Social Computing Project: Literature Survey

Data Augmentation for Hate Speech Detection

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

1 Introduction

In this survey, we provide a brief overview of the existing literature in the area of hate speech detection. We also review a collection of data augmentation techniques which have been applied to various related Natural Language Processing tasks. We then look at approaches to solve hate speech detection that use data augmentation methods.

2 Automated Hate Speech Detection

In this section, we will touch upon a variety of methods and procedures applied in attempts to solve the problem of hate speech detection. There are multiple definitions of hate speech in the existing literature. For our task, we stick to the one provided by the United Nations. Early approaches towards detection of hate speech using Bag-of-Words (BoW) models (Kwok and Wang, 2013) typically lead to a high number of false positives and suffer from data sparsity issues. In order to deal with the large number of false positives, efforts were made to better characterize and understand the nature of hate speech itself. This led to the formation of finer distinctions between the types of hate speech (Wang et al., 2014); in that, hate speech was further classified into 'profane' and 'offensive'. Features such as n-gram graphs (Themeli, 2018) or part of speech (POS) features (Chen et al., 2012) were also incorporated into the classification models leading to an observable rise in the prediction scores.

Later approaches used better representations of words and sentences by utilizing semantic vector representations such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). These approaches outshone the earlier BoW approaches as concepts with similar meanings are located closer together in the embedding space and thus, these models could deal with lexical items which were unseen during training. Thus, these continuous and dense representations replaced the earlier binary features resulting in a more effective encoding of the input data. Support Vector Machines (SVMs) with a combination of lexical and parse tree-based features have been shown to perform well for detecting hate speech as well (Chen et al., 2012).

The recent trends in deep learning led to better vector representations of sentences. With RNNs, it became possible to model variable-length sequences of text. Gated RNNs such as LSTMs (Sutskever et al., 2014) and GRUs (Chung et al., 2014) made it possible to better represent long term dependencies. This boosted classification scores, with LSTM and CNN-based models significantly outperforming character and word based N-gram models (Badjatiya et al., 2017). Character-based modelling with CharCNNs (Zhang et al., 2015) has been applied for hate speech classification. These approaches particularly shine in cases where the offensive speech is disguised with symbols like '*', '$' and so forth (Mehdad and Tetreault, 2016).

More recently, attention based approaches like Transformers (Vaswani et al., 2017) have been shown to capture contextualized embeddings for a sentence. Approaches such as BERT (Devlin et al., 2019) which have been trained on massive quantities of data allow us to generate robust and semantically rich embeddings which can then be used for downstream tasks including hate speech detection. Transformer networks pretrained on large multilingual corpora including mBERT (Devlin et al., 2019), XLM (Lample and Conneau, 2019), and XLM-RoBERTa (Conneau et al., 2019) have also proved useful for the task of detection and fine-grained classification of hateful content across a set of languages (Ghosh Roy et al., 2021). Recent ideas around task adaptive pretraining of Transformers (Gururangan et al., 2020) before utilizing them for classification tasks has also proven useful for the task of hostility detection in text

*Equal contribution. Order determined by roll number.

(Raha et al., 2021). Explanations by annotators can also be leveraged to improve hate speech detection (Mathew et al., 2020). Tokens in the rationale of why the post is classified as hate speech are given the value 1 and the rest of the tokens are marked as 0. This is then averaged over all annotators and passed through a softmax function to create the ground truth attention. This ground truth attention is then used to provide attention supervision to attention based approaches such BERT or BiRNN + Attention.

Hate speech detection can also be formulated in a multi-label setting based on the type of hate speech (obscene, toxic etc) or in a multi-class setting based on the severity (highly toxic, moderately toxic, not toxic, etc.). Adaptation approaches such as Multilabel-kNN (Zhang and Zhou, 2007) or HARAM (Benites and Sapozhnikova, 2015) have shown promise with multi-label hate speech detection (Mollas et al., 2020).

There have also been a variety of open or shared tasks to encourage research and development in hate speech detection. The TRAC shared task on aggression identification (Kumar et al., 2018) included both English and Hindi Facebook comments. Participants had to detect abusive comments and distinguish between overtly aggressive comments and covertly aggressive comments. OffensEval (SemEval-2019 Task 6) (Zampieri et al., 2019) was based on the Offensive Language Identification Dataset (OLID) containing over 14,000 tweets. This SemEval task had three subtasks: discriminating between offensive and non-offensive posts, detecting the type of offensive content in a post and identifying the target of an offensive post. At GermEval (Struß et al., 2019), there was a task to detect and classify hurtful, derogatory, or obscene comments in the German language. Two sub-tasks were continued from their first edition, namely, a coarse-grained binary classification task and a fine-grained multi-class classification problem. As a novel sub-task, they introduced the binary classification of offensive tweets into explicit and implicit.

