Towards Hate Speech Detection at Large via Deep Generative Modeling

Tomer Wullach
University of Haifa

Amir Adler
Braude College of Engineering and
Massachusetts Institute of Technology

Einat Minkov
University of Haifa

Abstract—Hate speech detection is a critical problem in social media platforms, being often accused for enabling the spread of hatred and igniting physical violence. Hate speech detection requires overwhelming resources including high-performance computing for online posts and tweets monitoring as well as thousands of human experts for daily screening of suspected posts or tweets. Recently, Deep Learning (DL)-based solutions have been proposed for automatic detection of hate speech, using modest-sized training datasets of few thousands of hate speech sequences. While these methods perform well on the specific datasets, their ability to detect new hate speech sequences is limited and has not been investigated. Being a data-driven approach, it is well known that DL surpasses other methods whenever a scale-up in train dataset size and diversity is achieved. Therefore, we first present a dataset of 1 million realistic hate and non-hate sequences, produced by a deep generative language model. We further utilize the generated dataset to train a well-studied DL-based hate speech detector, and demonstrate consistent and significant performance improvements across five public hate speech datasets. Therefore, the proposed solution enables high sensitivity detection of a very large variety of hate speech sequences, paving the way to a fully automatic solution.

I. INTRODUCTION

The rapidly growing usage of social media platforms and blogging websites has led to over 3.8 billion [1] active social media users worldwide (half of the world population) that use text as their prominent means of mutual communications. A fraction of the users use a language that expresses hatred towards a specific group of people, known as Hate Speech [2]–[6]. Hate speech is mostly used against groups of common race, ethnicity, national origin, religious affiliation, sexual orientation, caste, gender, gender identity, and serious disease or disability. Hate speech can be expressed explicitly, for example: "Refugees are the illness of the world and must be cured with antibiotics." or implicitly: "Black women on London transport...", which leads to a very large diversity of realistic hate speech sequences. Therefore, hate speech detection has become a critical problem for social media platforms, being often accused for enabling the spread of hatred and igniting physical violence in the real world. Hate speech detec-
Hate speech detection has been investigated extensively in recent years, and due to the outstanding success of Deep Learning (DL) [8] in natural language processing [9] it has become a leading approach for solving this problem. Among the proposed DL architectures, the most well-studied ones such as [10], were adopted from the related problem of sentiment analysis [11]. The main reason for the success of DL is its Generalization capability. Generalization [12] is defined as the ability of a classifier to perform well on unseen examples. In the supervised learning framework a classifier \( C : \mathcal{X} \rightarrow \mathcal{Y} \) is learned using a training dataset \( \mathcal{S} = \{(x_1, y_1), ..., (x_N, y_N)\} \) of size \( N \), where \( x_i \in \mathcal{X} \) is a data example (i.e. a text sequence) and \( y_i \in \mathcal{Y} \) is the corresponding label (i.e. hate or non-hate). Let \( \mathcal{P}(\mathcal{X}, \mathcal{Y}) \) be the true distribution of the data, then the expected risk is defined by \( \mathcal{R}(C) = E_{x,y \sim \mathcal{P}(\mathcal{X}, \mathcal{Y})} [L(C(x), y)] \), where \( L \) is a loss function that measures the misfit between the classifier output and the data label. The goal of DL is to find a classifier \( C \) that minimizes the expected risk, however, the expected risk cannot be computed since the true distribution is unavailable. Therefore, the empirical risk is minimized instead: \( \mathcal{R}_E(C) = \frac{1}{N} \sum_{i=1}^{N} L(C(x_i), y_i) \), which approximates the statistical expectation with an empirical mean. The Generalization Gap is defined as the difference between the expected risk to the empirical risk: \( \mathcal{R}(C) - \mathcal{R}_E(C) \). By using large and diverse datasets, DL has been shown to achieve a low generalization gap, where an approximation of the expected risk is computed using the learned classifier and a held-out testing dataset \( \mathcal{T} = \{(x_1, y_1), ..., (x_M, y_M)\} \) of size \( M \), such that \( \mathcal{S} \cap \mathcal{T} = \emptyset \). The implication of achieving a low generalization gap is good performance of the classifier on unseen data, namely, detecting accurately unseen hate speech.

Existing hate speech detection solutions utilize public datasets with representative hate and non-hate sequences, collected from social media. These datasets contain only a few thousands of sequences, thus, with a limited capability to faithfully represent the large diversity of hate and non-hate sequences. Therefore, while good results are achieved for existing solutions, these are limited only to the scope of the utilized datasets, and have not been tested on a larger variety of hate/non-hate sequences.

