A Meta-Model in NLP for Hatefulness

Daniel G. Kyrollos, Department of Systems and Computer Engineering
Ariel Lee, Department of Economics
Elio Velazquez, Project Supervisor
Institute of Data Science, Carleton University, Ottawa, Canada

Introduction

- We present MetaHate, a Natural Language Processing (NLP) meta-model for detecting hatefulness in tweets by combining predictors of hatefulness such as emotion (anger), sentiment (negativity), and offensiveness (offensive).
- We evaluate this meta-model with the TweetEval benchmark for hate speech detection.
- We perform preliminary tests on a real-world dataset: we detect hatefulness in a subset of tweets related to the Black Lives Matter (BLM) movement and its counter-movements, All Lives Matter, and Blue Lives Matter.

What is TweetEval?

- TweetEval is a benchmark for Tweet classification NLP tasks.
- Tasks include hate detection, offensive language detection, emotion detection, sentiment analysis, and stance detection, each with unique labeled datasets.
- It is an open-source, large-scale dataset with 41.8 million tweets containing one of the following keywords: BlackLivesMatter, AllLivesMatter, and BlueLivesMatter.
- The subset of tweets we study are filtered by type (no retweets, no replies) and by language (English only), and the dataset is updated as of March 2021.
- The tweets are restricted to tweets containing at least one of the above keywords.

Real-World Dataset & Limitations

- The real-world dataset we use is a Twitter corpus, which consists of 3,800,000 tweets containing one of the following keywords: BlackLivesMatter, AllLivesMatter, and BlueLivesMatter.
- The subset of tweets we study are filtered by type (no retweets, no replies) and by language (English only), and the dataset is updated as of March 2021.
- The tweets are restricted to tweets containing at least one of the above keywords.

Methodology for Meta-Model

- TweetRoBERTa models for hate detection, offensive language detection, emotion detection, and sentiment analysis were deployed on TweetEval hate speech dataset.
- XGBoost (Xtreme Gradient Boosting) method for model selection was used for the meta-model for its execution speed, its proven success in Kaggle competitions, and its interpretability.
- 5-fold cross-validation was performed using the hate speech training set defined by TweetEval, which was done to find the optimal hyperparameters on an XGBoost with an AUC evaluation metric.
- Full grid-search was performed for the parameters: `max_depth`, `min_child_weight`, `gamma`, and `colsample_bytree`. Total of 3840 parameter combinations were tested, using a score macro-averaged as the scoring method.

Observing the distribution of scores on the test set, it is clear MetaHate has increased the separability of the positive and negative classes. It is also evident that a threshold of 0.5 is not the ideal threshold for accurate classification; a higher threshold is required (0.7-0.8).

Using the TweetEval evaluation framework MetaHate achieves a maximum macro F1-score of 70.3% while the maximum reported score for the TweetRoBERTa model is 55.5%.

Results: TweetEval Benchmark

The most important features other than the hate-based feature were the anger score and the sadness score. Surprisingly, the offensive score feature had a low importance.

Conclusion

- MetaHate combines predictors of hatefulness such as emotion (anger), sentiment (negativity), and offensiveness (offensive).
- Performs better on the TweetEval benchmark than the TweetRoBERTa-base model for hate speech detection.
- Caution should be used when generalizing TweetEval benchmark results to real-world datasets.
- While it is difficult to evaluate unlabeled datasets, studying these results can still help point out where domain expertise would be useful.

Our project highlights the importance of generalizing a result obtained using the TweetEval benchmark.

Acknowledgements

Thank you to professors Majid Komeili, Elio Velazquez, and Michael Genkin from DATAS000, as well as the SGS support team for their help.

References