Large-Scale Hate Speech Detection with Cross-Domain Transfer

Anonymous ACL submission

Abstract

Hate speech towards people with different backgrounds is a major problem observed in social media. Although there are various attempts to detect hate speech automatically via supervised learning models, the performance of such models simply rely on limited datasets on which models are trained. In this study, we construct large-scale tweet datasets for supervised hate speech detection in English and Turkish, including human-labeled 100k tweets per each. Our datasets are designed to have equal number of tweets distributed over five domains; namely religion, gender, race, politics, and sports. We analyze the performance of state-of-the-art language models on large-scale hate speech detection with a special focus on model scalability. We also examine cross-domain transfer ability of hate speech detection.

1 Introduction

With the growth of social media platforms, hate speech towards people who do not share the same identity or community increases dramatically (Twitter, 2021). Consequences of online hate speech could be real-life violence against other people and communities (Byman, 2021). The need of automatically detecting hate speech text is thereby urging.

Existing solutions to detect hate speech mostly rely on supervised learning, resulting in a strict dependency on the quality and quantity of labeled data. Most of the datasets labeled by human experts for hate speech detection are not large in size due to the labor cost (Poletto et al., 2021), causing a lack of detailed experiments on model generalization and scalability. Indeed, most studies on hate speech detection report high performances on their test sets, while their generalization capabilities to other datasets are limited (Arango et al., 2019).

Existing datasets for hate speech detection are mostly prepared for non-agglutinative languages, e.g. around half of them are in English (Poletto et al., 2021). Agglutinative ones, such as Turkic and Uralic languages, have low or no resources for hate speech detection. We thereby construct large-scale human-annotated datasets for hate speech detection using English and Turkish tweets.

Hatred language can be expressed in various topics (we refer to topics as *hatred domains*). Domains vary depending on the target group. For instance, misogyny (targeting women) and homophobia (targeting different gender identities) are examples of the domain of gender-based hatred. Existing studies mostly consider a limited number of domains, and investigate hate speech in terms of an abstract notion including aggressive language, threats, slurs, and offenses (Poletto et al., 2021). We consider not only the hatred behavior in the definition of hate speech, but also five most frequently observed domains depending on target group; namely religion, gender, racism, politics, and sports-based hatred.

Supervised models trained on a specific learning dataset can fail to generalize their performance on the original evaluation set to other evaluation sets. However, this phenomenon is studied in cross-dataset[1] (Gröndahl et al., 2018; Karan and Šnajder, 2018), cross-lingual (Pamungkas and Patti, 2019), and cross-platform (Agrawal and Awekar, 2018) transfer. Transfer learning among hatred domains is not well studied due to the lack of large-scale datasets. In this study, with the help of our novel datasets including five hatred domains mentioned above, we analyze the generalization capability of hate speech detection in terms of hatred domains.

The **contributions** of this study are in three folds. (i) We construct large-scale human-labeled hate speech detection datasets for English and Turkish. (ii) We analyze the performance of various models for hate speech detection with a special

[1]In literature, the phrase "cross-domain" is mostly used for the transfer between two datasets that are published by different studies but not necessarily in different hatred domains. We refer to them as cross-dataset.

focus on model scalability. (iii) We examine the generalization capability of hate speech detection in terms of zero-shot cross-domain transfer.

The structure of the paper is as follows. In the next section, we provide a summary of related work. In Section 3, we explain our large-scale datasets. In Section 4, we report our experimental design and results. In Section 5, we provide a discussion on scalability, ablation study, and limitations of our study. We conclude the study in the last section.

2 Related Work

We briefly summarize related work on the methods, previous datasets, and transfer learning for hate speech detection.

2.1 Methods for Hate Speech Detection

Earlier studies on hate speech detection are based on matching hatred keywords using lexicons (Sood et al., 2012). The disadvantage of such methods is strict dependency on lexicons. Supervised learning with a set of features extracted from a training set is a solution for the dependency issue. Text content is useful to extract bag-of-words features; such as n-grams, Part-of-Speech tags, linguistic and syntactical features (Dadvar et al., 2013; Waseem and Hovy, 2016; Nobata et al., 2016; Waseem, 2016; Davidson et al., 2017). User-based features, such as content history, meta-attributes, and user profile (Dadvar et al., 2013; Waseem, 2016; Chatzakou et al., 2017; Unsvåg and Gambäck, 2018), can be used to detect hatred signals. Structural features of a social network, such as centrality and clustering, are studied as well (Chatzakou et al., 2017).

