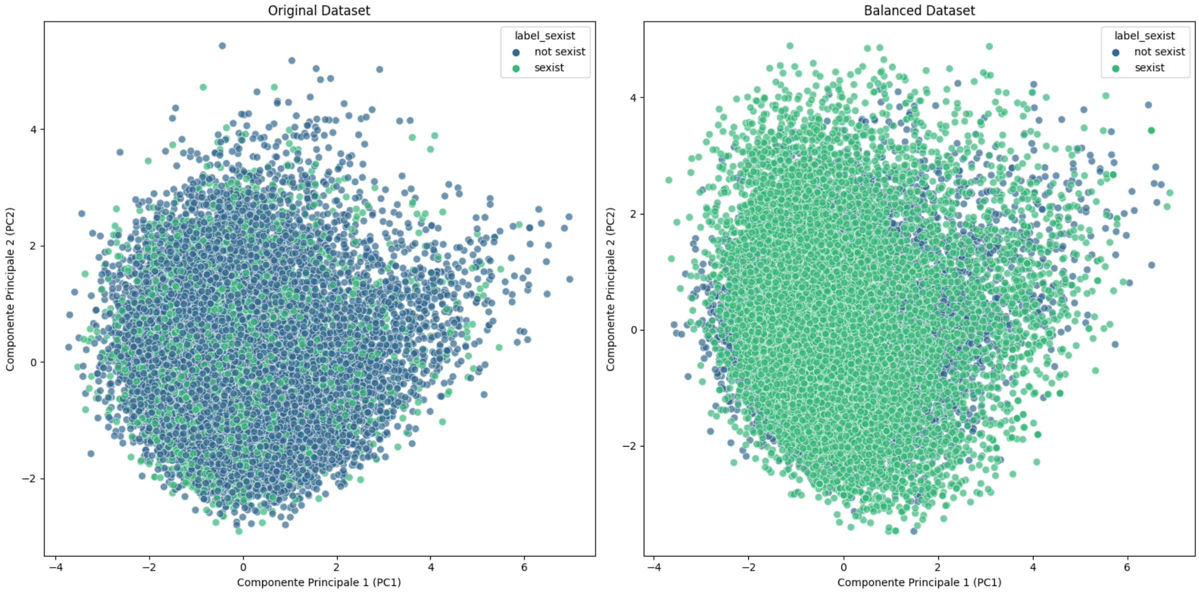

Figure 4: Embeddings distribution from original and balanced datasets

We used bert-base-uncased model from HuggingFace by setting the number of labels equal to 2 to get a binary classification. As we have performed this experiment on both datasets, we fine-tuned two BERT models called, respectively, *original bert model* and *balanced bert model* with the same loss function, which is the cross entropy. We have also defined two AdamW optimizers with the same learning rate of $2e - 5$. Both models have been trained only for 3 epochs with a final loss of 0.106 for the original model and 0.138 for the balanced one.

3.1.3. Logistic regression on TF-IDF

The second approach is a **TF-IDF-based** logistic regression that relies on the fact that both datasets are not linearly separable and we can only apply non-linear models to classify our sentences.

For this approach we firstly applied a TF-IDF vectorizer on the preprocessed text of both the datasets, then we had to convert our labels into numerical ones such that if a sentence is sexist, its value would be 1, otherwise 0.

Also in this case, we built two logistic regression models, one on the *original* and one on the *balanced* dataset, with the same number of iterations equal to 1000.

3.1.4. Text classification with ngrams embeddings

The third approach is an **ngram embedding-based** classification using majority voting on cosine similarity. This method does not need any training and it is shown in 5.

For this approach we used a sentence transformer called all-MiniLM-L6-v2 to obtain the embedding of the trigrams extracted from sentences from both datasets.

To deal with ngrams, we have performed this pipeline:

1. Extract trigrams from sexist sentences
2. Extract trigrams from not sexist sentences
3. Compute the intersection between sexist and not sexist trigrams and remove it from both sets
4. Extract trigrams embeddings with sentence transformer

