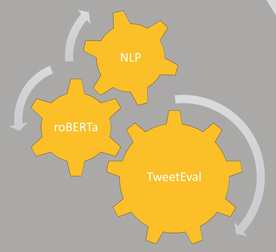


A Meta-Model in NLP for Hatefulness

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

Carleton UNIVERSITY
Canada's Capital University

Introduction

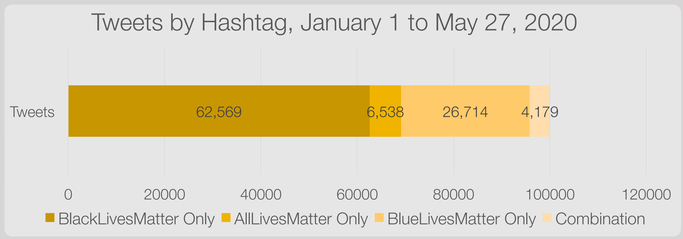
- We present **MetaHate**, a Natural Language Processing (NLP) **meta-model for detecting hatefulness in tweets** by combining predictors of hatefulness such as emotion (anger), sentiment (negativity), and offensiveness (offensive).
  - We **evaluate this meta-model with the TweetEval benchmark** for hate speech detection
  - We **perform preliminary tests on a real-world dataset**: we detect the 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, emoji detection, and stance detection, each with unique labeled dataset
- Unified criteria for splitting train/validation/test data and evaluation of models
- Baseline models were RoBERTa transformer models that were fine-tuned using Twitter data and trained on the downstream task using task-specific dataset

Real-World Dataset & Limitations

- The real-world dataset we use is a **Twitter corpus** [3], an open-source large-scale dataset with 41.8 million tweets which contains 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 public availability as of March 3, 2021.
- The tweets are **filtered temporally**, consisting of tweets starting from January 1 to May 27, 2020.
- A limitation is that we will not be able to evaluate our classification results for the Twitter corpus as it is an unlabeled dataset

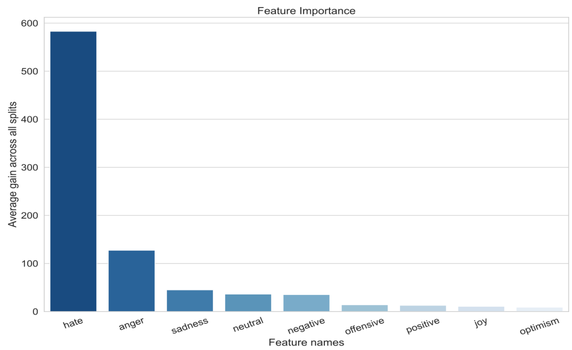


Methodology for Meta-Model

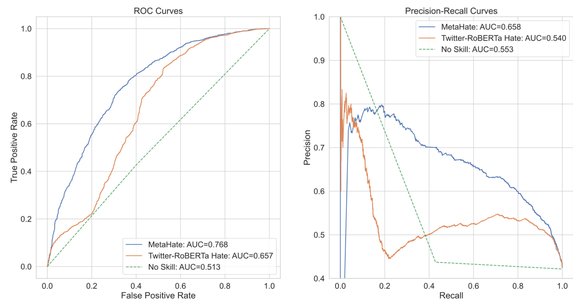
- Twitter-RoBERTa models for hate detection, offensive language detection, emotion detection, and sentiment analysis were deployed on TweetEval hate speech dataset
- XGBoost (eXtreme Gradient Boosting) method chosen for meta-model for its execution speed, its proven success in Kaggle competitions, and its interpretability
- 5-fold cross-validation using the hate speech training set defined by TweetEval was done to find the optimal hyperparameters on an XGBoost with an AUC evaluation metric.
- Full grid-search was performed for parameters: "learning_rate", "max_depth", "min_child_weight","gamma" and "colsample_bytree". Total of 3840 parameter combinations were tested, using F1-score macro-averaged as the scoring method.
- Best performing model had parameters:
  - 'colsample_bytree': 0.4
  - 'gamma': 0.4
  - 'learning_rate': 0.15
  - 'max_depth': 3
  - 'min_child_weight': 3