To capture word semantics better than bag-of-words; word embeddings, such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), are utilized to detect abusive and hatred language (Djuric et al., 2015; Nobata et al., 2016; Mou et al., 2020). To resolve the issues related to noisy text of social media, character and phonetic-level embeddings are studied for hate speech (Mou et al., 2020). Instead of extracting hand-crafted features; deep neural networks, such as CNN (Kim, 2014) and LSTM (Jozefowicz et al., 2015), are applied to extract deep features to represent text. Indeed, their application outperforms previous methods that employ lexicons and hand-crafted features (Badjatiya et al., 2017; Zimmerman et al., 2018; Mou et al., 2020; Cao et al., 2020).

Recently, Transformer architecture (Vaswani et al., 2017) is studied for hate speech detection, as in all other downstream tasks of NLP. Transformer employs self-attention for each token over all tokens, targeting to capture a rich contextual representation of whole text. Fine-tuning BERT (Devlin et al., 2019) for hate speech detection outperforms previous methods (Liu et al., 2019a; Caselli et al., 2021; Mathew et al., 2021; Aluru et al., 2021). We examine the performance of not only BERT, but also various Transformer language models for both multi-class and binary hate speech detection.

2.2 Resources for Hate Speech Detection

A recent survey summarizes the current state of datasets in hate speech detection by listing over 40 datasets, around half of which are tweets, and again around half of which are prepared in English language (Poletto et al., 2021). Benchmark datasets are also released as a shared task for hate speech detection (Basile et al., 2019; Zampieri et al., 2020).

There are efforts to create large-scale human-labeled datasets for hate speech detection. The dataset in Davidson et al. (2017) has around 25k tweets each labeled by three or more annotators for three classes; offensive, hate, and neither. The dataset in Golbeck et al. (2017) has 35k tweets labeled by at most three annotators per tweet for binary classification (harassing or not). The dataset in Founta et al. (2018) has 80k tweets each labeled by five annotators for seven classes including offensive and hate. However, our datasets differ in terms of the following aspects. We have 100k top-level tweets per two languages, English and Turkish. The datasets are clean, which will be explained in the next section. We have three class labels (hate, offensive, and normal), and five annotators per each tweet. Lastly, we design to have 20k tweets for each of five hatred domains, which would enable us to analyze zero-shot cross-domain transfer.

2.3 Transfer Learning for Hate Speech Detection

Generalization of a hate-speech detection model trained on a specific dataset to other datasets with the same or similar class labels, i.e. cross-dataset transfer, is widely studied (Gröndahl et al., 2018; Karan and Šnajder, 2018; Wiegand et al., 2018; Pamungkas and Patti, 2019; Swamy et al., 2019; Arango et al., 2019; Pamungkas et al., 2020; Markov and Daelemans, 2021). Using different datasets in different languages, cross-lingual transfer aims to overcome language dependency in hate speech detection (Pamungkas and Patti, 2019; Pamungkas et al., 2020; Markov et al., 2021; Nozza, 2021). There are also efforts to analyze platform-independent hate speech detection, i.e. cross-platform transfer (Agrawal and Awekar, 2018). In this study, we analyze whether hate speech detection can be generalized across hatred domains, regardless of the target and topic of hate speech.

3 Large-Scale Datasets for Hate Speech Detection

3.1 Dataset Construction

We used Full-Archieve Search provided by Twitter Premium API to retrieve more than 200k tweets; filtered according to language, tweet type, publish time, and contents. We filter English and Turkish tweets published in 2020 and 2021. The dataset contains only top-level tweets, i.e., not a retweet, reply, or quote. Tweet contents are filtered based on a keyword list. The list contains hashtags and keywords from five topics (i.e. hatred domains); religion, gender, racism, politics, and sports. We design to keep the number of tweets belonging to each hatred domain balanced.

For cleaning, we remove near-duplicate tweets by measuring higher than 80% text similarity between tweets using the Cosine similarity with TF-IDF term weighting (Sedhai and Sun, 2015). We restrict the average number of tweets per user in order not to exceed 1% of all tweets to avoid user-dependent modeling (Geva et al., 2019). We also remove tweets shorter than five words; excluding hashtags, URLs, and emoticons.

3.2 Dataset Annotation

Based on the definitions and categorization of hateful speech (Sharma et al., 2018), we label tweets as containing hate speech if they target, incite violence against, threaten, or call for physical damage for an individual or a group of people because of some identifying trait or characteristic. We label tweets as offensive if they humiliate, taunt, discriminate, or insult an individual or a group of people in any form, including visual and textual. Other tweets are labeled as normal.

Each tweet is annotated by five annotators randomly selected from a set of 16 undergrads and four grads. If consensus is not achieved on ground-truth, a human expert outside the initial annotator set determines the label. We provide annotation guidelines to all annotators. The guidelines document includes the rules of annotations; the definitions of hate, offensive, and normal tweets; and the common mistakes observed during annotation. The annotations started on February 15th, and ended on October 5th, 2021 (i.e. a period of 84 days). We measure inter-annotator agreement with Krippendorff's alpha coefficient and get a nominal score of 0.395 for English and 0.417 for Turkish.