In this paper we present a novel approach for significantly increasing the generalization capabilities of hate-speech detection solutions. The performance of the proposed approach is evaluated using a well-studied DL-based detector, trained using five different public datasets. First, we scale up significantly the training datasets in terms of size and diversity by utilizing the GPT-2 [13] deep generative language model. Each one of the five datasets was used to fine-tune the GPT-2 model, such that it generated 100,000 hate and 100,000 non-hate speech sequences, per dataset. Following the data generation stage, hate speech detection performance was evaluated by augmenting the baseline training sets with the generated data, demonstrating consistent and significant performance improvements of the DL detector, as compared to training using only the baseline (un-augmented) datasets.

The contributions of this paper are three-fold: (i) the utilization of deep generative language modeling for approximating the probability distribution of hate speech and generating massive amounts of hate speech sequences. (ii) a comparative study of the achievable generalization of hate speech detection using existing hate speech datasets. (iii) a detailed demonstration of significant hate speech detection generalization improvements, by utilizing the generated data.

The rest of this paper is organized as follows: an overview of hate speech detectors and datasets is provided in the related work section. The generation of hate and non-hate speech by the deep generative model is detailed in the proposed approach section. Hate speech detection results and a study of the achievable generalization are
provided in the performance evaluation section. The conclusions section summarizes the paper.

Related work

Researchers have proposed several approaches for hate speech detection over the recent years, ranging from classical learning methods, to modern deep learning. While these methods differ with respect to architecture and data representation, they share the need for a large, diverse, and high-quality data resource for training effective classification models. In the lack of such a resource, previous related works constructed datasets for this purpose, which were manually labeled, and therefore strictly limited in size. Below, we review existing hate speech detection approaches, and describe the datasets that have been used by researchers for training and evaluating such classifiers. We will leverage those labeled datasets which are publicly available for generating our MegaSpeech corpus.

Hate Speech Classifiers

Classical machine-learning methods require data to be transformed into some pre-defined feature representation scheme, which should ideally capture meaningful and relevant information for classification purposes. Given textual data, features typically pertain to raw lexical information (surface word forms), considering individual words as well as word sequences (n-grams). Other features may encode processed syntactic and semantic information. For example, Davidson et al. [14] represented tweets using word n-grams, having them weighted by word frequency (TF-IDF), part-of-speech and sentiment features. In addition, they encoded genre-specific features, denoting hashtag and user mentions, URLs, and the tweet’s length. They then trained a logistic regression classifier to distinguish between tweets labeled as hate or offensive speech, or neither. In another work, a Support Vector Machine (SVM) was trained to predict whether a tweet was hateful or not using feature representations of word n-grams and typed word dependencies [15].

Deep Learning (DL) methods can effectively exploit large-scale data, learning latent representations using multi-layered neural network architectures. Modern DL architectures of text processing generally consist of a word embedding layer, intended to capture generalized semantic meaning of words, mapping each word in the input sentence into a vector of low-dimension [16]. The following layers learn relevant latent feature representations, where the processed information is fed into a classification layer that predicts the label of the input sentence.

In a recent work, Founta et al. [17] experimented with a popular DL architecture for text processing, classifying tweets into the categories of hate, abusive, or offensive speech, sarcasm and cyberculling. They transformed the input words into Glove word embeddings [18]. They then used a recurrent layer comprised of Gated Recurrent Units (GRU) for generating contextual and sequential word representations, having each word processed given the representations of previous words in the input sentence. Following a dropout layer (intended to prevent over-fitting), the final dense layer outputs the probability that the sentence belongs to each of the target using a softmax activation function. Another popular choice of DL architecture for processing textual data is Convolution neural networks (CNN). Previously, Badjatiya et al. [19] used such an architecture for hate speech detection. CNN applies a filter over the input word representations, processing local features per fixed-size word subsequences. Hierarchical processing, which involves aggregation and down-sampling, consolidates the local features into global representations. These representations are finally fed into a fully-connected layer which predicts the probability distribution over all of the possible classes.

Another DL classification architecture presented by Zhang et al. [10] is comprised of both convolution and recurrent processing layers. This architecture has been shown to yield top-performance results on the task of hate speech classification, and is therefore the architecture of choice in our experiments. In brief, and as illustrated in Figure 1, this text classification network consists of a convolution layer applied to individual dimensions of the input word embeddings, which is down-sampled using a subsequent max-pooling layer. The following layer is recurrent (GRUs), producing hidden state representations per time step. Finally, a global max-pooling layer
is applied, and a softmax layer produces a probability distribution over the target classes for the given input. Further implementation details are provided later in this paper.

