Hate Speech Identification

Jake Pistotnik, Vijay Ranganatha (University of California, Berkeley, MIDS Apr-2023)

Abstract

Social media tools like Twitter, Reddit, and Youtube have facilitated the sharing of information, news, content, and communications with individuals from all age groups and walks of life, with minimal filtration of hateful content, making those spaces harmful for children and vulnerable demographic groups. It is thus crucial to be able to identify hateful posts and post spreading hateful messages in order to protect users from harassment. This paper presents the results of identifying and classifying hateful speech in social media posts. The work consisted creating models by leveraging and fine tuning modern, state-of-the-art, pre-trained transformer networks (BERT, ensemble models BERT with a CNN, and fine tuned models like roBERTa and a 2 stage roBERTa model) in order to identify posts which contain hateful content, as well as build a database to score users on their hatefulness based on their past posts. Results indicate a large improvement over the baseline of 35%, achieving 86% F1 score. The implications of this research extend to social media companies, and identifying hateful posts and hateful users to make their platforms safer and more inclusive places for productive interactions.

Introduction

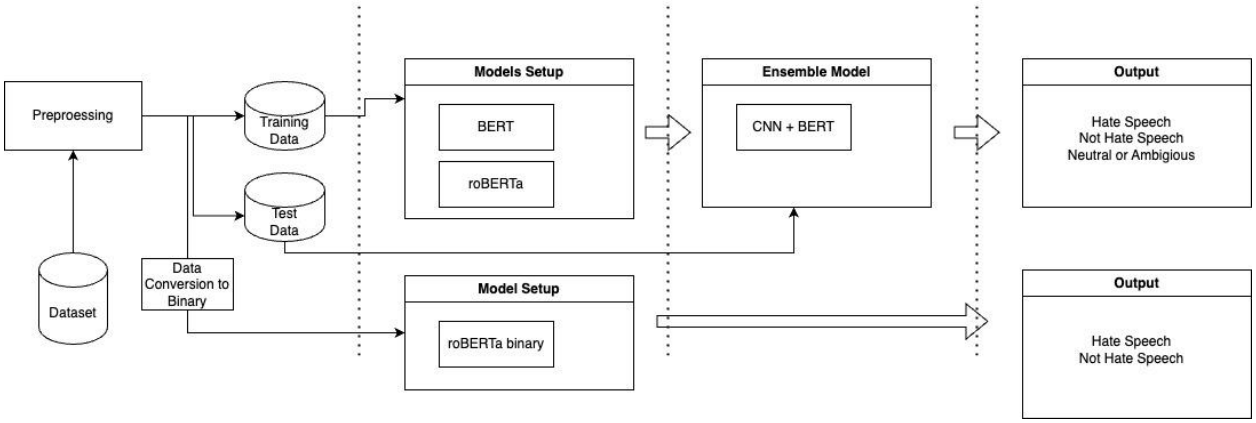
Social Media has grown rapidly since its inception in the late 2000s and has provided the world with a platform to opine, debate, display and discuss like never before. It has a major influence in research areas that analyze human behavior and social groups, and the phenomena of social interactions is even being used in areas such as the Internet of Things and connected devices. (Ortiz-Ospina, 2019) This constant stream of data connecting individuals and organizations across the globe has had a tremendous impact on the functioning of society and even has the power to sway elections.(Thomas, et al., 2022) Despite having numerous benefits, social media has certain issues, especially in the divisive political and social climate that exists in 2023, leading to the rise of hate speech posts on these platforms. Due to resource constraints, respect to freedom of speech, and a rise in content throughout these social media platforms, these issues continue to exist with little identification and repercussions. This leads to cyberbullying, defamation, and presents a concern to the health and well being of users. (United Nations, 2023)

Hate Speeches are poorly tracked, and mitigation efforts cannot be effective without properly identifying a user comment as discriminatory or not. U.S. residents alone experienced an average of 250,000 hate crime victimizations each year from 2004 to 2015 [1].
in the field of hate-speech text detection, our contribution in this paper is twofold:

1. We compare different pre-trained transformer-based neural network model's performance and explain model performance. This paper presents evaluation of hate speech detection methods to provide insights into their detection accuracy
2. An ensemble of the different models presented, including CNN+BERT, and a 2-step roBERTa Classifier.