Definition	**EN**	**TR**
Number of tweets	100,000	100,000
Number of offensive tweets	27,140	30,747
Number of hate tweets	7,325	27,593
Number of users	85,396	69,524
First tweet date	02/26/20	01/17/20
Last tweet date	03/31/21	03/31/21
Average tweets per user	1.2	1.4
Average tweet length (words)	29.20	24.37
Shortest tweet length	5	5
Longest tweet length	72	121
Number of hashtags	23,170	24,444
Number of URLs	76,006	72,233
Number of tweets with hashtags	12,751	17,390
Number of tweets with URLs	73,439	71,434

Table 1: Dataset statistics. We construct two large-scale datasets including English (EN) and Turkish (TR) tweets for hate speech detection in terms of three classes (hate, offensive, and normal).

Lang.	**Domain**	**Hate**	**Offens.**	**Normal**	**Total**
EN	Religion	1,427	5,221	13,352	20k
	Gender	1,313	6,431	12,256	20k
	Race	1,541	3,846	14,613	20k
	Politics	1,610	6,018	12,372	20k
	Sports	1,434	5,624	12,942	20k
TR	Religion	5,688	7,435	6,877	20k
	Gender	2,780	6,521	10,699	20k
	Race	5,095	4,905	10,000	20k
	Politics	7,657	4,253	8,090	20k
	Sports	6,373	7,633	5,994	20k

Table 2: Distribution of topics in our datasets with respect to three classes (hate, offensive, and normal).

3.3 Dataset Statistics

We report main statistics about our datasets in Table 1. Although we follow a similar construction approach, the number of tweets with hate speech in English is less than those in Turkish, which might indicate a tighter regularization for English content by Twitter. Normal tweets dominate in both languages, specifically in English, as expected due to the nature of hate speech and the platform regulations. The statistics of tweet length imply that our task is similar to a short text classification for tweets, where the average number of words is ideal to be 25 to 30 (Şahinuç and Toraman, 2021).

The distribution of tweets for each domain and language is given in Table 2. In English, the number of hatred tweets are similar in each domain; however, race has less number of offensive tweets than others. The number of hatred tweets are similar in Turkish, except gender and politics.

4 Experiments

We have two main experiments. First, we analyze the performance of various methods for hate speech detection. In the second part, we examine the generalization capability of hate speech detection in terms of cross-domain transfer.

4.1 Hate Speech Detection

4.1.1 Experimental Design

We apply 10-fold leave-one-out cross-validation, where each fold has 90k train instances; and report the average score of accuracy, precision, recall, and weighted F1 score. We fine-tune the following models that are pre-trained by using English text:

- **ALBERT** (Lan et al., 2020): Compared to BERT (Devlin et al., 2019), ALBERT has additional training data and lowers memory consumption with fewer parameters. Instead of next sentence prediction, sentence order prediction is used to focus on coherence between two sentences.
- **BART** (Lewis et al., 2020): BART is a seq2seq model that employs a bidirectional encoder and a left-to-right decoder. The advantage is to learn a model by reconstructing the input text. BART has sentences randomly shuffled in training, and text spans are masked instead of single words.
- **BERT** (Devlin et al., 2019): BERT uses bidirectional language modeling with masked language modeling and next sentence prediction.
- **BERTweet** (Nguyen et al., 2020): BERTweet is trained based on the RoBERTa (Liu et al., 2019b) pre-training procedure by using only tweets.
- **ConvBERT** (Jiang et al., 2020): ConvBERT architecture replaces the quadratic time complexity of the self-attention mechanism of BERT with convolutional layers. The number of self-attention heads are reduced by a mixed attention mechanism of self-attention and convolutions that would model local dependencies.
- **DeBERTa** (He et al., 2021): DeBERTa introduces a disentangled attention mechanism on top of the BERT architecture to emphasize relative word positions. The model also uses an enhanced mask decoder for absolute positions. DeBERTa employs BPE instead of WordPiece tokenization.
- **DistilBERT** (Sanh et al., 2019): DistilBERT is an efficient version of BERT with 40% less parameters while retaining 97% of its performance.
- **ELECTRA** (Clark et al., 2020): ELECTRA introduces the discriminator, a Transformer model that replaces the task of masked language modeling with replaced token detection. This new task predicts if a token is replaced by a generator network, enabling to run the task for all tokens rather than a subset as in masked modeling.
- **Megatron** (Shoeybi et al., 2019): Megatron introduces an efficient parallel training approach for BERT-like models to increase parameter size.
- **RoBERTa** (Liu et al., 2019b): RoBERTa is built on BERT architecture with modified hyperparameters and a diverse corpora in pretraining, and removes the task of next sentence prediction.
- **XLNet** (Yang et al., 2019): XLNet replaces the task of masked language modeling with permutation language modeling, and removes the task of next sentence prediction.

