Vote! Don't Retrain: Combating Hate in a Scalable Way

Mohammad Aflah Khan
aflah20082@iiitd.ac.in

Mohit Jain
mohit20221@iiitd.ac.in

Neemesh Yadav
neemesh20529@iiitd.ac.in

Sanyam Goyal
sanyam20116@iiitd.ac.in

Abstract

Social Media has given its users the ability to post anonymously, which has in-turn increased the amount of hate and offensiveness being spread. Hate Speech Detection and its Mitigation has thus become an increasingly important task in the field of Natural Language Processing (NLP) ever since the dawn of social media. We try to aid in the detection of hate speech by building better and more intuitive classifiers over pre-existing ones. The current state of the art works all seem to be throwing huge DL models at the problem without any significant gains while we try to capture different kinds of markers using specialized models and pool them to get the final results. We also show that training an end to end system which is much heavier may not be the best route as we get better performance using isolated systems combined via max-voting over the labels. This provides a more quickly adaptable system which doesn't require total re-training to incorporate new models.

1 Problem Definition

With the exponential growth of social media, the amount of offensive content has also grown significantly. The spreading of such content severely impacts communities, government organizations, individuals, and the platforms themselves. For years, social media platforms have spent millions of dollars on building countermeasures against hate speech. Still, most of them have been directed towards manual work, which is labor-intensive, time-consuming, and non-scalable. One of the most common strategies used now is the amalgamation of computational systems capable of identifying offensive language along with human moderation. Our task is to apply various architectures and find the best ones, which, when given an input text, can categorize it as offensive or non-offensive.

Given a tweet, the task is to identify whether it is offensive or not. Here offensive is used as an umbrella term to determine any form of profane, abusive, hateful, abusive language being used. We have to use the original OLID (Offensive Language Identification Dataset), and although the original dataset has three hierarchical tasks, our project will deal only with labels at the first level, which is a binary classification of whether a tweet is Hateful/Offensive or not.

2 Related Works

Hate Speech Detection Tasks have been studied for a long time now but now more focus has been placed on subtasks such as Implicit Hate Detection (Gao et al., 2017 ;ElSherief et al., 2021), Target Prediction (Sachdeva et al., 2022; Chiril et al., 2022), Explanation Generation (Mathew et al., 2020; Karim et al., 2020) and many others.

Due to the growth in social media and internet percolation a sharp spike in hate speech occurrences has been observed. Especially during the pandemic L1ght reported that there is a 900% increase in attacks against China and Chinese People on twitter. The rise is also significant as ADL's (Anti-Defamation League) 2018 survey showed a 12% increase over Pew Research Center (2017) analysis in 2017.

The dataset used in this work is the **Offensive Language Identification Dataset** (OLID) by Zampieri et al. (2019a). The dataset was used in the SemEval 2019 Competition Zampieri et al. (2019b) where several systems competed to attain the best Macro F1 score. The best performing system finetuned a BERT model with max length 64 and ran the model for 2 epochs and outperformed all other models.

The dataset also appeared in SemEval 2020 Task 12 Zampieri et al. (2020) where the best team used an ensemble of ALBERT models of varying sizes while the second best team used RoBERTa-large fine tuned on SOLID via MLM objective.

Apart from the OLID dataset used in our work there are various other datasets available as well which handle various subtasks of Hate Speech Detection. However most datasets suffer from some common issues such as Annotator Bias (Sap et al., 2022), Niche Distribution which doesn't extend well across datasets (Kim et al., 2022) and Varying Definitions and Terminologies (Schmidt and Wiegand, 2017). There has been work done to create standardized datasets/ checking systems to evaluate model performance however not everyone uses them. **HateCheck** (Röttger et al., 2021) is one such dataset which contains both challenging examples and diverse examples to evaluate the real model performance.

Some works have also tried a Human in the Loop Approach to attack Hate Speech and it's related tasks such as Counter Speech (Chung et al., 2019). Some works have tried incorporating real world knowledge into language models as well using Knowledge Graphs (Sridhar and Yang, 2022) while some other works tried to use prompt based fine tuning to create new data (Hartvigsen et al., 2022) to use for downstream tasks.

We refer to the implementation by Dai et al. (2020) where they use a **multi-task learning (MTL)** setup to train their model to train a MTL model in a similar fashion. We also use the **RoBERTa model** released by Sabry et al. (2022) on HuggingFace to finetune on our OLID dataset. These models were fine-tuned on OLID however we could not reproduce their results by off the shelf loading and hence decided to finetune it ourselves. Across all the works the best performing models are either variants of the same model such as **ALBERT** or use heavier models such as T5 which are also harder to train. We believe that instead of ensembling identical architectures with varying depth and breadth or outright choosing much heavier models we can leverage the wide variety of currently existing models to capture different contexts. As is evident from previous research using **heavy LLMs** is here to stay but we can still use them in a smart way to reduce the load while getting similar results.