Hate speech datasets

Table 1 details the statistics of five public datasets of hate speech, which are all manually curated and of modest size. Below, we discuss the generation process and characteristics of these datasets. In this work, we will leverage these datasets for tuning an automatic language generation process, resulting in a very large number of hate speech sequences, namely the Mega-speech resource.

The typical process that has been traditionally applied by researchers and practitioners for constructing labeled datasets in general, and of hate speech in particular, involves the identification of authentic hate speech sentences, as well as counter (non-hate) examples. Twitter is often targeted as a source of relevant data; this (and other) public social media platform allows users to share their thoughts and interact with each other freely, making it a fertile ground for expressing all kinds of agendas, some of which may be racist or hateful. The initial retrieval of hate speech examples from Twitter is based on keyword matching, specifying terms that are strongly associated with hate. Once candidate tweets are collected, they are assessed and labeled by human annotators into pre-specified categories. The manual annotation of the examples provides high-quality ground truth labeled datasets, yet it is costly

Table 1. Publicly available hate speech datasets, pertaining to examples labeled strictly as hate or non-hate.

| Dataset | Source | Hate | Non-Hate |
|---------|--------|------|----------|
| DV [14] | Tweets | 1,404 | 7,875 |
| FN [22] | Tweets | 4,020 | 49,775 |
| WS [20] | Tweets | 1,965 | 3,797 |
| WH [23] | Posts | 1,196 | 9,507 |
| SE [21] | Tweets | 4,217 | 5,758 |

We now turn to describe the individual datasets in Table 1 in more detail. The dataset due to Davidson et al. [14] (DV) includes tweets labeled by CrowdFlower\(^3\) workers into three categories: hate speech, offensive, or neither. For the purposes of this work, we only consider examples of the first and latter categories. Waseem and Hovy [20] created another dataset (WS), considering tweets of accounts which frequently used slurs and terms related to religious, sexual, gender and ethnic minorities; those tweets were manually labeled into the categories of racism, sexism or neither. Again, we only consider the first and latter categories as examples of hate and non-hate, respectively. Another dataset was constructed by SemEval conference organizers [21] for the purpose of promoting hate detection (SE). They considered the historical posts of identified hateful Twitter users, narrowed down to tweets that included hateful terms, and had examples labeled by CrowdFlower workers. We find that many of the tweets labeled as hateful in this dataset target women and immigrants. Founta et al [22] (FN) performed iterative sampling and exploration while having tweets annotated using crowdsourcing. Their resulting dataset is relatively large (~80K examples), and distinguishes between multiple flavors of offensive speech, namely offensive, abusive, hateful, aggressive, cyber bullying, spam and none. In order to maintain our focus on hate speech, we consider the labeled examples associated with the hateful and none categories. Finally, the dataset due to de Gilbert et al. [23] (WH) was extracted from the extremist StormFront Internet forum.\(^4\) This dataset aims to gauge hate in context, considering also cases where a sentence does not qualify as hate speech on its own, but is interpreted as hateful within a larger context comprised of several sentences.

\(^1\)https://www.welcome.ai/crowdflower
\(^2\)https://en.wikipedia.org/wiki/Stormfront_(website)

\(^3\)https://www.welcome.ai/crowdflower
\(^4\)https://en.wikipedia.org/wiki/Stormfront_(website)
The proposed approach: Scalable Hate Speech Detection via Data Generation

Hate Speech Data Generation

Ideally, DL-classifiers would be trained using large and diverse datasets, which more closely represent the true distribution of hate speech. We automatically develop large-scale datasets of hate speech text sequences and counter examples from seed labeled examples using the GPT-2 model [13].

The Generative Pretrained Transformer 2 (GPT-2) is a language model, trained to predict, or “generate”, the next token in a sequence given the previous tokens of that sequence in an unsupervised way. Being trained using 40GB of Web text, GPT-2 automatically generates text sequences of greater quality than ever seen before (CITE). We employed the large configuration of GPT-2 (764M parameters) in this work.

