Embeddia at SemEval-2019 Task 6: Detecting Hate with Neural Network and Transfer Learning Approaches

Andraž Pelicon, Matej Martinc, Petra Kralj Novak
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
{Andraz.Pelicon, Matej.Martinc, Petra.Kralj.Novak}@ijs.si

Abstract

SemEval-2019 Task 6 was OffensEval: Identifying and Categorizing Offensive Language in Social Media. The task was further divided into three sub-tasks: offensive language identification, automatic categorization of offense types, and offense target identification. In this paper, we present the approaches used by the Embeddia team, who qualified as fourth, eighteenth and fifth on the three sub-tasks. A different model was trained for each sub-task. For the first sub-task, we used a BERT model fine-tuned on the provided dataset, while for the second and third tasks we developed a custom neural network architecture which combines bag-of-words features and automatically generated sequence-based features. Our results show that combining automatically and manually crafted features fed into a neural architecture outperform transfer learning approach on more unbalanced datasets.

1 Introduction

Over the years, computer-mediated communication, like the one on social media, has become one of the key ways people communicate and share opinions. Computer-mediated communication differs in many ways, both technically and culturally, from more traditional communication technologies (Kiesler et al., 1984). However, the ability to fully or partially hide our identity behind an internet persona leads people to type things they would never say to someone's face (Shaw, 2011). Not only is hate speech more likely to happen on the Internet, where anonymity is easily obtained and speakers are psychologically distant from their audience, but its online nature also gives it a far-reaching and determinative impact (Shaw, 2011). Although most forms of intolerance are not criminal, hate speech and other speech acts designed to harass and intimidate (rather than merely express criticism or dissent), deteriorate public discourse and opinions, which can lead to a more radicalized society.

Online communities, social media platforms, and technology companies have been investing heavily in ways to cope with offensive language to prevent abusive behavior in social media. Social media companies Facebook, Twitter and Google's YouTube have greatly accelerated their removal of online hate speech, and report reviewing over two-thirds of complaints within 24 hours. It has been proven in practice that naive word filtering systems do not manage to scale well to different forms of hate and aggression (Schmidt and Wiegand, 2017). The most promising strategy for detecting abusive language is to use advanced computational methods. This topic has attracted significant attention in recent years as evidenced in recent publications (Waseem et al., 2017; Davidson et al., 2017; Malmasi and Zampieri, 2018).

The SemEval-2019 Task 6 — OffensEval: Identifying and Categorizing Offensive Language in Social Media (Zampieri et al., 2019b) is to use machine learning text classification methods to identify offensive content and hate speech. The task organizers have provided a new dataset (Zampieri et al., 2019a) comprised of Twitter posts which employs a three-level hierarchical labeling scheme, according to the three hierarchically posed sub-tasks, where each sub-task serves as a stepping stone for the next sub-task. Sub-task A aims to identify offensive content, Sub-task B aims to classify offensive content as a targeted or untargeted offense, while Sub-task C aims to identify the target of the offense.

In this paper, we present the approaches used by the Embeddia team to tackle the three sub-tasks of SemEval-2019 Task 6: OffensEval. The Embeddia team qualified as fourth, eighteenth and fifth on Sub-tasks A, B and C, respectively. The Embed-

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 604–610
Minneapolis, Minnesota, USA, June 6–7, 2019. ©2019 Association for Computational Linguistics

dia team used different neural architectures and transfer learning techniques (Devlin et al., 2018). We also explore if combining automatically generated sequence-based features with more traditional manual feature engineering techniques improves the classification performance and how different classifiers perform on unbalanced datasets. Our results show that a combination of automatically and manually crafted features fed into a neural architecture outperforms the transfer learning approach on the more unbalanced datasets of Subtasks B and C.

This paper is organized as follows. In Section 2, we present related work in the area of offensive and hate speech detection. Section 3 describes in more detail the provided dataset and the methodology used for the task. Section 4 reviews the results we obtained on the three sub-tasks with our models. Section 5 concludes the paper and presents some ideas for future work.

2 Related Work

A number of workshops that dealt with offensive content, hate speech and aggression were organized in the past several years, which points to the increasing interest in the field. Due to important contributions of publications from TA-COS[1], Abusive Language Online[2], and TRAC[3], hate speech detection became better understood and established as a hard problem. The report on shared task from the TRAC workshop (Kumar et al., 2018) shows that of 45 systems trying to identify hateful content in English and Hindi Facebook posts, the best-performing ones achieved weighted macro-averaged F-scores of just over 0.6.