Figure 1: Class Distribution for the OLID dataset.

3 Methodology

3.1 Data Preprocessing

We performed the following preprocessing steps on the given tweets; converting tweets to lowercase, replacing @usernames, URLs, and emojis with the <user>, <URL>, and <emoji> tokens, respectively, and removing punctuations, stopwords, and extra whitespaces. Then we split our given dataset into a training and validation set (80:20).

3.2 Data Visualization

We tried various visualization techniques to understand the distribution of the task better. We noticed that the data was somewhat skewed, with a majority of the labels in favor of one class (62% `NOT` Offensive and 32% `OFF`ensive) which we can also see in Figure 1. As with all hate speech datasets, we see a skewed distribution of class labels. However, the skewness in this dataset is not as bad as some other datasets where only around 5-10% of the total dataset is hateful. This prior bias also might contribute to poor results in some cases. We also tried topic modeling (Figures 2, 3) to see if we can see any clear topics across the 2 classes but surprisingly the most popular topics in both datasets are the same. This shows how non-trivial this task is.

3.3 Train-Validation Splitting

The OLID dataset only consists of the train-test samples, so we split our train samples using an 80-20 splitting ratio into training and validation sets.

3.4 Modeling

We first experimented with classical Machine Learning Models namely SVM, MLP, and Perceptron. For each of these models we convert the input

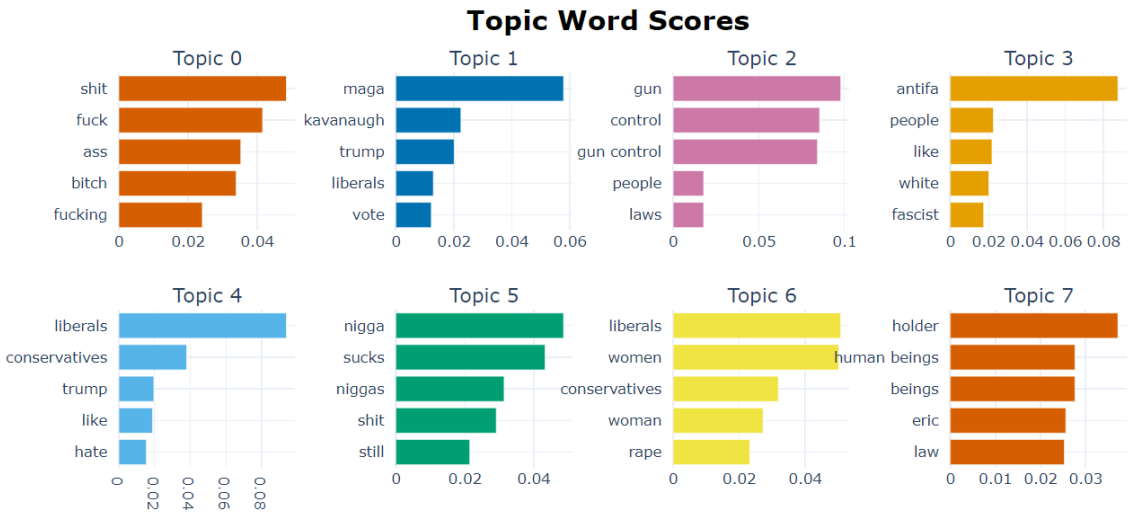

Figure 2: Topics as derived from BERTopic, for the NOT Offensive class.

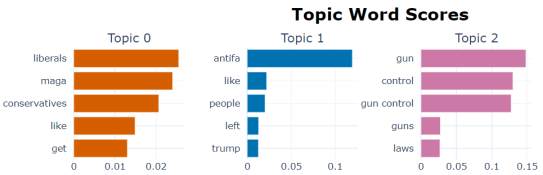

Figure 3: Topics as derived from BERTopic, for the OFFensive class.

text into sentence vectors by either first obtaining word embeddings and averaging them over the sentence in the case of non contextual embeddings or by directly using a Siamese Network exposed via SBERT (Reimers and Gurevych, 2019) to obtain the sentence embeddings. We also try RNN, LSTM and GRU based models for our work.