Our aim is to generate hateful rather than general text sequences, however. Concretely, we wish to generate hate (and, non-hate) speech sequences that: a) are perceived by humans as hate speech (or, non-hate) with high probability, b) introduce high language diversity, and c) exhibit high language fluency. In order to meet the first requirement, we exploit available labeled examples, tuning GPT-2 towards hate (and non-hate) speech generation. High diversity is achieved by leveraging the rich language model encoded into GPT-2, and by generating a very large number of candidate sequences. Last, we maintain high data quality by automatically evaluating the generated candidate sequences using a state-of-the-art text classifier, namely BERT [24], where generated sequences that are evaluated as low-quality texts by BERT are automatically removed. Next, we formalize and discuss each of these components of our approach for hate speech generation.

Domain fine-tuning

We wish to approximate the language distribution of a specific class, namely hate speech or non-hate, as opposed to generating random text sequences. We therefore fine-tune the pre-trained GPT-2 model, where we continue training the model from its distribution checkpoint, serving it with relevant text sequences. We randomly selected 80% of the labeled examples per dataset and category for this purpose (see Table 1). (The remaining labeled examples are held out for evaluating hate detection performance in our experiments.) Let us denote the examples sampled from dataset $d_i$ that are labeled as hate and non-hate by $d^i_{hate}$ and $d^i_{non-hate}$, respectively. Accordingly, we fine-tune two distinct GPT-2 models per dataset $d_i$:

- $G^i_{hate}$—fine-tuned using $d^i_{hate}$, and
- $G^i_{non-hate}$—fine-tuned using $d^i_{non-hate}$.

Importantly, fine-tuning a separate GPT-2 model per dataset and class (yielding 10 models overall) is expected to bias the language generation process according to the topics and terms that are characteristic to that combination of dataset and class. Once text sequences are generated from each fine-tuned model, they are unified so as to obtain a diverse distribution of automatically generated text samples.

Candidate sequence Generation

Given the fine-tuned GPT-2 models, we generate a large number of sequences (600K) per model. Similarly to the labeled examples, we generate text sequences that are relatively short, limiting the sequence length to 30 tokens. We provide no prompt to the GPT-2 model, that is, text is generated unconditionally, starting from an empty string.

Sequence selection

One should not expect all of the sequences automatically generated by the fine-tuned GPT-2 models to be perceived as hateful (or non-hateful). We found it crucial to assess and filter the generated sequences by their quality and relevance to the focus class. We employ BERT [24], a powerful DL transformer-based model, which has been shown to achieve state-of-the-art results on many natural language processing tasks, for automatically classifying the generated sequences as hateful (or not). BERT, which has been trained using large amounts of text, projects given text sequences into contextual embedding representations, which are useful for text classification; adding a softmax layer on top of BERT, one may learn to predict
target classes of interest by further training the model using labeled examples [24], [25].

We utilize the pool of labeled training examples for fine-tuning BERT on the task of hate speech detection, using a balanced set of hate and non-hate labeled examples (by down-sampling the latter) for this purpose. Similarly to other works, we found it beneficial to first fine-tune BERT on the related task of sentiment classification, using labeled examples from the Stanford Sentiment Treebank (SST) [26], [27]. We used the popular HuggingFace platform [28] for fine-tuning BERT, and consequently—predicting the class probabilities using the fine-tuned BERT for each generated sequence. We maintained the top-scoring 100K sequences per dataset and class—their union comprises our 1M MegaSpeech sequence corpus.

**MegaSpeech: A Large Hate Speech resource**

We created *MegaSpeech*, a large-scale resource comprised of 1M text sequences, automatically generated and assessed as hate and non-hate speech. This resource exceeds the size of currently available datasets (each comprised of thousands of examples) by magnitudes of order. We will make this resource available for research purposes upon request.

A sample of 1,000 randomly selected sequences was independently evaluated by the authors, manually assigning the generated texts to the classes of hate and non-hate. We measured inter-annotator agreement based on a subset 250 co-annotated random examples and obtained a score of 0.712 in terms of Fleiss kappa [29], indicating on strong agreement between the annotators. Overall, the ratio of generated hate sequences perceived as hate speech was as high as 65%, and similarly, the ratio of generated non-hate sequences perceived as non-hate was estimated at 86%. We further assessed the text sample with respect to coherence, or ‘readability’ using a scale of 1–5 (ranging from poor to excellent text quality, respectively), where this evaluation yielded an average score of 3.62, i.e., overall the text quality was good. We note that the generated text resembles the seed examples genre-wise, making use of slang, abbreviations, and hashtags. Table 5 includes several representative examples of hate and non-hate generated sequences.

| Source dataset | Generated hate | Generated non-hate |
|----------------|----------------|--------------------|
| WS             | 0.12           | 0.05               |
| DV             | 0.07           | 0.05               |
| FN             | 0.11           | 0.14               |
| WH             | 0.09           | 0.16               |
| SE             | 0.05           | 0.03               |