There have been some efforts to construct lexicons for the purpose of aiding the task of hate speech detection. Hurtlex, a multilingual lexicon of hurtful words created by expert annotators leveraging inputs from linguists has proven useful for detecting hate against immigrants and misogyny in tweets (Bassignana et al., 2018). This resource is available publicly[1]. In further experiments, these features were utilized in form of lexicon-based encodings at the sentence-level and as word-level embeddings to improve over baseline BERT models (Koufakou et al., 2020).

The escalation in derogatory posts on the internet has prompted certain agencies to make toxicity detection modules available for web developers as well as for the general public. A notable work in this regard is Google's Perspective API[2] which uses machine learning models to estimate various metrics such as 'toxicity', 'insult', 'threat', etc., given a span of text as input.

A majority of the publicly available datasets on hate speech detection have tweets as their primary data source. Vidgen and Derczynski (2020) review a collection of abusive language datasets focusing on their creation, content, and impact. Their reviewed collection is well maintained and openly accessible[3]. Qian et al. (2019) present two fully-labeled large-scale hate speech intervention datasets collected from Gab and Reddit which provide conversation segments, hate speech labels, as well as intervention responses. Due to the choice of their platforms, the data language style is different from that of tweets and in that, is more grammatical and structurally sound.

3 Data Augmentation

Data augmentation is the process of adding additional samples to the training set to encourage a machine learning model to learn generalisable patterns instead of superficial features that are specific to the training data (Jha et al., 2020). The method has found great success in the computer vision community, and has been used as a part of the training process since the earliest approaches to image classification using deep learning (Krizhevsky et al., 2012). However, while simple transformations like cropping, tilting, and flipping an image work well for augmenting data in vision tasks, language tasks pose a much stronger challenge due to the strong interdependency of syntactic and semantic features in text data (Liu et al., 2020).

There have been a variety of approaches to data augmentation for language tasks. These approaches range from simple meaning-agnostic perturbations of the training data to approaches based

[1] github.com/valeriobasile/hurtlex
[2] www.perspectiveapi.com
[3] hatespeechdata.com

on conditional language generation which can generate entirely new sentences. Here, we review approaches to augmenting data for training text classification models, as this is the principal paradigm for hate speech detection tasks.

Wei and Zou (2019) present a method called Easy Data Augmentation (EDA), where they propose a set of simple transformations that can be applied to generate perturbed samples with the same label. They study replacing words with synonyms (from WordNet (Miller, 1995)) at random, and randomly inserting, deleting, or swapping words. They find that even such simple transformations provide gains in performance on benchmark datasets, while also ensuring that the transformations do not inadvertently cause label changes frequently.

Wang and Yang (2015) use word embedding similarity to replace a word with similar words. They replace words with one of the their k nearest neighbours in the word embedding space. They apply this approach to a text classification task on Twitter data.

While synonym replacement in prior work was performed agnostic to the context of the word, Kobayashi (2018) propose a method to substitute words based on the predictions of a language model obtained based on the left and right context of the word. This allows for contextual information to be incorporated into the augmentation process.

While most of the above approaches transform sentences in the training data to obtain new sentences with the same class label, Liu et al. (2020) propose a reinforcement learning approach that fine-tunes pretrained conditional language models to generate new samples for each class. Their reward model for generation incorporates rewards for words that are closely associated with a particular label, and for fluency as determined by the pretrained language model. They validate the effectiveness of their generation on benchmark classification tasks, and the quality of the generated data with human evaluations.

Xu et al. (2020) perform a systematic comparison of various approaches to data augmentation for text classification, emphasising the importance of data augmentation in cases where the number of samples of each label is not balanced. They investigate approaches that add no additional data like resampling from the existing data, word-level transformations (EDA), sequence-to-sequence generation, and generation with variational autoencoders (VAE) (Kingma and Welling, 2014). They compare all the approaches on the same tasks under the same setting, providing a direct comparison of the methods. They find that augmentation with VAE consistently provides improvements in performance, and is the best performing approach in most tasks. They also investigate the problem of the amount of generated data that is used, and find that there is a point of extremum for each task which provides the highest performance, beyond which point augmentation becomes detrimental.

To elaborate on VAE in a more mathematical sense, it is a probabilistic framework which finds an efficient approximate ML or MAP estimation of given parameters that shape the distribution of i.i.d data samples, and an efficient approximate posterior inference of a latent variable implied by given observed values. In order to fulfill these objectives, Kingma and Welling (2014) provide a lower bound for the likelihoods in question and derives a loss function that interprets a KL-divergence term as regularizer. Moreover, an algorithm on how to update the parameters during training (auto-encoding VB algorithm) is provided.