Schmidt and Wiegand (2017) note in their survey that supervised learning approaches are predominantly used for hate speech detection. Among those, the most widespread are support vector machines (SVM) and recurrent neural networks, which are emerging in recent times (Pavlopoulos et al., 2017). Zhang et al. (2018) devised a neural network architecture combining convolutional and gated recurrent layers for detecting hate speech, achieving state-of-the-art performance on several Twitter datasets. Malmasi and Zampieri (2018) used SVMs with different surface-level features, such as surface n-grams, word skip-grams and word representation n-grams induced with Brown clustering. They concluded that surface n-grams perform well for hate speech detection but also noted that these features might not be enough to discriminate between profanity and hate speech with high accuracy and that deeper linguistic features might be required for this scenario.

A common difficulty that arises with supervised approaches for hate speech and aggression detection is a skewed class distribution in datasets. Davidson et al. (2017) note that in the dataset used in the study only 5% of tweets were labeled as hate speech. To counteract this, datasets are often resampled with different techniques to improve on the predictive power of the systems over all classes. Aroyehun and Gelbukh (2018) increased the size of the used dataset by translating examples to four different languages, namely French, Spanish, German, and Hindi, and translating them back to English. Their system placed first in the Aggression Detection in Social Media Shared Task of the aforementioned TRAC workshop.

A recently emerging technique in the field of natural language processing (NLP) is the employment of transfer learning (Howard and Ruder, 2018; Devlin et al., 2018). The main idea of these approaches is to pretrain a neural language model on large general corpora and then fine-tune this model for a task at hand by adding an additional task-specific layer on top of the language model and train it for a couple of additional epochs. A recent model called Bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018) was pretrained on the concatenation of BooksCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2,500M words) and then successfully applied to a number of NLP tasks without changing its core architecture and with relatively inexpensive fine-tuning for each specific task. According to our knowledge, it has not been applied on a hate speech detection task yet, however it reached state-of-the-art results in the question answering task on the SQuAD dataset (Rajpurkar et al., 2016) as well as beat the baseline models in several language inference tasks.

3 Methodology and Data

This section describes the tasks, the dataset, the methodology used and the experiments.

[1] http://ta-cos.org/

[2] https://sites.google.com/site/abusivelanguageworkshop2017/

[3] https://sites.google.com/view/trac1/home

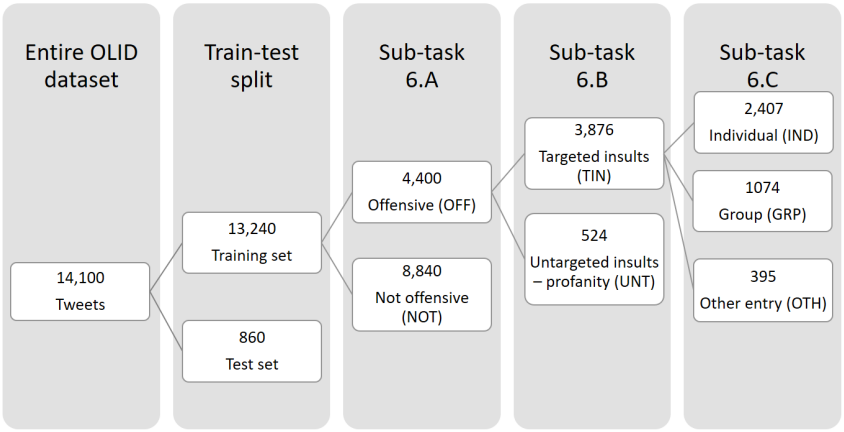


Figure 1: Schema of SemEval-Task 6: OffensEval: Identifying and Categorizing Offensive Language in Social Media. The hierarchy of the sub-tasks and respective dataset sizes.

3.1 Dataset

The SemEval-2019 Shared Task 6: Identifying and Categorizing Offensive Language in Social Media was divided into three sub-tasks, namely offensive language identification (Sub-task A), automatic categorization of offense types (Sub-task B) and offense target identification (Sub-task C). The organizers provided a new dataset called OLID (Zampieri et al., 2019a) which includes tweets labeled according to the three-level hierarchical model. On the very first level, each tweet is labeled as offensive (OFF) or not offensive (NOT). All the offensive tweets are then labeled as targeted insults (TIN) or as untargeted insults (UNT), which simply contain profanity. On the last level, all targeted insults are categorized as targeting an individual (IND), a group (GRP) or other entity (OTH). The dataset contains 14,100 tweets split into training and test sets. The training set containing 13,240 tweets and the test set without labels were made available to the participants for the task. The inspection of the dataset reveals that the classes at first level are slightly imbalanced with the imbalances between classes getting more prominent with each subsequent level. A more detailed breakdown of the dataset is presented in Figure 1. We didn't use any additional datasets in any of the three sub-tasks.

3.2 Methodology

According to the findings from the related work, we decided to test two different types of architectures. First was a pretrained BERT model, which was fine-tuned on the provided dataset for distinguishing offensive and not offensive posts in the Sub-task A. For the sub-tasks B and C, a neural network architecture was developed, which tried to achieve synergy between two types of features that both proved successful in the past approaches to the task at hand, by basing its predictions on a combination of classical bag-of-words features and automatically generated sequence-based features. The three models, as well as their source code, are available for download in a public repository[4].