There are already fine-tuned models for hate speech detection in English (we find no fine-tuned model for Turkish hate speech detection). We use the following fine-tuned models for zero-shot inference, as well as fine-tuning again with our data.

- **HateXplain** (Mathew et al., 2021): HateXplain fine-tunes BERT-base, using a novel dataset with 20k instances, 9k of which are tweets. The model can be used for zero-shot inference on multi-class (hate, offensive, and normal) detection.
- **HateBERT** (Caselli et al., 2021): HateBERT re-trains BERT-base, using around 1.5m Reddit messages published by suspended communities due to promoting hateful content. The model can be used for zero-shot inference on binary classification (hateful or not).

For Turkish, we fine-tune the same models used in English listed above, except already fine-tuned ones, to understand cross-lingual generalization capability from English and Turkish. Besides, we fine-tune the following models that are pre-trained by using only Turkish text.

- **BERTurk** (Schweter, 2020): The model re-trains BERT architecture for Turkish data.
- **DistilBERTurk** (Schweter, 2020): A distilled version of BERTurk with a smaller training data.
- **ConvBERTurk** (Schweter, 2020): Based on ConvBERT (Jiang et al., 2020), but using a modified training procedure and Turkish data.

- **ELECTRA (TR)** (Schweter, 2020): Based on ELECTRA (Clark et al., 2020), but using Turkish data. We refer to it as ELECTRATurk.

To understand generalization capability of from multi-lingual models to both English and Turkish, we fine-tune the following multi-lingual models.

- **mBERT** (Devlin et al., 2019): mBERT is built on BERT architecture, but using multilingual data covering 100 languages.
- **XLM-R** (Conneau et al., 2020): XLM-R is built on RoBERTa architecture, but using multilingual data covering 100 languages. The model is trained on more data than mBERT, and removes the task of next sentence prediction.

Our dataset is prepared for fine-tuning multi-class (hate, offensive, and normal) detection. However, to understand the performance of models in binary setup, we merge offensive and hate instances into a single hate class. We report performances in both multi-class and binary setups for all models listed above, if fine-tuning is available accordingly.

To get fair comparison, all models are set to the same hyper-parameters: Batch size is 32, learning rate is 1e-5, the number of epochs is 10, maximum input length is 128 tokens, using AdamW optimizer. Only exception is Megatron, due to its large size, we reduce batch size to 8 and epochs to 5. We use GeForce RTX 2080 Ti for fine-tuning the models.

4.1.2 Experimental Results

In Table 3, we report the performance of multiclass (hate, offensive, and normal) and binary (hate + offensive vs. normal) hate speech detection along with model sizes, pretraining domains, and the average time in minutes of 10-folds for fine-tuning. The highest performing models in English are those with the highest number of parameters (Megatron and BART) regardless of multi-class or binary setups. BERTweet achieves higher performance than BERT which would highlight the importance of the domain of the pretrain corpus.

The highest performing model in Turkish is ConvBERTurk both in multi-class and binary setups. Pretraining in the same language with the downstream task helps increase the performance. However, the performance difference between XLM-R and BERTurk models are not substantial. We thereby argue that one can utilize multilingual models in low-resource setups. The models pretrained in English demonstrate a capability of cross-lingual transfer, e.g. ELECTRA achieves competitive performance with multi-lingual and Turkish models, when fine-tuned for Turkish.

Zero-shot models fine-tuned for hate speech detection on other datasets underperform on our data, and do not achieve highest performances when fine-tuned further. This observation would show that already fine-tuned models have limited capability of generalization to new data.

The performance of binary detection is higher than multi-class detection in both languages, as expected. Binary detection dramatically improves the performance in Turkish, which would show the poor performance of detecting offensive tweets in Turkish (see class-based analysis in Section 5).

4.2 Cross-Domain Transfer

4.2.1 Experimental Design

We test cross-domain transferability with fine-tuning a model on a source domain and testing it on a target domain. We design to set a fixed hatred domain as target, and remaining ones as source. The performance can be measured by relative zero-shot transfer ability (Turc et al., 2021). We refer to it as *recovery ratio*, since it represents the ratio of how much original performance is recovered by changing source domain, given as follows.

$$recovery(S,T) = \frac{M(S,T)}{M(T,T)} \quad (1)$$

where $M(S,T)$ is a model performance for the source domain S on the target domain T. In the case of source and target domains are the same, recovery would be 1.0.

