**Baseline system for shared task SemEval2023**

Group 2 – Sub-task A

**Mehdi Amiri, Behrooz Nikandish, Lynne Zhang**

We will investigate sexism identification for the baseline system. The whole process comprises data preprocessing, modeling, and evaluation. For this baseline, we examined three different machine learning models and the result is provided in the conclusion section.

Data: Data is imbalanced. Around 75 % of data have been labeled nonsexist and consist of not only alphabets and numbers but also special characters, punctuations, URLs, HTML tags, abbreviations, and emojis. The first step is to clean the data as much as possible. In this step, we take two major steps into account: data cleaning and TF-IDF vectorization.

Data cleaning consists of removing the spacy stop words, additional stopwords removal like #, !, $, deleting URLs and HTML tags, eliminating emojis Unicode includes emoticons, symbols and pictographs, transport and map symbols, and flags.

Then, the TF-IDF vectorization method makes a vector for each word. This vector will be used as a feature and be embedded into the model.

Data is split into training set and development set with the rate of 80% -20 % in a stratified 5-fold cross validation way to select the samples from all classes equally.

Method:

We examined three models(Logistic Regression(LR), SVM, and Random Forest(RF)) for the baseline system each of which with TF-IDF against the data. As the data is imbalanced, the macro average will be considered as an accuracy measure.

All models were examined with a set of parameters as follows that are not fine-tuned.

|  |  |
| --- | --- |
| Model | Parameters |
| LR + TF-IDF | C=5e1, solver='saga', multi\_class='ovr', random\_state=17, n\_jobs=4) |
| SVM + TF-IDF | C=1.0, kernel='linear', degree=3, gamma='auto' |
| RF + TF-IDF | max\_depth=50 |
| Naïve Bayes +TF-IDF |  |

Expriment and Evaluation:

Results:

The results are shown below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Precision | Recall | F1-score | Accuracy |
| LR + TF-IDF | 0.74 | 0.71 | 0.72 | 0.81 |
| SVM + TF-IDF | 0.86 | 0.70 | 0.73 | 0.84 |
| RF + TF-IDF | 0.86 | 0.59 | 0.60 | 0.80 |
| Naïve Bayes +TF-IDF | 0.53 | 0.54 | 0.50 | 0.54 |

Analysis:

From the table above, the SVM and random forest have the best precision. However, SVM shows a better efficiency for recall. Indeed the major difference relates to the recall of the sexist label in the SVM model which is 41% is more efficient than that of random forest which is 19%. Therefore, we select the SVM for the baseline system and will try to design a system to overcome the SVM accuracy.