We further wished to verify that the generated sequences depart from the seed labeled examples, forming new and diverse language usage, as opposed to duplicating or repeating parts of the input examples. We therefore assessed the lexical similarity between the automatically generated sequences and the source examples per dataset and class [30]. Concretely, we employed ROUGE-L [31], a popular text similarity measure which indicates how much is common between two word sequences—ranging from 0 (nothing in common) to 1 (identical sequences). Table 2 reports the computed average ROUGE-L similarity scores. As shown, the similarity scores are low, indicating that the rich language model encoded within GPT-2 was effectively leveraged for generating new and different text sequences.

**Can MegaSpeech boost hate speech detection?**

In the following section, we report the results of an empirical study, where we train deep learning hate detection models using MegaSpeech. Let us denote those hate and non-hate text sequences in MegaSpeech that were generated given the training examples of dataset $d_i$ by $gen_{hate}^i$ and $gen_{non-hate}^i$, respectively. In our experiments, we augment the training examples of $d_i$ with those automatically generated examples, ideally providing hate detection classifiers that are trained using this dataset with a more general and truthful approximation of the underlying language distribution.

Our results show consistent improvements in hate detection performance by using the MegaSpeech text sequences. In particular, we show a major boost in classification recall
(aka, *sensitivity*), which is crucial for effective hate speech detection—this impact is due to the greater diversity of hate speech introduced by MegaSpeech compared with the existing datasets.

**Generalized Hate Detection: Evaluation**

This section presents a comparative study of the generalization capabilities of DL-based hate speech detection, where we augment hate speech datasets with massive amounts of realistic text sequences, sampled from the approximated distribution. Concretely, each one of the datasets in Table 1 was randomly split to a training set (80%) and a testing set (20%). The data generation processes and all the training procedures described below were performed using the training sets, where hate detection performance is evaluated exclusively using the corresponding testing sets. We learn hate detection classification models using two types of training sets: the *baseline*, which includes several thousands of human-generated text samples per dataset (Table 1) and *augmented*, constructed by adding $gen_{i\text{hate}}$, 100,000 model-generated hate speech samples and $gen_{i\text{non-hate}}$, 100,000 model-generated non-hate speech samples, to the corresponding baseline training set.

Hate speech detection results were obtained using the high-performing DL-based detector [10], as detailed in Figure 1. The DL detector was implemented in TensorFlow [32] and trained using an NVIDIA K-80 GPU, with the Adam [8] optimizer, minimizing the binary cross-entropy loss, with early stopping and mini-batch size of 32 samples. The Convolutional layer consists of 100 filters with a window size of 4, the MaxPooling1D layer uses a pool size of 4 and the GRU outputs a 100 hidden units for each timestep. Note that in this study we compare hate detection performance given different datasets using a single DL detector, where it is not claimed that the chosen detector achieves state-of-the-art results for each of the datasets.

Performance was evaluated using standard classification metrics: accuracy (A), precision (P), recall (R) and F1 (the harmonic mean of precision and recall). We note that the utilized datsets are generally imbalanced—and so are the derived test sets, including smaller proportions of examples labeled as hate speech vs. non-hate speech. We report evaluation with respect to the hate speech class: Recall corresponds to the proportion of true hate speech examples that were automatically identified as hate speech, and Precision is the proportion of correct predictions within the examples identified as hate speech by the detector. While false positive predictions may be tracked relatively easily by means of further human inspection, it is impossible to track false negatives at scale. Therefore, increasing recall by means of improved generalization is of great importance in practice.

**Experiments**

Our study included two types of experiments, which we refer to as intra-dataset and cross-dataset hate speech detection. The *intra-dataset* experiments applied the DL detector to each dataset independently: performance of the DL detector was measured using the held-out test examples of every dataset $D_i$, having trained the detector with either the baseline training set, or with the augmented training set available for $D_i$. In addition, we formed a *combined* dataset comprised of the union of all datasets (having joined all training and testing sets, respectively). The *cross-dataset* experiments aim to evaluate the realistic condition of data shift, where the hate detector is applied to new examples drawn from a data distribution that is different from the one used for training. In these experiments, having the detector trained using the training set of dataset $D_i$, it was evaluated on the held-out testing set of dataset $D_j \not= D_i$. Again, we trained the DL detector using either the baseline or the augmented training set available for dataset $D_i$, aiming to gauge the impact of dataset augmentation on hate detection generalization; ideally, hate detection should maintain good performance when applied to new examples, where, as discussed above, it is crucial to maintain high recall during test time.