Finally we use a combination of task-specific BERT variants for extracting different subspace representations which capture different markers. We use BERT (Devlin et al., 2019) to capture high-level semantics without any task specific filtering, HateBERT (Caselli et al., 2020) to capture signals which are indicative of hatefulness or its absence and BERTweet (Nguyen et al., 2020) to capture social media text specific signals. We combine them by concatenating or averaging their last layer representations and only allow the last 3 layers to train to preserve their original skills and avoid catastrophic forgetting. We also experiment with combining them with the HaT5 RoBERTa model and MTL model optionally in a max voting fashion.

We also tried combining external knowledge with the help of commonsense or stereotype knowledge graphs, by concatenating the linearised texts of the tuples to the input sentences, but were unfortunately not able to get any good results.

We rank all our models based on the Macro F1 scores obtained as used in the original SemEval task. We believe these scores are not accurate as we report the scores provided by kaggle which only uses 25% of the data. For all our simple ML baselines we use Grid Search with 2 Fold Cross Validation to obtain optimal hyperparameters. We use the implementation from SKLearn. For obtaining our non-contextual word embeddings we use the loading wrappers provided by Gensim for Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and FastText (Joulin et al., 2016).

4 Experimental Results

We have reported our final best-performing model's Macro-F1 scores in Table 1.

5 Analysis

Overall we observe some clear trends. Simple ML Models accompanied by good sentence embeddings can also outperform simple DL baselines. For instance SVM and Perceptron with FastText are better than any other traditional DL Model. MTL, HateBERT and DistilBERT, BERT.

HateBERT and TweetBERT all perform well alone however combining and training BERT, HateBERT and BERTweet together does not perform as well. This is infact good for us as this is much more costly than performing max voting which we show yields much higher results. Our Best Max Voting Configuration used HateBERT + MTL Model + HaT5 together. Max Voting simply means take the class which is predicted by a majority of the models.

This yields us 2 interesting results:

- There are some signals which are captured by these separately finetuned language models which makes them stand out. Combining these results brings significant gain to our performance
- Simple max voting performs much better than fancy heuristics, is scalable and more explainable as we can see which models think the tweet should be classified as hateful and which ones think it should not.

6 Contribution of each member

All the team members divided the models to be run equally amongst themselves. Aflah and Neemesh performed preprocessing while Sanyam and Mohit performed EDA. Later the model coding for the DL and ML models was equally split among all 4 members and everyone met from time to time to keep track and help each other debug/implement things.

Classifier	Kaggle Reported F1 Score
Perceptron + SentenceTransformer	0.70821
MLP + FastText	0.7621
SVM + FastText	0.7621
RNN	0.64193
GRU	0.74411
LSTM	0.6915
MTL Model	0.82661
HateBERT	0.8199
DistilBERT	0.78348
BERTweet	0.79471
BERT + HateBERT + Bertweet (Averaging Embeddings) with MLP on top	{0.7529}
Max Voting (BERT + Bertweet + MTL Model + HaT5)	0.85833
Max Voting (HateBERT + MTL Model + HaT5)	0.86562
Max Voting (HateBERT + BERT+ MTL Model)	0.86349
Max Voting (HateBERT + BERT+ MTL Model + HaT5)	0.84994
Max Voting (HaT5 + BERT + MTL Model)	0.85512

Table 1: Macro F1 Scores as were reported on Kaggle, for various models, over 4 types of word embeddings (GloVe, FastText, Word2Vec, SentenceTransformer)

References

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. HateBERT: Retraining BERT for abusive language detection in english.

Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. 2022. Emotionally informed hate speech detection: A multi-target perspective. *Cognitive Computation*, 14(1):322–352.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Wenliang Dai, Tiezheng Yu, Zihan Liu, and Pascale Fung. 2020. Kungfupanda at SemEval-2020 task 12: BERT-based multi-TaskLearning for offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2060–2066, Barcelona (online). International Committee for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 774–782, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification.

Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Bharathi Raja Chakravarthi, Md Azam Hossain, and Stefan Decker. 2020. DeepHateExplainer: Explainable hate speech detection in under-resourced bengali language.

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. 2022. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. HateXplain: A benchmark dataset for explainable hate speech detection.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Pew Research Center. 2017. Online harassment 2017. Technical report.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Sana Sabah Sabry, Tosin Adewumi, Nosheen Abid, György Kovacs, Foteini Liwicki, and Marcus Liwicki. 2022. HaT5: Hate language identification using text-to-text transfer transformer.

Pratik Sachdeva, Renata Barreto, Claudia Von Vacano, and Chris Kennedy. 2022. Targeted identity group prediction in hate speech corpora. In *Proceedings*

of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 231–244, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Rohit Sridhar and Diyi Yang. 2022. Explaining toxic text via knowledge enhanced text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–826, Seattle, United States. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval).

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.