Methodology

The overall experimentation methodology includes a three-stage process: (i) data collection (ii) Pre-Processing (iii) experimenting using machine learning (ML) models. The experiment environment is the same for all experiments (e.g., data preprocessing, ML architecture, test data). See Figure 1 for a high level description of our methodology whose details are presented in the following subsections



Dataset

In order to train our models and compare our results, we leveraged a the publicly available subset of the dataset described in Kennedy et al. (2020) and Sachdeva et al. (2022), consisting of 39,565 comments annotated by 7,912 annotators, for 135,556 combined rows. The primary outcome variable is the "hate speech score" [link to dataset]. The dataset has comments sourced from three major social media platforms: YouTube, Twitter, and Reddit. The outcome variable measures hate speech, where higher = more hateful and lower = less hateful.

- > 0.5 is approximately hate speech,
- < -1 is counter or supportive speech
- -1 to +0.5 is neutral or ambiguous.

In the initial phases of preprocessing, we decided to aggregate the 135,556 rows into their unique 39,565 comments by taking the average hate speech score given by the unique annotators. By doing so, we felt that this would help generalize better what was considered a hateful comment as compared to neutral and supportive comments. We also labeled the output variable from continuous float values to categorical values, such that we have one label for each class rather than having a float variable. I.e. All values less than -1 were made 2 for supportive text, the values between -1 to 0.5 were made 0 for neutral text, and greater than 0.5 were changed to 1 for hateful text.

Before moving into the models, we looked at the embedding sequence sizes that came out of a standard whitespace tokenizer. Interestingly, we found that the maximum sequence size for the embeddings was exactly 128, with an average length of 19 tokens . Given this maximum size, we decided to move forward using this maximum embedding length for our BERT embeddings

as a way to minimize training time while ensuring that we are not restricting our models ability to analyze entire sequences.

Models Setup

Current research has moved towards using pre-trained transformer encoding models and CNN models. (Kennedy et. al, 2020) This is due to the state of the art performance they achieve due to the massive corpus they are pre-trained on and ability to contextualize. (Weidemann et. al, 2020) We have used a set of well acknowledged models in sentiment detection tasks. Three types of encoder transformers we are leveraging for our classification are: BERT (Devlin et al, 2018), an ensemble BERT with CNN model, a roBERTa model (Jahan et. al, 2018), as well a 2 step roBERTa Model.

Baseline Model

Given that the amount of posts posted to social media sites would be practically impossible to hand label, we decided to make a baseline model that classifies the type of post based on the proportion in each class in our dataset. Given that the original distribution of classes was 44% Supportive, 23% Neutral, and 32% Hateful, we used those same percentages as our random baseline guesser. This gave us a f1 score of 0.34

Transformer Network Models

We first decided to leverage the BERT (Bidirectional Encoder Representations from Transformers) transformer, as they are the seminal transformer based language model. By applying an attention mechanism, BERT provides the ability to learn contextual relations between words in a text sequence by using a Masked Language Model as well as Next Sequence Prediction (NSP). We also decided to build off of BERT, creating an ensemble network by adding a Convolutional Neural Network to the back end of the model to improve our models ability to accurately classify our text examples between hateful and neutral posts.

We also decided to leverage the roBERTa model, a replication of BERT developed by Facebook, also known as Robustly Optimized BERT, which follows a similar pretraining approach with the following modifications (Liu et al, 2019):

1. training the model longer with bigger batches as well as more, cleaner data while discarding the NSP objective
2. dynamically changes the masking patterns, e.g. taking care of masking complete multiword units.

In leveraging BERT, we wanted to compare Model Performances between the various BERT models in order to determine where we could leverage the different mode

Dataset Adjustments

In order to improve the performance of our models, we made some adjustments to our input text as well as evaluating the original labels given in our dataset. After some evaluation, we made a few main adjustments in order to improve our models accuracy.

We first looked at our labels generated from the various scorers of the posts. We decided to move the bound for hateful score from a score of .5 as described in the original dataset to a score of .3. The reason for this was because after analyzing many of the posts that existed on the edge of hateful and neutral, we found that there were many posts, such as:

- "Number killed from car accidents: 1,000,000. Number killed by the word n****r: 0 Therefore, n****r is no big deal"
    - Hate speech score = .31

Thus we moved the bounds of label, which helped our model immensely in being able to classify between hateful and neutral tweets, increasing overall f1 score by 7% on average across models. For the sake of keeping a platform safe, we felt that this decision to increase the bound on what was considered 'hateful' was justified as a way to, at minimum, flag that content for further inspection.

In addition, we wanted to clean up the input posts in order to remove any data within the text that could distract from the actual content of the post, so we removed all mentions, url/https links, '!flair' and brackets from reddit posts, and extra white spaces from the text in order to have more consistent input data to train on.