We also set a fixed hatred domain as source, and remaining ones as target. The performance can be measured by cross-lingual transfer gap (Hu et al., 2020). We modify it to normalize, and refer to it as *decay ratio*, since it represents the ratio of how much inference performance is decayed by replacing target domain, given as follows.

$$decay(S,T) = \frac{M(S,T) - M(S,S)}{M(S,S)} \quad (2)$$

In the case of source and target domains are the same, there would be no decay or performance drop, so decay would be zero. In the cross-domain experiments, we measure weighted F1; and employ BERT for English, and BERTurk for Turkish.

Lang.	Model	Params	Pretrain	Multi-class					Binary				
				Acc.	Prec.	Recall	F1	Time	Acc.	Prec.	Recall	F1	Time
	ALBERT	11.7m	W,B	0.806	0.680	0.806	0.731	138.3	0.853	0.736	0.853	0.789	139.1
	BART	139.4m	W,B	**0.819**	0.692	**0.819**	0.745	163.0	**0.866**	0.755	**0.866**	0.805	162.0
	BERT	108.3m	W,B	0.808	0.679	0.808	0.732	135.5	0.858	0.743	0.858	0.794	136.5
	BERTweet	134.9m	M	0.815	0.686	0.815	0.741	133.2	0.863	0.750	0.863	0.801	134.8
	ConvBERT	105.7m	Web	0.812	0.684	0.812	0.738	156.3	0.861	0.747	0.861	0.798	157.2
	DeBERTa	138.6m	W,B,Web,M,S	0.811	0.681	0.811	0.736	171.7	0.862	0.750	0.862	0.801	172.0
	DistilBERT	65.2m	W,B	0.807	0.679	0.807	0.732	67.2	0.856	0.739	0.856	0.792	67.7
	ELECTRA	108.9m	W,B	0.809	0.679	0.809	0.734	139.0	0.861	0.747	0.861	0.798	132.7
EN	Megatron	345m	W,S,N,Web	0.817	**0.703**	0.817	**0.749**	295.6	0.864	**0.765**	0.864	**0.807**	287.3
	RoBERTa	124.6m	W,B,N,Web,S	0.814	0.687	0.814	0.741	134.2	0.864	0.765	0.864	0.807	134.0
	XLNet	116.7m	W,B,N,Web,CC	0.810	0.681	0.810	0.735	179.7	0.859	0.745	0.859	0.797	178.5
	mBERT	177.9m	W	0.805	0.677	0.805	0.730	144.9	0.855	0.738	0.855	0.790	140.1
	XLM-R	278.0m	CC	0.816	0.689	0.816	0.742	145.3	0.863	0.752	0.863	0.802	146.0
	HateXplain	109.5m	W,B,M	0.681	0.637	0.681	0.647	zero-shot	-	-	-	-	-
	HateXplain	109.5m	W,B,M	0.782	0.643	0.782	0.700	133.7	-	-	-	-	-
	hateBERT	109.5m	M	-	-	-	-	-	0.654	0.652	0.654	0.653	zero-shot
	hateBERT	109.5m	M	-	-	-	-	-	0.859	0.745	0.859	0.796	132.3
	ALBERT	11.7m	W,B	0.691	0.499	0.691	0.575	135.6	0.806	0.659	0.806	0.723	145.6
	BART	139.4m	W,B	0.721	0.544	0.721	0.614	159.7	0.826	0.691	0.826	0.750	175.4
	BERT	108.3m	W,B	0.726	0.548	0.726	0.620	129.7	0.826	0.691	0.826	0.751	141.1
	BERTweet	134.9m	M	0.739	0.569	0.739	0.639	139.8	0.834	0.704	0.834	0.762	142.3
	ConvBERT	105.7m	Web	0.732	0.560	0.732	0.629	151.8	0.826	0.690	0.826	0.750	164.1
	DeBERTa	138.6m	W,B,Web,M,S	0.726	0.549	0.726	0.620	168.6	0.826	0.692	0.826	0.751	177.6
	DistilBERT	65.2m	W,B	0.722	0.543	0.722	0.614	66.7	0.825	0.689	0.825	0.748	72.9
	ELECTRA	108.9m	W,B	0.748	0.581	0.748	0.650	129.9	0.842	0.716	0.842	0.772	135.8
TR	Megatron	345m	W,S,N,Web	0.725	0.562	0.725	0.625	303.9	0.826	0.704	0.826	0.755	288.8
	RoBERTa	124.6m	W,B,N,Web,S	0.728	0.552	0.728	0.623	130.5	0.831	0.701	0.831	0.758	135.5
	XLNet	116.7m	W,B,N,Web,CC	0.730	0.556	0.730	0.626	187.4	0.828	0.695	0.828	0.754	177.9
	mBERT	177.9m	W	0.744	0.576	0.744	0.644	134.0	0.839	0.711	0.839	0.768	135.5
	XLM-R	278.0m	CC	0.761	0.600	0.761	0.667	143.7	0.856	0.739	0.856	0.791	142.4
	BERTurk	110.6m	W,B,Web	0.767	0.606	0.767	0.673	129.3	0.863	0.752	0.863	0.802	132.8
	DistilBERTurk	67.5m	W,B,Web	0.759	0.596	0.759	0.663	67.7	0.851	0.732	0.851	0.785	71.1
	ConvBERTurk	106.8m	W,B,Web	**0.770**	**0.610**	**0.770**	**0.677**	154.5	**0.867**	**0.758**	**0.867**	**0.807**	157.4
	ELECTRATurk	110.0m	W,B,Web	0.767	0.608	0.767	0.674	133.7	0.864	0.754	0.864	0.804	132.0

