Social Computing Project: Scope Document

Data Augmentation for Hate Speech Detection

Sayar Ghosh Roy* 20171047 **Souvik Banerjee*** 20171094 **Saujas Vaduguru*** 20171098 **Ujwal Narayan*** 20171170

1 Conspectus

In this document, we outline the overall scope of our proposed work. Our aim is to investigate approaches to incorporating data augmentation techniques into the task of hate speech detection.

We plan to utilize two publicly available hate speech detection datasets collected from the Reddit and Gab platforms (Qian et al., 2019) having binary classification labels for each post. We prefer the use of source text which is grammatical and closer to formal English as compared to the social media style of short texts typically seen in Tweets. We also hope to evaluate our established approaches on a Twitter-based hate speech detection dataset.

We plan to study the following three promising approaches for data augmentation, namely, (a) Various lexical augmentation methods by Rizos et al. (2019), (b) Cao and Lee (2020)'s reinforcement learning-based HateGANs, and (c) Variational Auto-Encoders for natural language generation by Bowman et al. (2016), in detail and select two suitable methods from these works as our baselines. We then hope to introduce thought-out variations into our implemented baselines. We produce updated pipelines which leverage data augmentation for hate speech detection. Finally, we will present a thorough analysis of the behaviour of each augmentation technique and how that translates to downstream improvements in the classification performance.

2 Problem Statement

Our primary problem is one of hate speech detection in English. We can formally define the task as: given a piece of text, compute a binary label indicating the presence of hate speech i.e label it as '1' if the text contains at least one instance of hate speech, and else, label it as '0'. There are multiple definitions of hate speech in the existing literature. For our task, we stick to the one provided by the United Nations which goes as follows. Hate speech encompasses "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor."

Our problem scope includes the usage of data augmentation techniques to enhance the performance of hate speech detection modules. Based on our literature review, we found two broad classes of approaches for creating additional data which are listed as follows.

1. Alteration-based: These approaches make use of various schemes to perturb token sequences in the existing training data to obtain new sequences (Wei and Zou, 2019). The perturbations are performed in a way such that class labels remain invariant.

2. Generation-based: These approaches make use of explicit natural language generation frameworks including but not restricted to variational autoencoders (Bowman et al., 2016), generative adversarial networks (Cao and Lee, 2020), and language model-based language generation (Rizos et al., 2019).

We plan to understand, analyze, experiment with two existing data augmentation methods, and propose modifications to derive newer techniques for the data augmentation process. The goal being to produce synthetic natural language data in order to increase the number of training examples for the underlying binary classification machine learning models that use encoded input text representations as features.

*Equal contribution. Order determined by roll number.

3 Problem Scope

Throughout our project, we will only focus on hate speech detection i.e. coarse binary classification of cleaned sequences of tokens as hate speech or not. We will not explore fine-grained multi-class classification based on severity or multi-label classification tasks into classes such as sexism, homophobia, anti-Semitism, and Islamophobia. We select two suitable baseline methods for data augmentation from Rizos et al. (2019) and Cao and Lee (2020), and Bowman et al. (2016). We plan to establish baselines on the benchmarks we have selected, and introduce certain modifications to our baselines to arrive at newer data augmentation techniques.

We will make the choice of baselines as well as the modifications based on properties (like style, syntax, etc.) of Reddit and Gab posts, which differ from the text typically seen on Twitter. These platforms allow us to explore the applicability of data augmentation methods to hate speech detection in domains with typically longer text and additional context as compared to Tweets (Qian et al., 2019). We also plan to evaluate our formulated approaches on a Tweet-based hate speech detection dataset[1] (Mandl et al., 2019) to understand how our architectures deal with the social media style of short texts, albeit with limited context and reduced grammaticality.