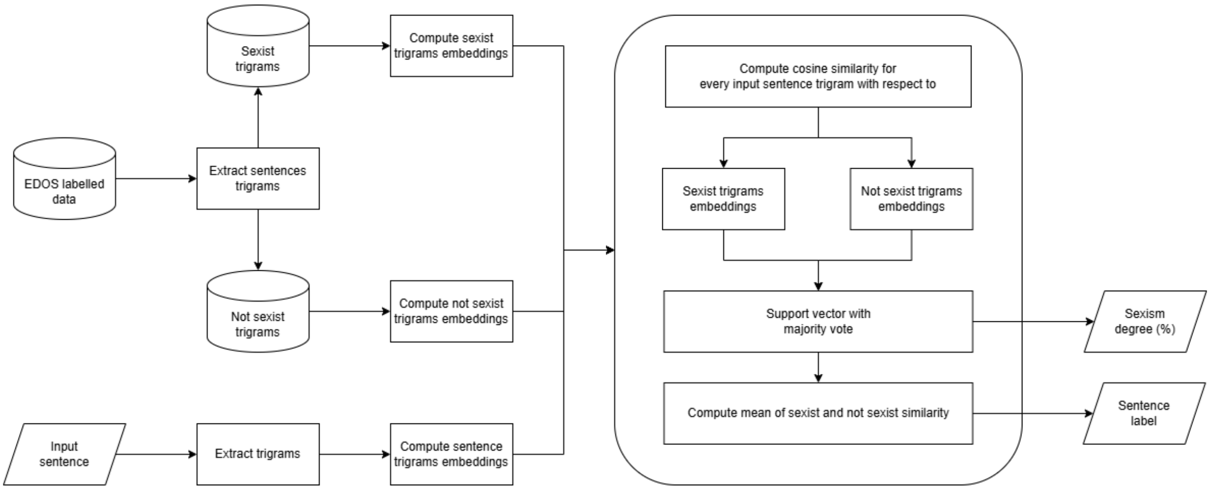
Sexist trigrams
Compute sexist trigrams embeddings
EDOS labelled data
Extract sentences trigrams
Not sexist trigrams
Compute not sexist trigrams embeddings
Input sentence
Extract trigrams
Compute sentence trigrams embeddings
Compute cosine similarity for every input sentence trigram with respect to
Sexist trigrams embeddings
Not sexist trigrams embeddings
Support vector with majority vote
Compute mean of sexist and not sexist similarity
Sexism degree (%)
Sentence label

Figure 5: Simplified version of the proposed system

Now we will only use two vectors, one for the trigrams and one for the corresponding trigram embedding. We will use these vectors to compute the **cosine similarity** between trigrams and their label, if sexist or not, by checking their similarity scores.

This model does not require any training, so we just need to preprocess an input sentence extracting trigrams from it and computing their embeddings. At this point, we will find the most similar sexist and not sexist trigram embedding and we will build a **support vector** with the same dimension of the number of trigrams from the input sentence. In this support vector we will set the corresponding value to 1 if the most similar embedding retrieved with the higher score is sexist, 0 otherwise.

In this example 1 on the original dataset we obtain the support vector

$$[10011]$$

which is converted to percentages in 60% sexist and 40% not sexist. We also compute the **global mean** of both sexist and not sexist similarity scores getting 0.18 for the sexist similarities and 0.17 for the not sexist ones. The final answer will be:

*The sentence 'If she breaths she's a t***' is 60.00% sexist. Possible detected sexist trigram is: s a t****

In the example 2 on the balanced dataset we obtain the support vector

$$[11111]$$

which is converted to percentages in 100% sexist and 0% not sexist. We also compute the global mean of both sexist and not sexist similarity scores getting 0.18 for the sexist similarities and 0.17 for the not sexist ones. The final answer will be:

*The sentence 'If she breaths she's a t***' is 100.00% sexist. Possible detected sexist trigram is: s a t****

This system will rely on the value of the mean of the similarities and will provide the **degree of sexism** detected, the same will happen for a not sexist sentence but with the not sexist percentage.

4. Evaluation

We report here all the evaluations of the three approaches.

Trigram	Sexist trigram	Not sexist trigram	Support vector value
If she breathes	If she breathes (0.91)	If she breathes (0.79)	1
she breathes she	she calls her (0.66)	breath she says (0.74)	0
breathes she s	mouth she s (0.67)	breath she says (0.72)	0
she s a	he s A (1.0)	he 's an (0.92)	1
s a t***	s a t*** (1.0)	" T*** " (0.80)	1

Table 1
Example of trigram extraction with similarities scores and support vector on the original dataset

Trigram	Sexist trigram	Not sexist trigram	Support vector value
If she breathes	If she breathe (0.91)	If she blows (0.76)	1
she breathes she	she breathe she (0.87)	breath she says (0.74)	1
breathes she s	mouth she a (0.74)	breath she says (0.72)	1
she s a	he s A (1.0)	he as a (0.83)	1
s a t***	s a t*** (1.0)	how is some (0.76)	1

Table 2
Example of trigram extraction with similarities scores and support vector on the balanced dataset

	Precision	Recall	F1-Score	Support
Not sexist	0.88	0.95	0.92	3045
Sexist	0.79	0.60	0.68	955
Accuracy			0.87	4000
Macro average	0.84	0.77	0.80	4000
Weighted average	0.86	0.87	0.86	4000

Table 3
Evaluation on fine-tuning BERT on the original dataset

	Precision	Recall	F1-Score	Support
Not sexist	0.85	0.83	0.84	2005
Sexist	0.84	0.86	0.85	2017
Accuracy			0.85	4022
Macro average	0.85	0.85	0.85	4022
Weighted average	0.85	0.85	0.85	4022

Table 4
Evaluation on fine-tuning BERT on the balanced dataset

4.1. Fine-tuning BERT for text classification

The results on the original dataset are reported in 3 while the results on the balanced dataset are reported in 4.

