194.093 NLP and IE — Milestone 1

Sven Gerloff, Theresa Mayer, Yasmine Khajjou, Thomas Klar

10.11.2024

1 Introduction

The data for the project is obtained from the EDOS project (Explainable Detection of Online Sexism)[1]. The project aims to recognize and classify sexist content in English texts on platforms such as Reddit and Gab. For Milestone 1, the focus was primarily on the `edos_labelled_aggregated.csv` data set, which contains the features `rewire_id`, `text, label_sexist`, `label_category`, `label_vector` and split. `label_categroy` shows whether the text is labeled as sexist or not.

2 Text preprocessing

In order to prepare the dataset for analysis, we took the following steps:

- As described in the lecture, first important libraries were imported and downloads were performed. Based on the pre-existing labels (dev, train, test) in the column split of the
- `edos_labelled_aggregated.csv` data, a split was performed.To prepare the data, the `label_sexist` column was recoded into 0 (not sexist) and 1 (sexist). Additionally, the text part was modified with a "clean function" where `[USER]` is removed. The text is then further cleaned up by applying the `process_text` function and processed with `nlp_pipeline`.
- The sentences are segmented into sentences and tokens. Here, the lemma is extracted for each word, and the POS tag is returned. Returned is the processed document (doc) lemmas, and POS tags. The output is returned in CoNLL-U format.
- The text part was modified with a cleaning function where `[USER]` is replaced with `USERTOKEN` and `[URL]` with `URLTOKEN`. This was done to avoid confusion with actual words in the comments, while still counting their occurrences.
- Punctuation was removed, filtering out strings that do not contain characters a-z.
- Stopword filtering was applied, retaining personal pronouns and other relevant words, while removing common stopwords. For instance, while "she" is typically a stopword, it remains in our custom list due to its relevance in context.
- The text is then further cleaned up by applying the `process_text` function and processed with `nlp_pipeline`. The sentences are segmented into tokens, extracting lemmas and POS tags. The output is returned in CoNLL-U format.

3 Data exploration

Along with preprocessing, we also engaged in data exploration along several different axes in order to understand the data better and document findings that could be potentially relevant for the tasks in the later milestones.

The first thing we looked at is of course the class distribution within the dataset. All rows belong to one of two classes, which are labeled *sexist* and *non-sexist*. The data was split into a training, test and dev set by the original authors. We investigated the class distribution over all three splits. Figure 1 shows that all three splits show the exact same percentage of classes, suggesting that the authors performed a stratified split. Only about a quarter of instances belong to class 1 (*Sexist*), meaning that it will likely be necessary to use class weights and adjust metrics to account for this imbalance.

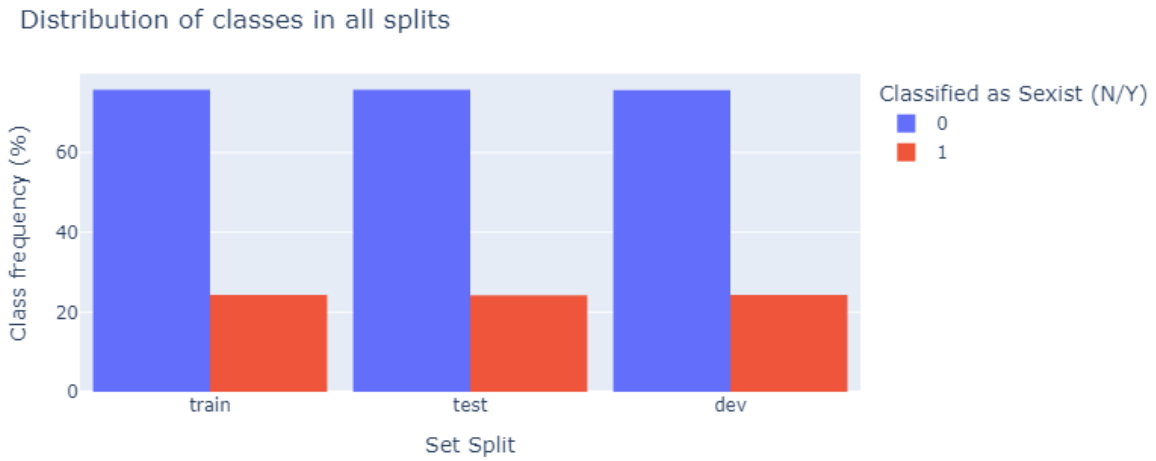


Figure 1: Class Distribution across the three splits.