Results: TweetEval Benchmark

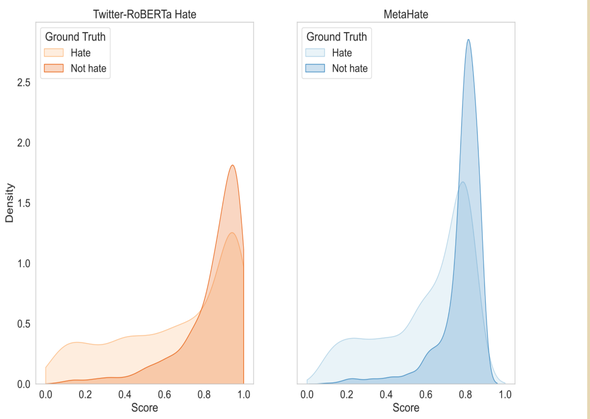
The most important features other than the hate-based feature were the anger score and the sadness score. Surprisingly, the offensive score feature had a low importance.



ROC and precision-recall curves show that the XGboost meta-model, which we have named MetaHate, significantly outperforms the Twitter-RoBERTa model for hate detection.



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)

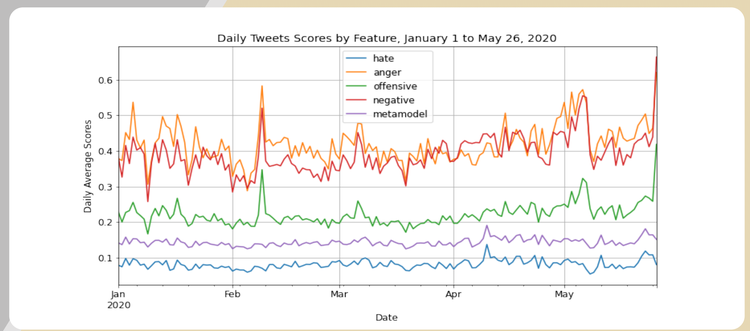


Using the TweetEval evaluation framework MetaHate achieves a maximum macro F1-score of 70.3% while the maximum reported score for the Twitter-RoBERTa model is 55.5%.

Results: Real-World Preliminary

We apply our meta-model to our real-world data subset. Using a threshold of 0.7-0.8 for classification:

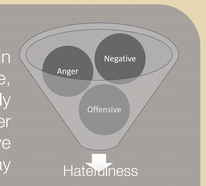
- TweetEval: 2.16–4.18 % of tweets labeled as hateful
- MetaHate: 2.5–7.65% of tweets labeled as hateful



- The TweetEval hate speech training set used on the MetaHate was focused on hate towards women and immigrants. This was reflected in the tweets that our meta-model scored as most hateful.
- MetaHate scores better with the TweetEval benchmark, but caution must be exercised when using the model outside this context.
  - There are tweets classified as hateful by MetaHate (but not TweetEval) that are arguably not hateful, though verification should occur through experts on the BLM movement and counter-movements, race, policing, and hate speech.

Considerations for a Meta-Model

- Studying results from unlabelled data can be useful in pointing out where domain expertise is helpful. For example, some tweets that score high on anger may not necessarily be hateful. An expert could help differentiate types of anger for the next iteration of the MetaHate, such as aggressive anger versus other types of anger, including anger that may be justified.
- Do not underestimate the importance of domain experts in classification and model building



Conclusion

- MetaHate combines predictors of hatefulness such as emotion (anger), sentiment (negativity), and offensiveness (offensive)
  - Performs better on the TweetEval benchmark than the TweetEval pre-trained Twitter-RoBERTa-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 limitations of generalizing a result obtained using the TweetEval benchmark.

Acknowledgements

Thank you to professors Majid Komeili, Elio Velazquez, and Michael Genkin from DATA5000, as well as the SCS support team for their help.

References

[1] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. InProceedings of Findings of EMNLP, 2020.
[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
[3] Salvatore Giorgi, Sharath Chandra Guntuku, Muhammad Rahman, McKenzie Himelein-Wachowiak, Amy Kwarteng, and Brenda Curtis. Twitter corpus of the blacklivesmatter movement and counterprotests: 2013 to 2020, 2020.
[4] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi,Pierric Cistac, Tim Rault, R´emi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick vonPlaten, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, MariamaDrame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural languageprocessing. InProceedings of the 2020 Conference on Empirical Methods in Natural Language Pro-cessing: System Demonstrations, pages 38–45, Online, October 2020. Association for ComputationalLinguistics.