Observing these results we can state that the balanced dataset helps the fine-tuning process with a good precision and recall on both sexist and not sexist sentences while the original one is too unbalanced, leading to an high accuracy due to the high number of not sexist samples.

We tested our two fine-tuned models on some english sentences with the following results 5.

As said before, the high accuracy on the first fine-tuned BERT is not reliable as it predict as not sexist a lot of sexist sentences while the second fine-tuned BERT performs better on the same sentences.

Sentence	Original model	Balanced model	Actual label
Women are not suited for leadership roles	Sexist	Sexist	Sexist
The team worked together to achieve their goal	Not sexist	Not sexist	Not sexist
Go back in the kitchen and make me a sandwich	Not sexist	Sexist	Sexist
Women are better than men	Not sexist	Sexist	Sexist
Woman driving, peril thriving	Not sexist	Sexist	Sexist
Stop talking like a Karen	Not sexist	Sexist	Sexist
If she breaths she's a t***	Sexist	Sexist	Sexist

Table 5
Testing both BERT models on same sentences

	Precision	Recall	F1-Score	Support
Not sexist	0.82	0.97	0.89	3045
Sexist	0.78	0.34	0.47	955
Accuracy			0.82	4000
Macro average	0.80	0.65	0.68	4000
Weighted average	0.81	0.82	0.79	4000

Table 6
Evaluation of Logistic Regression on the original dataset

	Precision	Recall	F1-Score	Support
Not sexist	0.77	0.87	0.81	2005
Sexist	0.85	0.74	0.79	2017
Accuracy			0.80	4022
Macro average	0.81	0.80	0.80	4022
Weighted average	0.81	0.80	0.80	4022

Table 7
Evaluation of Logistic Regression on the balanced dataset

Sentence	Original model	Balanced model	Actual label
Women are not suited for leadership roles	Not sexist	Sexist	Sexist
The team worked together to achieve their goal	Not sexist	Not sexist	Not sexist
Go back in the kitchen and make me a sandwich	Not sexist	Sexist	Sexist
Women are better than men	Not sexist	Sexist	Sexist
Woman driving, peril thriving	Not sexist	Sexist	Sexist
Stop talking like a Karen	Not sexist	Not sexist	Sexist
If she breaths she's a t***	Not sexist	Not sexist	Sexist

Table 8
Testing both logistic regression models on same sentences

4.2. Logistic regression on TD-IDF

The results on the original dataset are reported in 6 while the results on the balanced dataset are reported in 7.

In this case we get a very low recall on the original dataset but it increases on the balanced one, showing again how a good dataset can improve training results.

We tested our two logistic regression models on some english sentences with the following results 8.

In this case we have an improvement on the second model based on the balanced dataset, but the results are worse than the fine-tuned approach, taking into account that this model uses TF-IDF and it

	Precision	Recall	F1-Score	Support
Not sexist	0.96	0.06	0.11	3045
Sexist	0.25	0.99	0.40	955
Accuracy			0.28	4000
Macro average	0.61	0.53	0.25	4000
Weighted average	0.79	0.28	0.18	4000

Table 9
Evaluation of ngrams similarity on the original dataset

	Precision	Recall	F1-Score	Support
Not sexist	0.88	0.06	0.11	2005
Sexist	0.51	0.99	0.68	2017
Accuracy			0.53	4022
Macro average	0.69	0.53	0.39	4022
Weighted average	0.69	0.53	0.40	4022

Table 10
Evaluation of ngrams similarity on the balanced dataset

Sentence	Original model	Trigram	Actual label
Women are not suited for leadership roles	Sexist (100.00%)	Women are not	Sexist
The team worked together to achieve their goal	Sexist (0.00%)	to achieve their	Not sexist
Go back in the kitchen and make me a sandwich	Sexist (62.50%)	Go back in	Sexist
Women are better than men	Sexist (0.00%)	better than men	Sexist
Woman driving, peril thriving	Sexist (0.00%)	Woman driving peril	Sexist
Stop talking like a Karen	Sexist (33.33%)	talking like a	Sexist
If she breaths she's a t***	Sexist (60.00%)	s a t***	Sexist
Yesterday I went to the store and I bought a cake	Not sexist (100.00%)		Not sexist

Table 11
Testing system based on original dataset on some sentences

is lightweight with respect to the transformer-based one.