There generally exists a trade-off between precision and recall classification performances [33]. Given an input text sequence $x$, the output of the DL detector is the posterior probability $P(\text{class}(x) = "\text{hate}" | x)$, namely, the probability that the input belongs to the "hate" class. The input is classified as "hate" if $P(\text{class}(x) = "\text{hate}" | x) > \tau$, and otherwise as "non-hate", for some pre-defined threshold $\tau$. Different threshold choices alter the trade-off between precision and
Figure 1. The DL hate speech detector [10]: an input text sequence is represented by a bag-of-words model, where each word is a one-hot vector. The Embedding layer reduces the dimension of each word vector from the input vocabulary dimension to 300, and the sequence of embedded word vectors are processed by a 1D convolutional layer which extracts features. The 1D MaxPooling layer down-samples the convolutional features, and the resulting down-sampled sequence enters a Gated-Recurrent Unit (GRU) which learns the time-dependencies between words. The GRU output is down-sampled by a 1D GlobalMaxPooling layer and converted to a posterior probability (that the input sequence is hate speech) by the output layer.

Figure 2 provides the complete Precision-Recall curves for all datasets, computed by altering the detector threshold in the range of $\tau = 0.5$ to 0.95 in steps of 0.05. These curves clearly demonstrates that training data augmentation yields significantly improved Recall (and consequently F1) across this range.

Cross-Dataset Experiments. We further evaluated the generalization of the DL detectors in conditions of data shift. Figure 3 shows the results of applying the DL detectors trained using the baseline (non-augmented) training set $D_i$ on the held-out test set $D_i$ (intra-dataset), as well as on the held-out test sets of other datasets, $D_j \neq D_i$ (cross-dataset). As shown, there is a significant drop in performance measurements between the first and latter setups—we observe precision drops, and poor recall performance (below 0.2) on the target test sets in conditions of data shift. Therefore, provided with (un-augmented) train sets of few thousands of hate speech examples, the achievable generalization to other hate speech datasets is very limited, resulting in low levels of recall (i.e., high ratio of missed hate speech).

In another set of experiments, we re-trained the detector with each one of the correspond-
Table 3. Intra-Dataset classification results, comparing the baseline and augmented-baseline training sets

| Dataset | Accuracy | Precision | Recall | F1     |
|---------|----------|-----------|--------|--------|
|         | Baseline | Augmented (%) | Baseline | Augmented (%) | Baseline | Augmented (%) | Baseline | Augmented (%) |
| WS      | 0.967    | +1.03     | 0.902  | +2.11  |
| WH      | 0.891    | -2.13     | 0.672  | -14.87 |
| SE      | 0.715    | +6.85     | 0.263  | +73.02 |
| DV      | 0.922    | +1.41     | 0.802  | +8.10  |
| FN      | 0.956    | -1.46     | 0.795  | -19.77 |
| Combined| 0.904    | +0.11     | 0.935  | +48.26 |

Figure 2. Precision-Recall curves of the DL Hate Speech detector, trained separately using each dataset (a)-(e) and using the combined dataset (f), comparing baseline and augmented training data results.

ing augmented training set, and evaluated on the same cross-dataset combinations of the first experiment. Table 4 summarizes the results of the Cross-dataset experiment. The average metrics of all cross-dataset pairs reveal a consistent increase in accuracy (+6.9%), precision (+3.5%), recall (+182.8%) and F1 (+179.7%). That is, augmenting the original labeled training sets with the automatically generated text sequences resulted in a dramatic increase of recall, while maintaining good levels of precision. Overall, this trend is reflected by a steep increase in the combined measure of F1. These results thus provide another approval of the benefits of augmenting the data with generated sentences and using a broader data distribution by utilizing a pre-trained language model.
Figure 3. Cross-Dataset performance without training set augmentation: in each experiment the DL detector was trained using the training set of one dataset (the left initials of each pair). The trained detector was evaluated on the held-out test set of the same dataset (Intra-Dataset) and on the held-out test set of a different dataset (Cross-Dataset, according to the right initials of each pair). The results clearly indicate a significant decrease in cross-data performance, especially in the Recall and F1 metrics. Therefore, the proportion of false-negatives (i.e. undetected hate speech) significantly increases, as compared to the intra-dataset results.