4 Data Augmentation for Hate Speech Detection

A straightforward way of tackling the problem of hate-speech detection would involve a supervised approach that heavily depends on labeled datasets for training, which turns out to be a challenge. Existing hate speech datasets are highly imbalanced, as shown by Davidson et al. (2017). The researchers manually annotated a large Twitter corpus to differentiate offensive tweets from hate tweets. However less than 12% of the total data was labeled as hateful. Thus, it becomes necessary to perform data augmentation before carrying out any neural network-based training for hate speech detection.

Rizos et al. (2019) conduct the first study of data augmentation methods for the task of hate speech detection. They present three methods of data augmentation, in combination with methods to harness deep learning approaches to perform better hate speech detection. These approaches are called *ThreshAug*, *PosAug*, and *GenAug* respectively. *ThreshAug* performs word substitution with words that have embeddings with a cosine similarity above a fixed threshold. *PosAug* shifts and warps tokens within a padded sequence to provide

additional samples with the same syntax and semantics. *GenAug* uses an RNN-based language model trained on the data belonging to a specific class to generate additional samples of that class. They find that *ThreshAug* and *PosAug* improve the performance of their detection models, but do not see benefits with *GenAug*, which they attribute to the simplicity of the approach, and the small size of the training data used to train the language models from scratch.

One effective method for such data augmentation is proposed by Cao and Lee (2020). Their model named HateGAN adopts a reinforcement learning based generative adversarial network architecture to generate hate speech for data augmentation. The discriminator is trained to guide the generator to synthesize tweets that are indistinguishable from the real tweets. Since the end goal is to use the generated tweets for hate speech detection, it utilises a metric which rewards the hateful sentiment of the tweets. Therefore, a pretrained toxicity scorer quantifies the hatefulness of the synthesized tweets as hate scores. The synthesized tweets are also scored with respect to how 'realistic' they are. Subsequently, the realistic scores and hate scores are used as rewards to guide and update the parameters in the generator for more realistic hateful Tweet generation. Moreover, the model is trained with policy gradient to overcome the problem of differentiation in sequence generation.

Delving somewhat deeper into the model, the architecture consists of a word embedding layer followed by a LSTM, topped off with a fully connected layer. The LSTM layer is made up of two stacked LSTM sequences. Maximum and average pooling operations are applied to all hidden states of the second LSTM layer. The two vectors are connected to a fully connected layer to generate the final vector for multilabel classification into six polarities: 'toxicity', 'obscene', 'threat', 'insult', 'identity attack', and 'sexually explicit'. As for the generator, it adopts a sequence generation framework where reinforcement learning and monte carlo search is utilized. The discriminator here is a binary classifier trained to evaluate the 'realisticness' of the generated sentence. The discriminator weights are optimized to distinguish the generated tweets from the real ones.

References

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.

E. Bassignana, Valerio Basile, and V. Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *CLiC-it*.

F. Benites and E. Sapozhnikova. 2015. Haram: A hierarchical aram neural network for large-scale text classification. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 847–854.

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.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

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.

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

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks.

Rohan Jha, Charles Lovering, and Ellie Pavlick. 2020. Does data augmentation improve generalization in nlp?

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105. Curran Associates, Inc.

Ritesh Kumar, Atul Kr. Ojha, Marcos Zampieri, and Shervin Malmasi, editors. 2018. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Association for Computational Linguistics, Santa Fe, New Mexico, USA.

I. Kwok and Y. Wang. 2013. Locate the hate: Detecting tweets against blacks. In *AAAI*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Ruibo Liu, Guangxuan Xu, Chenyan Jia, Weicheng Ma, Lili Wang, and Soroush Vosoughi. 2020. Data boost: Text data augmentation through reinforcement learning guided conditional generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9031–9041, Online. Association for 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. *arXiv preprint arXiv:2012.10289*.

Yashar Mehdad and Joel Tetreault. 2016. Do characters abuse more than words? pages 299–303.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Ethos: an online hate speech detection dataset.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

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

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.

Julia Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Sissy Themeli. 2018. *Hate Speech Detection using different text representations in online user comments*. Ph.D. thesis.

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.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2014. Cursing in english on twitter. In *Proceedings of the 17th ACM conference on*

Computer supported cooperative work & social computing, pages 415–425.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, Lisbon, Portugal. Association for Computational Linguistics.

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.

Binxia Xu, Siyuan Qiu, Jie Zhang, Yafang Wang, Xiaoyu Shen, and Gerard de Melo. 2020. Data augmentation for multiclass utterance classification – a systematic study. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5494–5506, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Min-Ling Zhang and Zhi-Hua Zhou. 2007. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.