Table 3: **Multi-class and binary hate speech detection**. Average of 10-fold cross-validation is reported. Highest score is given in bold. Time is the average minutes of 10-fold fine-tuning. Models are divided into sub-groups in terms of English, multi-lingual, already fine-tuned, and Turkish language models. For pretraining datasets; W stands for Wikipedia, B for BooksCorpus (Zhu et al., 2015), M for Social Media (Twitter or Reddit), Web for OpenWebText (Gokaslan and Cohen, 2019) or ClueWeb (Callan et al., 2009), S for Stories (Trinh and Le, 2018), N for News (RealNews (Zellers et al., 2019) or Giga5 or CCNews), CC for CommonCrawl.

4.2.2 Experimental Results

Table 4 answers the question of "To what extent target domain is recovered by different source domains?" Recovery performances between domains are quite effective, such that all recovery performances are above 80% for both languages. The reason might be the similar hate speech patterns in the domains. Recovering gender domain is particularly more difficult than other domains in English. We argue that speech patterns in gender-based hatred text can be differentiated from general hate patterns, i.e. gender-based hatred is more unpredictable by other domains in English. We observe the same argument for politics in Turkish. We expect to fully recover when source is all domains, since the original source is already covered. Indeed, using all domains does not deteriorate recovery.

Table 5 shows the decay scores when tested on a different domain. When gender is used as source, there is no decay in other target domains in English, but not in Turkish. Recall that gender recovery in English is poor as well. We argue that gender-based hatred language is not easily transferred from other domains, but it can transfer hatred language to others. This could be important for data scarcity in hate speech detection. In addition, the performance of sports decays much when used as a source in both languages, showing that sports-based hatred cannot easily generalize to other domains.

We note that recovery and decay ratio can be interpreted together. For instance, in English, the domain transfer from religion to gender has 89% recovery, and its decay ratio is -12%. While the domain transfer from sports to gender has the same

Lang.	Source/Target	Religion	Gender	Racism	Politics	Sports	All
EN	Religion	0.712	89%	96%	97%	95%	96%
	Gender	101%	0.700	97%	99%	98%	99%
	Racism	99%	89%	0.750	94%	91%	94%
	Politics	97%	85%	94%	0.720	97%	95%
	Sports	95%	89%	91%	99%	0.782	95%
	All	101%	99%	100%	99%	99%	0.732
TR	Religion	0.637	91%	94%	90%	93%	93%
	Gender	90%	0.666	92%	84%	90%	90%
	Racism	94%	90%	0.676	88%	93%	93%
	Politics	85%	84%	88%	0.656	85%	88%
	Sports	88%	83%	88%	81%	0.705	88%
	All	101%	102%	100%	100%	101%	0.673

Table 4: Cross-domain transfer for hate speech detection in terms of **column-wise recovery ratio**. The results should be interpreted column-wise, e.g. 89% recovery from religion to gender in EN means that we recover 89% of 0.700 (gender to gender), but not 0.712 (religion to religion). Source domains are given in rows, targets in columns. Diagonal gray cells have weighted F1 where target and source is the same. As recovery increases, green color gets darker.

Lang.	Source/Target	Religion	Gender	Racism	Politics	Sports	All
EN	Religion	0.712	-12%	0%	-2%	0%	-1%
	Gender	0%	0.700	0%	0%	0%	0%
	Racism	-6%	-17%	0.750	-10%	-5%	-8%
	Politics	-4%	-17%	-2%	0.720	0%	-4%
	Sports	-14%	-20%	-13%	-9%	0.782	-11%
	All	-2%	-5%	0%	-2%	0%	0.732
TR	Religion	0.637	-5%	-0.3%	-8%	0%	-2%
	Gender	-14%	0.666	-7%	-18%	-5%	-9%
	Racism	-11%	-11%	0.676	-14%	-3%	-8%
	Politics	-18%	-14%	-9%	0.656	-9%	-10%
	Sports	-21%	-22%	-15%	-25%	0.705	-16%
	All	-5%	0%	0%	-2%	0%	0.673

Table 5: Cross-domain transfer for hate speech detection in terms of **row-wise decay ratio**. The results should be interpreted row-wise, e.g. -12% decay from religion to gender in EN means that we lose -12% of 0.712 (religion to religion), but not 0.700 (gender to gender). Source domains are given in rows, targets in columns. Diagonal gray cells have weighted F1 where target and source is the same. As decay increases, red color gets darker.

recovery ratio, its decay is -20%, which shows that the same recovery values do not necessarily mean the same performance.