Table 4. Cross-Dataset performance, comparing detector training using baseline to augmented-baseline training sets

| Trainset-Testset | Accuracy | Precision | Recall | F1  |
|------------------|----------|-----------|--------|-----|
| FN-SE            | 0.613    | 0.609     | 0.155  | 0.323|
| FN-WH            | 0.846    | 0.853     | 0.086  | 0.114|
| DV-SE            | 0.794    | 0.820     | 0.189  | 0.225|
| DV-WH            | 0.652    | 0.670     | 0.107  | 0.174|
| SE-FN            | 0.926    | 0.924     | 0.031  | 0.059|
| WH-SE            | 0.584    | 0.593     | 0.050  | 0.124|
| WH-DV            | 0.752    | 0.777     | 0.153  | 0.217|
| WS-WH            | 0.749    | 0.824     | 0.255  | 0.357|
| WS-SE            | 0.822    | 0.895     | 0.135  | 0.177|
| Average          | 0.740    | 0.791     | 0.151  | 0.168|

CONCLUSIONS

In this work we demonstrated the problems that may occur when training a hate speech classifier on small datasets. We showed the performance decrease that arises when training a hate speech classifier on a small dataset and tested on a dataset of a different hate speech language distribution. We suggested a new framework for augmenting hate speech dataset by using data that was generated from a deep generative model trained on the small hate speech datasets that were available. The proposed framework was evaluated using a high-performing DL hate speech detector, demonstrating a significant boost in the generalization capability of the DL detector, intra-dataset and cross-dataset as well.

ACKNOWLEDGMENT

This research was supported partly by Facebook Content Policy Research on Social Media Platforms Research Award.

REFERENCES

1. S. Kemp, “Digital 2020: Global digital overview,” DataReportal, January 2020. [Online]. Available: https://datareportal.com/reports/digital-2020-global-digital-overview
2. Twitter, “Twitter hateful conduct policy,” [Online]. Available: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
3. ——, “Updating our rules against hateful conduct,” March 2020. [Online]. Available: https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html
Table 5. Examples of generated sequences by the GPT-2 language model, fine-tuned by each source training set.

| Dataset | Hate Speech Generated by GPT-2                                                                 | Non-Hate Speech Generated by GPT-2                                                                 |
|---------|-----------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------|---|
| WS      | What does Islam have to do with democracy and human rights islam is a hateful religion that has inspired the murders of for starters i know some of you never read this canadian quran is full of excuses why so many muslims are murdering islamists the writer of the sharia law book given to the muslim world in the 1980s is calling for a religious civil war | The haters are all over this comment This story originally appeared on National Geographic News US wants Zambias political crisis to end |
| DV      | #AltRight is so important to white people these days they want to rebrand white genocide as immigration amnesty #WhiteGenocide #NoMoreD Fox News allowed these other nut jobs on their network to spout hatred towards women and minorities so that they could sell more toys to these scum pigs we must stop lawless illegal black hearted sand nigger lovers theyre ruining our country my countrymen dont like those people | A member of the Liberty Bell Choir sings the National Anthem before the game between the New York Yankees Who do you think of when you think of one of sports lowest paid employees in league history #yankees #mariners #mlb Afghanistanans military training looks to be on the brink of a revolution after President Ashraf Ghani pledged to make it so special that it is |
| FN      | this new administration is willing to start a world war over corruption while these illegal immigrants destroying our country are not taken into consideration i wish someone would tell me what the hell is going on with this video i hate it man that fucking nigga Democrats and their media sycophants do not want to talk about why illegal border crossers are stealing from citizens who pay taxes #buildthatwall | On Thursday evening a collection of posts appeared on a website with the slogan stop the racist terrorist threats to burma #facebook #burmaco what the fuck does feeling good imply hes not feeling good anymore ill shake his hand or give him a hug mfw It was initially deemed a case of falling victim to men in power who are trying to change the #innovation #technology |
| WH      | The number of refugees and immigration controls are clear, now is it time to go to extremes in a desperate attempt to stop them the left celebrates the death of freedom as it is the road to the extermination of whites and the rebirth of the baboon Mexicans wonder why their country is going to hell when they send their sons and daughters to fight for such disgusting filth | thanks for your reponse this is what our school system is becoming as an asian guy living in canada i have to say i love it here welcome to stormfront i remember when my 3 year old son made a face like the one on the far left in the above cartoon |
| SE      | Afghan refugees now run our towns and streets #buildthatwall #secureourborders #unleashtheinsane illegal aliens must be deported or even sued if not they will keep coming illegally #nodaca #noamnesty #endbirthrightcitizenship | Police say a man in his 20s is dead and three others were injured after a brawl broke out at a political rally in Athens To celebrate the launch of the Fund for Refugees I am inviting you to become a guardian of this important right for every Syrian refugee as we defend refugee rights Republican presidential candidate Donald Trump has vowed to ensure anything is possible with all of our great employees #muslimimmigration partners #bayareamus |