Another very basic analysis we performed considered the 15 most common words occurring in each class (after lemmatization). It is not surprising that "be", "not" and personal pronouns are the most commonly occurring words. Interesting to note is that "woman" occurs the same number of times in both classes (despite the class imbalance), but "man" is not among the top 15 *Nonsexist* words and "girl" not among the top 15 *Sexist* words. A classifier could be expected to make use of the underlying information.

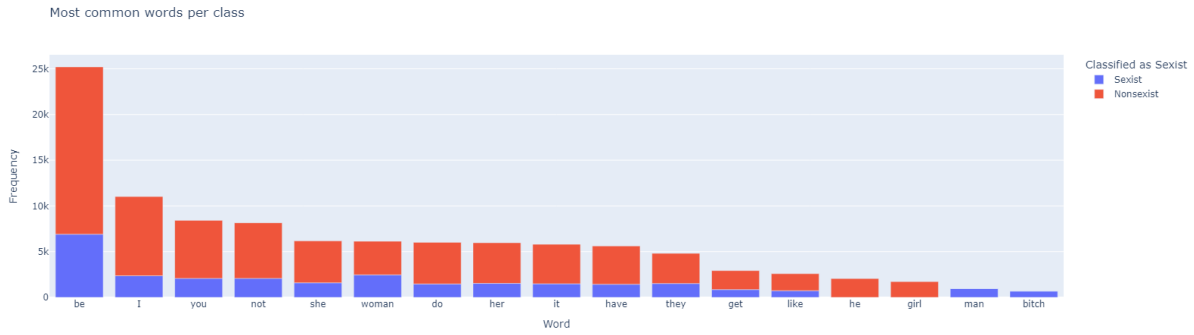
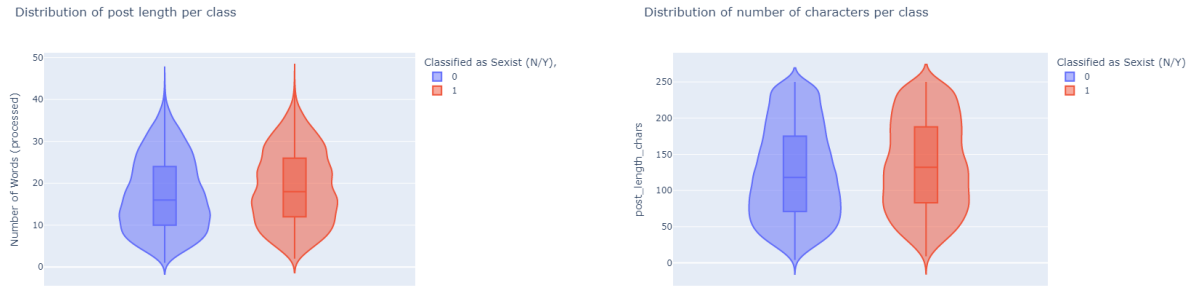


Figure 2: Top 15 most frequent words in both classes

Another feature we can measure directly on the posts that is not captured by the text is the length of the post. We measured this in two ways, firstly by the number of characters

in the unprocessed post and secondly by the number of lemmatized words, with stopwords and punctuation removed. Looking at the results displayed as violin plots in figure 3, we see that the general trend is similar, with two long and tail-heavy distributions with similar endpoints. The median number of words/characters is slightly higher for *Sexist* posts as well as the 25 and 75 percent quantiles. This might however be because for very short posts without context it is hard to classify them as being sexist. Most posts seem to be rather short, with up to 50 significant words and up to 250 characters. This general trend for longer posts to be more likely to be *Sexist* is however very weak and likely not enough to classify on alone.