4 Solution Overview

We will utilize only the cleaned natural language (NL) text as our encoder input throughout all our experiments. Tweets typically contain much higher numbers of non-NL tokens such as hashtags, mentions, emojis, reserved words, etc. as compared to Reddit or Gab posts (Qian et al., 2019) where only a fraction of the posts contain such special tokens. Our lexical data augmentation approaches rely only on NL tokens and we thus make a conscious decision to identify hate conditioned solely on the lexical semantics of a post without considering cues from special tokens such as hashtags.

In recent shared tasks on hate speech detection (Poletto et al., 2020; Mandl et al., 2020, 2019), utilizing Transformer models (Vaswani et al., 2017) pretrained using specific Masked Language Modeling (MLM) objectives on large amounts of natural language have proven useful (Raha et al., 2021; Ghosh Roy et al., 2021) beating the more traditional encoding techniques techniques based on term-frequencies (Gaydhani et al., 2018; Gitari et al., 2015), aggregated word vectors (Arora et al., 2016), and RNNs such as LSTMs and GRUs (Badjatiya et al., 2017; Bisht et al., 2020). For our classification architectures, we will experiment with Transformer based text encoder models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) with classification heads, fine-tuned upon our datasets.

All our augmentation approaches will leverage datapoints from only the examples within the training splits. This is to ensure that we do not indirectly use a sample from the test set for training our models. New training samples are generated from each implemented augmentation technique and we evaluate the contributions of these synthetic training examples to the overall performance of our classifiers. Lastly, we plan to conduct further studies where multiple augmentation techniques are used in synergy.

The modifications we plan to explore fall into the two broad classes we mentioned above. Along the lines of alteration-based techniques, we plan to investigate whether WordNet-based[2] (Miller, 1995) perturbations can be introduced to create newer samples.

To investigate generation-based approaches, we also plan to experiment with State-of-the-art Transformer language models in order to extend LM-based generation approaches. More specifically, we would like to incorporate Transformer encoder-decoder models such as BART (Lewis et al., 2019), and T5 (Raffel et al., 2020) into the model workflows in order to leverage the latest advances in large scale language model pretraining. Overall, our experiments with data augmentation can be divided up into the following four sets.

1. No augmented samples utilized
2. Two existing baseline methods
3. Our proposed modifications
4. Multiple techniques combined

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings.

[1] hasocfire.github.io/hasoc/2019/dataset.html

[2] wordnet.princeton.edu

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 759–760, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Akanksha Bisht, Annapurna Singh, H. S. Bhadauria, Jitendra Virmani, and Kriti. 2020. *Detection of Hate Speech and Offensive Language in Twitter Data Using LSTM Model*, pages 243–264. Springer Singapore, Singapore.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

Rui Cao and Roy Ka-Wei Lee. 2020. HateGAN: Adversarial generative-based data augmentation for hate speech detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6327–6338, Barcelona, Spain (Online). International Committee on 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.

Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat. 2018. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach.

Sayar Ghosh Roy, Ujwal Narayan, Tathagata Raha, Zubair Abid, and Vasudeva Varma. 2021. Leveraging multilingual transformers for hate speech detection.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

T. Mandl, Sandip Modha, P. Majumder, Daksh Patel, Mohana Dave, Chintak Mandalia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. *Proceedings of the 11th Forum for Information Retrieval Evaluation*.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Amit Kumar Jaiswal, Durgesh Nandini, Daksh Patel, Prasenjit Majumder, and Johannes Schäfer. 2020. Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages). In *Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation*. CEUR.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Fabio Poletto, Valerio Basile, M. Sanguinetti, Cristina Bosco, and V. Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. In *LREC 2020*.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Tathagata Raha, Sayar Ghosh Roy, Ujwal Narayan, Zubair Abid, and Vasudeva Varma. 2021. Task adaptive pretraining of transformers for hostility detection.

Georgios Rizos, Konstantin Hemker, and Björn Schuller. 2019. Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 991–1000, New York, NY, USA. Association for Computing Machinery.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.