We also added a new binary label, '2_class_hate_score,' in order to give us the ability to classify solely between hateful tweets and other tweets. This was done by labeling all tweets with a hate score of >.3 as hateful, while the rest were classified as non-hateful.

Results

The results below are for the various models we tested, in addition to comparing F1 scores before and after the dataset adjustments we made.

Data	Model	Precision	Recall	F1 Score	Pre-Adjustment Precision, Recall, F1
Multiclass Labels	Baseline	.35	.35	.34	**.35**
Multiclass Labels	BERT	.86	.86	**.86**	.70
Multiclass Labels	BERT+CNN	.88	.88	**.87**	.84
Multiclass Labels	roBERTa	.84	.83	**.83**	.73
Binary Labels	roBERTa Binary	.93	.93	**.93**	.92
Binary and Multiclass Labels	roBERTa 2 step	.86	.86	**.86**	.73

One thing to note is that during our experimentation, we tested multiple different model parameters and found no real model improvements were made when changing parameters like the batch size and dropout rate. We did however see a slight boost in model performance when using a learning rate of .00001, as well as adding l1 regularization to our models, as they experienced some overfitting (especially the CNN BERT model) on the training data. Most of the models, as expected, performed better post data adjustments as compared to pre-adjustment performance.

When looking at the mislabeled examples from our models (examples can be found at the latter portion of the Combined Roberta Classification Portion of the notebook), it became clear that much of the difficulty of labeling hateful content came with the edge cases. Many of these comments could be considered hateful (derogatory, racist, sexist, etc.) so one important aspect of improving this model moving forward would be to get a larger sample of hate speech scores labels for each post to get a more generalized picture of if the post is considered hateful or not.

Conclusion

Hate Speech in social media platforms continues to be a pervasive problem, as even the UN has made calls just this year to make social media companies more accountable for the hate speech that exists on their platforms. (United Nations, 2023) It is important for the platforms to proactively identify and flag the content for any incitement. This paper presents an evaluation of Hate Speech detection using BERT and ensemble models, taking us one step closer in understanding the models which are beneficial in flagging hate speech and thus help alleviate the problem. Our conclusion based on the results is that the ensemble model using BERT+CNN achieves a better f1 score for Multiclass than roBERTa for classification and proves to provide a higher accuracy, however it struggled with slight overfitting so we worry about the model's ability to generalize. Overall, both roBERTa and the BERT + CNN model would be helpful in flagging hate speech content, roBERTa holding the edge in terms of generalizability, and the BERT+CNN holding the edge for accuracy and f1 score.

References

[1] - Hate Crime Victimization Report from Department of Justice

[2] - Dataset: https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech

[3] Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing Hate Speech Classifiers with Post-hoc Explanation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5435–5442, Online. Association for Computational Linguistics.

[4] Esteban Ortiz-Ospina (2019) - "The rise of social media". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/rise-of-social-media'

[5] Fujiwara, Thomas, et al. (2022) “Princeton University.” The Effect of Social Media on Elections: Evidence from the United States∗, https://www.princeton.edu/~fujiwara/papers/SocialMediaAndElections.pdf.

[6] G. Wiedemann, S. M. Yimam, C. Biemann, Uhh-lt & lt2 at semeval-2020 task 12: Finetuning of pre-trained transformer networks for offensive language detection, arXiv preprint arXiv:2004.11493 (2020).

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018)

[8] Kennedy, C. J., Bacon, G., Sahn, A., von Vacano, C. (2020, September 22). Constructing interval variables via faceted Rasch measurement and multitask deep learning: A hate speech application. arXiv.org. Retrieved April 16, 2023, from https://arxiv.org/abs/2009.10277

[9] Liu, Yinhan, et al. “Roberta: A Robustly Optimized Bert Pretraining Approach.” ArXiv.org, 26 July 2019, https://arxiv.org/abs/1907.11692.

[10] M. S. Jahan, M. Oussalah, A systematic review of hate speech automatic detection using natural language processing, arXiv preprint arXiv:2106.00742 (2021).

[11] Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The Measuring Hate Speech Corpus: Leveraging Rasch Measurement Theory for Data Perspectivism. In Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, pages 83–94, Marseille, France. European Language Resources Association.

[12] “'Urgent Need' for More Accountability from Social Media Giants to Curb Hate Speech: UN Experts | UN News.” *United Nations*, United Nations, (2023) https://news.un.org/en/story/2023/01/1132232.

[13] Y. Kim, Convolutional neural networks for sentence classification, arXiv preprint arXiv:1408.5882 (2014).

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).