2020 11
4. Facebook, “Community standards: Hate speech.” [Online]. Available: https://www.facebook.com/communitystandards/hate_speech

5. J. Peters, “Twitter now bans dehumanizing remarks based on age, disability, and disease,” The Verge, March 2020. [Online]. Available: https://www.theverge.com/2020/3/5/21166940/twitter-hate-speech-ban-age-disability-disease-dehumanize

6. M. Zuckerberg, “Hate speech has no place on facebook,” The Guardian. [Online]. Available: https://www.theguardian.com/technology/2016/feb/26/mark-zuckerberg-hate-speech-germany-facebook-refugee-crisis

7. I. Lunden, “Facebook to add 3,000 to team reviewing posts with hate speech, crimes, and other harming posts,” TechCrunch, May 2017. [Online]. Available: https://techcrunch.com/2017/05/03/facebook-to-hire-3000-to-review-posts-with-hate-speech-crimes-and-other-harming-posts/

8. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

9. T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent trends in deep learning based natural language processing [review article],” IEEE Computational Intelligence Magazine, vol. 13, no. 3, pp. 55–75, 2018.

10. Z. Zhang, D. Robinson, and J. Tepper, “Detecting hate speech on twitter using a Convolution-GRU based deep neural network,” in European Semantic Web Conference. Springer, 2018, pp. 745–760. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-93417-4_48

11. L. Zhang, S. Wang, and B. Liu, “Deep learning for sentiment analysis: A survey,” WIREs Data Mining and Knowledge Discovery, vol. 8, no. 4, p. e1253, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1253

12. K. Kawaguchi, L. P. Kaelbling, and Y. Bengio, “Generalization in deep learning,” ArXiv, 2017. [Online]. Available: https://dspace.mit.edu/bitstream/handle/1721.1/115274/gedl.pdf?sequence=2&isAllowed=y

13. A. Radford, J. Wu, R. Child, D. Luan, J. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” OpenAI Blog, vol. 1, no. 8, p. 9, 2019.

14. T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” in Eleventh international AAAI conference on web and social media (ICWSM), 2017.

15. P. Burnap and M. L. Williams, “Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making,” Policy & Internet, vol. 7, no. 2, pp. 223–242, 2015.

16. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Advances in neural information processing systems, 2013, pp. 3111–3119.

17. A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontidis, “A unified deep learning architecture for abuse detection,” in Proceedings of the 10th ACM Conference on Web Science, 2019.

18. J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

19. P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,” in Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp. 759–760.

20. Z. Waseem and D. Hovy, “Hateful symbols or hateful people? predictive features for hate speech detection on twitter,” in Proceedings of the NAACL student research workshop, 2016, pp. 88–93.

21. V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti, “Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter,” in Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 54–63.

22. A. M. Founta, C. Dijouvas, D. Chatzakou, I. Leontidis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, “Large scale crowdsourcing and characterization of twitter abusive behavior,” in International AAAI Conference on Web and Social Media, 2018.

23. O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros, “Hate speech dataset from a white supremacy forum,” in Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 11–20. [Online]. Available: https://www.aclweb.org/anthology/W18-5102

24. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in NAACL-HLT, 2019.

25. C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to fine-tune BERT for text classification?” in China National Conference on Chinese Computational Linguistics. Springer, 2019, pp. 194–206.

26. R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, “Recursive deep models for
semantic compositionality over a sentiment treebank,” in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.

27. A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, “Glue: A multi-task benchmark and analysis platform for natural language understanding,” in *EMNLP*, 2018.

28. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, “Huggingface’s transformers: State-of-the-art natural language processing,” *ArXiv*, vol. abs/1910.03771, 2019.

29. J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

30. S. Witteveen and M. Andrews, “Paraphrasing with large language models,” in *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Association for Computational Linguistics, 2019.

31. C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://www.aclweb.org/anthology/W04-1013

32. [Online]. Available: https://www.tensorflow.org/

33. E. Minkov, R. C. Wang, A. Tomasic, and W. Cohen, “Ner systems that suit users preferences: adjusting the recall-precision trade-off for entity extraction,” in *Proceedings of the Human Language Technology Conference of the NAACL Companion Volume: Short Papers*, 2006, pp. 93–96.