Literature Review

The study 'Detection and Classification of Sexism on Social Media Using Multiple Languages, Transformers, and Ensemble Models' discusses detecting and categorizing sexist content in social media posts. For this study, the researchers used well-known transformer topologies such as BERT, Roberta, Electra, GPT2, XLM-Align, and Info-XLM. Their research methodology is based on two critical tasks: detecting and categorizing sexism. The study relied on textual data in English and Spanish that included both single-language and bilingual versions. Data collection, data processing, exploratory data analysis, model implementation and hyperparameters analysis, final model implementation, and model comparison are the six systematic processes in this research technique. The models were trained using the dataset EXIST 2022 shared task, which included 11345 labeled posts from Twitter and Gab in English and Spanish, and then evaluated using 1058 tweets from January 2022. The accuracy and f1-Macro metrics were used to measure the model's performance, while Python was employed as the programming language. The models BERT and RoBERTa performed far better than the other designs, and multilingual models outperformed single-language models. RoBERTa received an F1 macro score of 0.954 in English, while Info-XLM had a score of 0.830 in Spanish.

In the paper 'Hate Speech and Counter Speech Detection: Conversational Context Does Matter' the researchers worked on the conversational context in identifying hate speech and counter speech. Hate speech is expressed in a way that degrades any particular individual or group or provokes them to become violent. It often leads to violence and inequity. As the spread of hate speech is increasing day by day, it has become important to take precautions. Two approaches like disruption and counter speech can be effective in this regard. The authors have given a dataset of Reddit comments annotated with hate, neutral, or counter-hate labels that showcased the impact of context on human judgments. Then again, experiments with context-aware classifiers gives us evident improvements in detecting hate speech and counter speech. In his paper a comprehensive analysis was done on language patterns and the benefits of including context in hate speech detection. A neural network model that had a pretrained RoBERTa transformer, that had a fully connected layer for classification purposes. Two textual inputs are taken into consideration: the Target alone and the Parent and Target together. The paper uses a dataset consisting of Gold instances. The dataset was randomly split into training (70%), validation (15%), and testing (15%) sets. Subsequently for training, silver instances are used only. The authors did experiment with two strategies to evaluate performance: blending Gold and Silver annotations and pretraining the

models with related tasks. The results prove that if blending Gold and Silver annotations are used, the performance gets better compared to using only Gold instances. The percentage of F1 weighted average is 61 % and 58 % consequently. If pertaining is done using stance detection data, the result of F1 improves. Incorporating the Parent comment in the training results into higher F1 scores for all classes and a higher weighted average (F1 weighted average: 0.63 for models trained on both Parent and Target, compared to 0.58 for Target-only models). The best performance is achieved by blending Gold and Silver annotations and pretraining with stance detection (F1 weighted average: 0.64). The paper surely has limitations such as considering the parent comment as context and potential biases through keyword sampling. Future works can be done to explore additional context strategies and present biases through community-based sampling.

The paper 'CoRoSeOf - An Annotated Corpus of Romanian Sexist and Offensive Tweets' introduces CoRoSeOf, a large corpus of Romanian social media manually annotated for sexist and offensive language. Detecting sexism is extremely important because it allows us to promote fairness and equality, create spaces where everyone feels safe and respected, prevent harm caused by gender-based discrimination, increase awareness about the issue, hold individuals and institutions responsible for their actions.. The purpose of our study is to address and prevent sexist language online. In this paper the aim was to create a Romanian corpus annotated specifically for sexist language on Twitter. By recognizing and detecting sexist language, users can be protected and fostering more inclusive digital spaces will be possible. To achieve the goals, authors identified the need for a publicly available corpus for classification tasks. They introduce CoRoSeOf, a large Romanian corpus manually annotated for sexist language on Twitter. A team of ten annotators, consisting of seven females and three males, carried out the annotation task. They were all native speakers of Romanian and university students majoring in Languages and Literature or Modern Applied Languages. Before starting the annotation task, the annotators underwent a training period. To annotate the 40,000 tweets, Google Forms was used. The data was randomly split into 160 forms, each containing 250 tweets, and each form was assigned to three annotators. At the end of the annotation process, each tweet received three annotations. The binary classification results on the CoRoSeOf corpus show an accuracy of 83.14% (F1 score), with precision and recall at 83.07% and 83.24% respectively. These scores surpass the best results achieved on the French sexist language corpus. False positives for offensive tweets are 4.69\%, while false negatives occur due to the absence of explicit sexist keywords. The classifier aligns better with

unanimous annotator agreement, resulting in fewer mistakes. Three-way classification into sexist language types also performs well, with macro-averaged precision, recall, and F1 scores at 71.62%, 69.29%, and 70.02% respectively. The dataset shows consistent annotations and will benefit researchers and NLP system development. Future work involves further experiments, improvements, and exploration of different forms of sexism. Data collection will expand to capture a broader range of sexist and offensive language.