Baseline System for Shared Task SemEval 2023

October 2022

Mehdi Amiri, Behrooz Nikandish, Lynne Zhang
{m.amiri.2, b.nikandish, L.zhang.50 }@student.rug.nl

Introduction This document describes our baseline system architecture for shared task SemEval 2023, task A, which is focused on text misogyny detection. The whole process comprises data pre-processing, modeling, and evaluation. To this end, we examined four different machine learning models in the same setting.

Data preprocessing The training data that the organizers provide is imbalanced. 75.73% of data has been labeled as not sexist (10602) and 24.27% as sexist (3398). We performed seven forms of data preprocessing, namely, lower casing, removing punctuations, stopwords, URL, HTML tags, emojis, emoticons, replacing chat words with actual ones, and lemmatization [1]. We experimented using four combinations of the above pre-processing techniques to see how they impact our results. As shown in table 2, various pre-processing combinations don't differ all that much. Though combination 4 offers the best performance, we go with it.

Models We built different models using four machine learning algorithms, namely, Logistic Regression(LR), Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB), in each of which we used TF-IDF vectorization to create features by extracting information from the data. We use the TF-IDF vectorizer with a limit of 10,000 words capturing uni-grams and bi-grams. Since the data is imbalanced, it is split into the training and development sets with the rate of 80% -20% using stratified 5-fold Cross-Validation in the Scikit-learn library. Thus, we ensure that each class is properly represented in the sample. Finally, we evaluate the models using accuracy and macro averaged F1-score, precision, and recall to compare the results.

Results and Discussion Table 1 shows the results of our experiments. Our findings show that the SVM model outperformed other models in all metrics. Therefore, we select the SVM for the baseline system. However, the model fails to correctly discriminate some sexist terms, according to some samples. The confusion matrix displays 41 False Positive and 418 False

[1] `https://www.nltk.org/_modules/nltk/stem/wordnet.html`

Table 1: The performance of four different machine learning models

Models	Accuracy	F1	Precision	Recall
LR + TF-IDF	65%	83%	68%	82%
SVM + TF-IDF	**69%**	**85%**	**73%**	**84%**
RF + TF-IDF	68%	84%	71%	83%
NB + TF-IDF	54%	55%	51%	54%

Table 2: The performance (F1-score) of four different machine learning models using four combinations of data pre-processing steps: **Combination 1**: lower casing. **Combination 2**: Combination 1 + punctuation and stop words removal. **Combination 3**: Combination 2 + lemmatization + URL / HTML tags removal. **Combination 4**: Combination 3 + emoji removal + replacing chat words.

Models	Combination 1	Combination 2	Combination 3	Combination 4
LR + TF-IDF	64.48%	66.18%	67.52%	67.36%
SVM + TF-IDF	**71.18**	**71.54**	**71.93%**	**71.93%**
RF + TF-IDF	71%	71.30%	71.44%	71.36%
NB + TF-IDF	50.22%	52.00%	51.93%	52.05%

Negative cases after testing the SVM model with a test set of size 2800. For example, among false negative cases, we found seven samples containing the word "bitch" but incorrectly mispredicted as not sexist.

Conclusion We examined four different machine learning models with TF-IDF vectorization as the baseline for the Shared Task SemEval 2023. Our findings show that SVM performed the best, which is aligned with other observations on machine learning algorithm investigations. [Fernández-Delgado et al., 2014, Akinsola, 2017]

References

J E T Akinsola. Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, 48:128 – 138, 06 2017. doi: 10.14445/22312803/IJCTT-V48P126.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real-world classification problems? *The journal of machine learning research*, 15(1):3133–3181, 2014.