5 Discussion

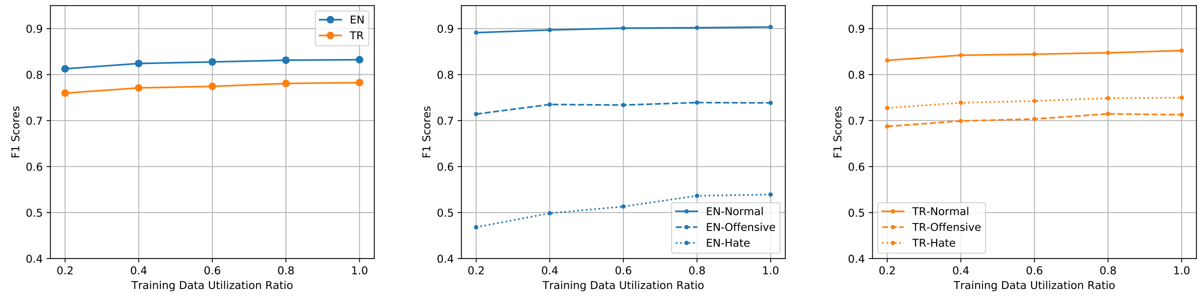
5.1 Scalability

We examine scalability as the effect of increasing training size on model performance. Since labeling hate speech data is costly, the data size of hate speech detection becomes important. Our large-scale datasets are available to analyze scalability. To do so, we split 10% of data for testing, 10% for validation, and remaining 80% for training. From the training split, we set five scale values starting from 20% to 100%. To obtain reliable results, we repeat this process five times, and report the average scores. At each iteration, training and validation datasets are randomly sampled. We re-run BERT for English, and BERTurk for Turkish.

We train the models for five epochs. However, we use the number of epochs that gives the best performance on the validation set, given in Table 6. The motivation is to have a fair comparison by neglecting the positive effect of having more training data, since more number of instances means more number of steps. We observe that using smaller number of instances (e.g. 20% of data size) needs more epochs to converge, compared to larger data.

The results for overall detection performance are given in Figure 1a. We observe that the performance slightly improves as training data increases in both English and Turkish. We also investigate the scalability performance of individual classes in Figure 1b for English, and Figure 1c for Turkish.

For English, normal tweets are the best predicted, while hate tweets are the worst predicted class. Interestingly, the performance of hate class improves



(a) Weighted F1 scores for multi-class hate speech detection with respect to increasing training data. There is a slight performance increase in both languages.

(b) Weighted F1 scores for different classes in English. The performance of normal class saturates early, and hate class benefits the most.

(c) Weighted F1 scores for different classes in Turkish. There is a slight performance increase in all classes.

Figure 1: Scalability analysis for hate speech detection.

Lang./Ratio	20%	40%	60%	80%	100%
EN	3.50	2.30	2.20	1.90	2.08
TR	3.90	3.70	3.33	2.28	2.52

Table 6: Number of epochs when the best model is obtained on validation set for scalability. Maximum epochs is set to 5.

significantly as training data increases. Normal and offensive tweets exhibit a slightly increasing pattern. This result emphasizes the importance of the data size in hate speech detection. Given that the main bottleneck in hate speech detection task is misprediction of hate speech rather than normal tweets, using higher number of data instances has significant effect on hate speech detection performance. On the other hand, the performance of all classes slightly increase in Turkish. Hate tweets are better predicted compared to offensive tweets, showing that language is important to detect hate speech. A reason could be the different speech patterns in different languages. Note that the number of hate tweets in Turkish is larger than those of English, however the performance of English is still worse than Turkish when similar number of training instances are considered (e.g. hate score of ratio 100% in Figure 1b is still worse than the score of 20% in Figure 1c). Overall, collecting hate speech data in large scale contributes to model performance, but not with a substantial degree. However, the best improvement by increasing the train size is observed for the hate class in English.