(a) By number of words, cleaned.

(b) By number of characters, not cleaned.

Figure 3: Distribution of post lengths across both classes

Lastly, we used a Sentiment Analysis tool from the nltk library to perform sentiment analysis of the raw post text. This is a simple but surprisingly effective tool to judge text intent and thus very applicable for our specific problem. Figure 4 shows the resulting distribution. We observe that the *Nonsexist* class is zero-centered (neutral), with the quantiles dipping more into negative than positive. For the *Sexist* class, the median is negative at around -0.3 and also lower quantiles. The upper quantile however is still positive. While it was expected that the 1 class generally exhibits negative sentiment, it is surprising the 0 class does not achieve higher scores. Still, the is a clear trend in the sentiment scores and this feature could reasonably be included in a classifier.

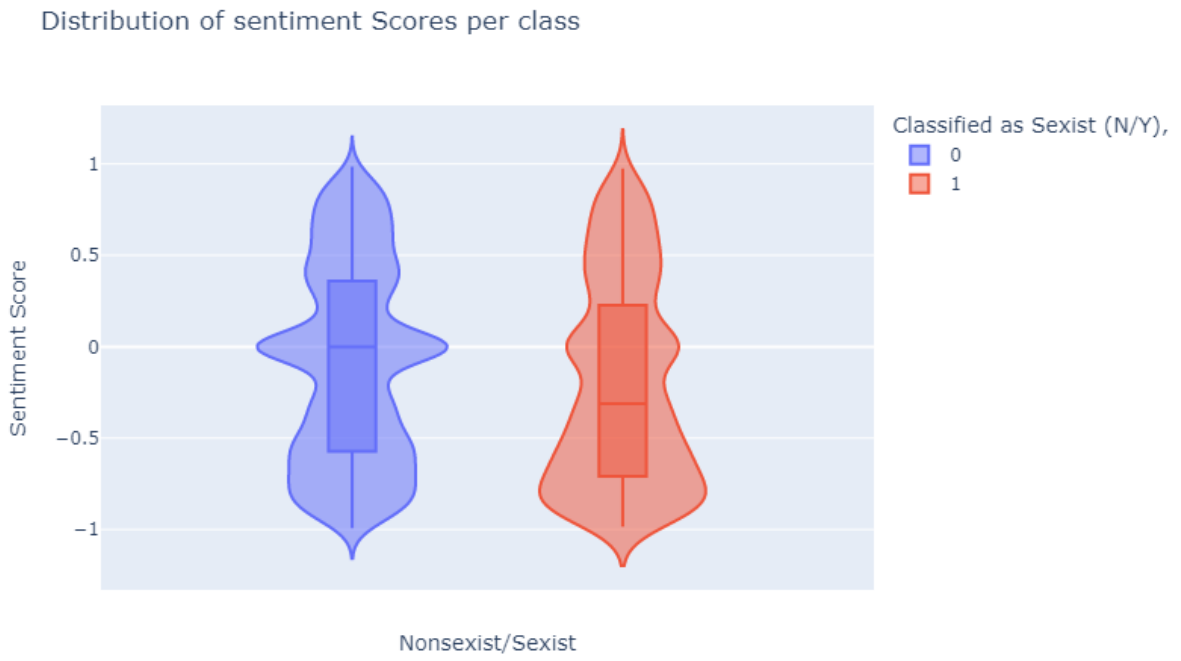


Figure 4: Sentiment scores for both classes

References

[1] Rewire Online. *EDOS Dataset (Labelled Aggregated Data)*. Retrieved from `https://github.com/rewire-online/edos/blob/main/data/edos\_labelled\_aggregated.csv`.