4.3. Text classification with ngrams embeddings

The results on the original dataset are reported in 9 while the results on the balanced dataset are reported in 10.

These are the most interesting results. We have the highest recall on sexist sentences and the lowest one on those which are not sexist in both datasets. A slightly improvement can be appreciated on the balanced dataset which shows an higher accuracy with respect on the original one. Nevertheless we get an accuracy of 0.53 on the balanced system, we have to consider that this system is able to detect the sexist or not sexist degree of the prediction and that it is also able to show the trigram with the highest similarity score for each prediction. Furthermore, this system has no training process and it is a non-parametric model.

We tested our two systems on some english sentences with the following results for the original dataset 11 and for the balanced dataset 12.

Again, we have very interesting results also in this section. The system based on the original dataset states that all the sentences are sexist except for the last one, but the degree of sexism is the interesting

Sentence	Original model	Trigram	Actual label
Women are not suited for leadership roles	Sexist (80.00%)	Women are not	Sexist
The team worked together to achieve their goal	Sexist (50.00%)	to achieve their	Not sexist
Go back in the kitchen and make me a sandwich	Sexist (75.00%)	in the kitchen	Sexist
Women are better than men	Sexist (66.67%)	better than men	Sexist
Woman driving, peril thriving	Sexist (50.00%)	Woman driving peril	Sexist
Stop talking like a Karen	Sexist (33.33%)	talking like a	Sexist
If she breaths she's a t***	Sexist (100.00%)	s a t***	Sexist
Yesterday I went to the store and I bought a cake	Sexist (66.67%)	went to the	Not sexist

Table 12
Testing system based on balanced dataset on some sentences

part. The second, the fourth and the fifth sentences show a 0.00% of sexist degree on that label, meaning that there is no sexism in them. On the third sentence, the system states that the most sexist trigram of that sentence is "Go back in" while it should be the "kitchen" part.

Moving on to the system based on the balanced dataset, we have different results. The second sentence shows a 50% of sexism in it, which could be treated as a neutral sentence because it has 50% of sexism and 50% of not sexism. On the third sentence we have the trigram "in the kitchen" as the most sexist, as we expected. Another interesting prediction can be found on the fifth sentence, detecting "woman driving peril" as the most sexist trigram but shows a 50% degree of sexism, that could be due to the sentence construction as the first part seems neutral but the last part shows the sexism. On the last sentence we get a wrong prediction with the 67% degree of sexism in it with respect to the one from the original dataset which performs better on not sexist sentences achieving a 100% degree of not sexism.

5. Conclusions and Limitations

These three approaches on the same problem shows interesting results both on the predictions and on the influence of the dataset used for solving the task.

The most interesting approach is the last one based on ngrams because it has the worst evaluation results but it seems to return some kind of explaination on the prediction while the other could only return some probabilities on the output. As a non-parametric system, it requires better data and more time for the embeddings extraction but the explaination based on the sexism degree is its strength. For this approach could be useful to define a new metric to understand better its performances on sentences classifications.

On the fine-tuning approach, choosing better parameters and more epochs may lead to better results, using tested parameters for this task that can be found on cited papers.

References

[1] H. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 task 10: Explainable detection of online sexism, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2193–2210. URL: https://aclanthology.org/2023.semeval-1.305. doi:10.18653/v1/2023.semeval-1.305.

[2] M. Padmavathi, Ds at semeval-2023 task 10: Explaining online sexism using transformer based approach, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 1102–1106.

[3] D. Obeidat, H. Nammas, M. Abdullah, et al., Just_one at semeval-2023 task 10: Explainable detection of online sexism (edos), in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 526–531.

[4] R. H. Rifat, A. Shruti, M. Kamal, F. Sadeque, Acsmkrhr at semeval-2023 task 10: Explainable online sexism detection (edos), in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 724–732.

[5] M. E. V. Rodrguez, F. M. P. Del Arco, L. A. U. Lopez, M. T. Martín-Valdivia, Sinai at semeval-2023 task 10: Leveraging emotions, sentiments, and irony knowledge for explainable detection of online sexism, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 986–994.

[6] A. Rydelek, D. Dementieva, G. Groh, AdamR at SemEval-2023 task 10: Solving the class imbalance problem in sexism detection with ensemble learning, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1371–1381. URL: https://aclanthology.org/2023.semeval-1.190. doi:10.18653/v1/2023.semeval-1.190.