5.2 Ablation Study

To assess the effect of tweet-specific components on the performance of hate speech detection, we remove each component from tweets, and re-run BERT for English, and BERTurk for Turkish. Tweet-specific components are URLs, hashtags, and emoji symbols. Table 7 reports the experimental results of the ablation study. The results show that removing tweet-specific components has almost no effect on the performance in English. Similar observation is valid for Turkish, but using hashtags has a slight performance improvement.

Data	Model	Acc.	Prec.	Recall	F1
EN	Raw text	0.808	0.679	0.808	0.732
	w/o URL	0.808	0.680	0.808	0.733
	w/o Hashtags	0.807	0.679	0.807	0.732
	w/o Emoji	**0.809**	**0.681**	**0.809**	**0.734**
	w/o All	0.808	0.679	0.808	0.732
TR	Raw text	**0.767**	**0.606**	**0.767**	**0.673**
	w/o URL	**0.767**	**0.606**	**0.767**	**0.673**
	w/o Hashtags	0.763	0.601	0.763	0.668
	w/o Emoji	0.766	0.605	0.766	0.672
	w/o All	0.763	0.601	0.763	0.668

Table 7: The ablation study: Effect of tweet-specific components. The average of 10-fold cross-validation is reported. Highest scores are given in bold.

6 Conclusion

We construct large-scale datasets for hate speech detection in English and Turkish to analyze the performances of state-of-the-art models. With the help of such available data, we also analyze model scalability. We design our datasets to have equal size of instances for each of five hatred domains; so that we report zero-shot cross-domain transfer results in hate speech detection. Future work would focus on a detailed error analysis of hate speech detection. The scalability results are limited to Transformer-based language models, one can further analyze other models. The generalization capability of hatred domains can be examined in other languages.

References

Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *Advances in Information Retrieval*, pages 141–153, Cham. Springer International Publishing.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2021. A deep dive into multilingual hate speech classification. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V*, pages 423–439. Springer International Publishing.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 45–54, New York, NY, USA. Association for Computing Machinery.

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*, pages 759–760.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. 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*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Daniel L. Byman. 2021. How hateful rhetoric connects to real-world violence. Accessed: 2021-10-15.

Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. 2009. Clueweb09 data set.

Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. 2020. Deephate: Hate speech detection via multi-faceted text representations. In *12th ACM Conference on Web Science*, WebSci '20, page 11–20, New York, NY, USA. Association for Computing Machinery.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on Twitter. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 13–22, New York, NY, USA. Association for Computing Machinery.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

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.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. WWW '15 Companion, page 29–30, New York, NY, USA. Association for Computing Machinery.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of Twitter abusive behavior.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. 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 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. Accessed: 2021-10-15.

Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 229–233, New York, NY, USA. Association for Computing Machinery.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, AISec '18, page 2–12, New York, NY, USA. Association for Computing Machinery.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: decoding-enhanced BERT with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Zihang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. 2020. Convbert: Improving BERT with span-based dynamic convolution. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. 2015. An empirical exploration of recurrent network architectures. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 2342–2350. JMLR.org.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Ping Liu, Wen Li, and Liang Zou. 2019a. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilia Markov and Walter Daelemans. 2021. Improving cross-domain hate speech detection by reducing the false positive rate. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 17–22, Online. Association for Computational Linguistics.

Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2021. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Online. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Guanyi Mou, Pengyi Ye, and Kyumin Lee. 2020. Swe2: Subword enriched and significant word emphasized framework for hate speech detection. In

Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pages 1145–1154.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 9–14. Association for Computational Linguistics.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in Twitter: A multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy. Association for Computational Linguistics.

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.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Furkan Şahinuç and Cagri Toraman. 2021. Tweet length matters: A comparative analysis on topic detection in microblogs. In *Advances in Information Retrieval*, pages 471–478, Cham. Springer International Publishing.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Stefan Schweter. 2020. BERTurk - BERT models for Turkish. Accessed: 2021-10-15.

Surendra Sedhai and Aixin Sun. 2015. Hspam14: A collection of 14 million tweets for hashtag-oriented spam research. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 223–232, New York, NY, USA. ACM.

Sanjana Sharma, Saksham Agrawal, and Manish Shrivastava. 2018. Degree based classification of harmful speech using Twitter data. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 106–112, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Sara Sood, Judd Antin, and Elizabeth Churchill. 2012. Profanity use in online communities. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1481–1490.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China. Association for Computational Linguistics.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of English in zero-shot cross-lingual transfer. *CoRR*, abs/2106.16171.

Twitter. 2021. Twitter Transparency Report. Accessed: 2021-10-15.

Elise Fehn Unsvåg and Björn Gambäck. 2018. The effects of user features on Twitter hate speech detection. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 75–85.

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.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the*

NAACL Student Research Workshop, pages 88–93, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

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

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 9054–9065.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving hate speech detection with deep learning ensembles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.