NLP Management Summary

Label Legends
26.01.2025

Executive Summary

Our project aims to build a model capable of accurately identifying whether a given text is sexist. This involves analyzing labeled training data to create a classifier, evaluating its performance on a test set, and improving its ability to handle complex, ambiguous cases of sexism. The project achieved significant milestones, ending in an ensemble model based on the DeBERTa architecture, which provided the highest recall and precision. Key challenges such as addressing sarcastic and stereotypical content were identified and partially mitigated.

Project Overview

Dataset
60k rows (20k unique sentences) of social media text from Reddit and Gab were used for the project. These were then annotated by trained female annotators (75% non-sexist) and split into ⅔, ⅓ training and test set. Challenges included imbalance data and subtle linguistic nuances like sarcasm and slang.

External Resources used
We imported some baseline models from Hugging Face, more specifically two BERT models (DeBERTa-v3-base and DistillBERT) and the VADER sentiment analysis. For other purposes (Regex baseline model and final ensemble solution) we downloaded a Swear Word Dataset from Kaggle (https://www.kaggle.com/datasets/tushifire/ldnoobw).

Milestones
For the **First Milestone** we preprocessed and explored the dataset, followed by lemmatization and tokenization. For the **Second Milestone** we implemented baseline models (Regex, Most Frequent, XGBoost, DistillBERT and DeBERTa) and calculated, logged and compared their performance metrics. For the **Final Milestone** we developed an ensemble model that enhanced recall by incorporating rules for female-specific terms, sentiment analysis and a curated database of derogatory terms.

Challenges and Insights

One of the first challenges we faced was dealing with the **data being filled with unnecessary placeholders** in the dataset like an empty link e.g. **[URL]**. Another one of the first challenges we noticed was the **imbalance between labels** categorized as sexist and non-sexist.
After running our baseline models we faced new challenges especially within our results focusing on the false negatives. Among them were several texts where the **text** was **written in a sarcastic or stereotypical way** which turned out to be a **huge obstacle for our models.** Another key challenge was **identifying** whenever **slang expressions** were inserted in the texts. Understanding them was very difficult for our models and many

misclassification cases could be traced back to them. For our final solution we focused on trying to solve this issue and with that improve our score for recall.

Key Results

For the second milestone we needed to work on some baseline models to detect sexism. We trained all these baseline models on the assigned training set and predicted for the test set. Our best baseline model turned out to be **DeBERTa-v3-base** since it achieved the highest accuracy and also the highest recall - that was the performance measure we mostly focused on, since we wanted the model to try to detect the sexist comments more than we wanted it to detect those that are not sexist.
After identifying our key challenges based on the wrongly classified instances, we came up with a solution that can mostly improve our recall score. Our solution is an **ensemble** model based on our best baseline model **DeBERTa-v3-base**. After we used the model to predict the labels, we reevaluated the non-sexist predictions in the following way: if a text was female specific and either had negative annotation or had bad words in it, we overwrote the non-sexist prediction to sexist. However our ensemble model achieved lower scores overall, the recall got better - which was our goal. The scores can be seen in Table 1.

	DeBERTa-v3-base	Ensemble Test Set	Ensemble on own set (n=60)	DeBERTa-v3-base on own set (n=60)
Precision	0.7235	0.5364	0.8888	0.8750
Recall	0.6738	0.7470	0.2857	0.2500
F-Score	0.6978	0.6244	0.4324	0.3889
Accuracy	0.8485	0.7668	0.6500	0.6333

Table 1

After creating our ensemble model, we still needed to try this final solution on an independent dataset to see how it works in ‘real life’. For this independent dataset we collected 60 social media comments from several sites and labelled them ourselves. We then ran our best baseline and our ensemble solution on this dataset expecting somewhat worse results than before. What we did not expect was the very low recall score - it can be, because we only tried to focus on negativity against women and not the stereotypes. From the 60 comments 28 were sexist and from these 28 the model only detected 8. With this result we can say that our solution worked well on our original dataset but it could not generalize well. But our final solution worked better than the baseline on this new test set.

Possible next steps

One of the next possible steps could be the enhancement of the dataset with additional annotated data focusing on sarcastic, stereotypical and slang filled content. Another one of these steps could be the real life implementation of our model, for example finding a reddit thread which we could scrape daily and further analyse our model on and repeat the process beginning from Milestone 2 to further develop our model. The workflow of the ensemble model could also be further developed with steps focusing on other challenges or with more wrought solutions for detecting sarcastic or stereotypical comments against women.