Three models were trained using the provided dataset, one for each sub-task. In the Sub-task A, the large pretrained BERT transformer with 24 layers of size 1024 and 16 self-attention heads was used for generating predictions on the official test set. A linear sequence classification head responsible for producing final predictions was added on top of the pretrained language model and the whole classification model was fine-tuned on the SemEval input data for 3 epochs. For training, a batch size of 8 and a learning rate of 2e-5 were used. The training dataset for the Sub-task A was randomly split into a training set containing 80% of the tweets and a validation set containing 20% of the tweets. Only a small amount of text preprocessing was needed on the data for the Sub-task A since the dataset already had all Twitter user mentions replaced by @USER tokens and all URLs by URL tokens. Additionally, we lowercased and tokenized the tweets using BERT's built-in tokenizer.

For Sub-task B, the non-offensive tweets were first filtered out of the original dataset. The re-

[4] https://gitlab.com/Andrazp/embeddia-semeval2019

duced dataset had 4400 tweets. To offset the lower quantity of data, we decided to split the dataset into a training set containing 90% of the data and a validation set containing 10% of the data. The second issue with the data was a severe class imbalance as only 12% of tweets in the filtered dataset were labeled as untargeted insults. We decided to resample the dataset in order to minimize the impact of the imbalance on our training. The approach that yielded the best results based on the validation set performance was to randomly remove the instances of the majority class until the classes were balanced. The remaining instances were lowercased and tokenized with the tweet tokenizer from the NLTK package (Bird et al., 2009). Stopwords were also removed from every tweet using an English stopwords list provided in the NLTK package.

As the BERT model was showing worse performance on the resampled data according to the validation set results, a new neural network architecture was devised for this sub-task (Figure 2). The neural architecture takes two inputs. The first input is a term frequency-inverse document frequency (tf-idf) weighted bag-of-words matrix calculated on 1- to 5-grams and character 1- to 7- grams using sublinear term frequency scaling. N-grams with document frequencies less than 5 were removed from the final matrix. Furthermore, the following additional features are generated for each tweet in the training set and added to the tf-idf matrix:

- The number of insults: using a list of English insults,[5] the insults in each tweet are counted and their number is added to the matrix as a feature.

- The length of the longest punctuation sequence: for every punctuation mark that appears in the Python built-in list of punctuations, its longest sequence is found in each tweet. The length of the sequence is then added as a feature.

- Sentiment of the tweets: the sentiment of each tweet is predicted by an SVM model (Mozetič et al., 2016) pretrained on English tweets. The model classifies each tweet as positive, neutral or negative. The predictions are then encoded and added as features.

The second input is word sequences, which are fed into an embedding layer with pretrained 100-dimensional GloVe (Pennington et al., 2014) embedding weights trained on a corpus of English tweets. The pretrained embeddings are additionally fine-tuned during the training process on the dataset for the task. The resulting embeddings are fed to an LSTM layer with 120 units, on the output of which we perform global max pooling. We perform a dropout operation on the max pooling output and the resulting vectors are concatenated with the tf-idf vectors. The resulting concatenation is sent to a fully-connected hidden layer with 150 units, the output of which is fed to a rectified linear unit (RELU) activation function. After performing dropout, final predictions are produced by a fully-connected hidden layer with a sigmoid activation function. For training, we use a batch size of 16 and Adam optimizer with a learning rate of 0.001. We trained the model for a maximum of 10 epochs and validated its performance on the validation set after every epoch. The best performing model was later used for generating predictions on the official test set.

For Sub-task C, the dataset was additionally filtered by removing the tweets that were labeled as non-targeted insults. The class imbalance for this task was even more prominent with only 28% of tweets being labeled as insults targeted towards groups and 10% as targeted insults that do not target an individual or a specific group of people. In light of such class imbalance, the dataset was again undersampled by removing 75% of tweets from the majority class and 50% percent of tweets from the middle class. Due to the dataset being even more aggressively filtered, the 90-10% split from the previous sub-task was kept. A modified version of the neural architecture from Sub-task B was used for prediction. We tried to capture the relationship between insults and their targets using sentence structure information. To this end, we added a third input to the neural architecture that accepts sequences of part-of-speech (POS) tags. First, all the tweets were POS-tagged using the POS tagger from the NLTK package and the resulting POS tag sequences were then fed to a randomly initialized embedding layer. Output embeddings are then fed to an LSTM layer with 120 units, on the output of which we performed global

[5] http://metadataconsulting.blogspot.com/2018/09/Google-Facebook-Office-365-Dark-Souls-Bad-Offensive-Profanity-key-word-List-2648-words.html

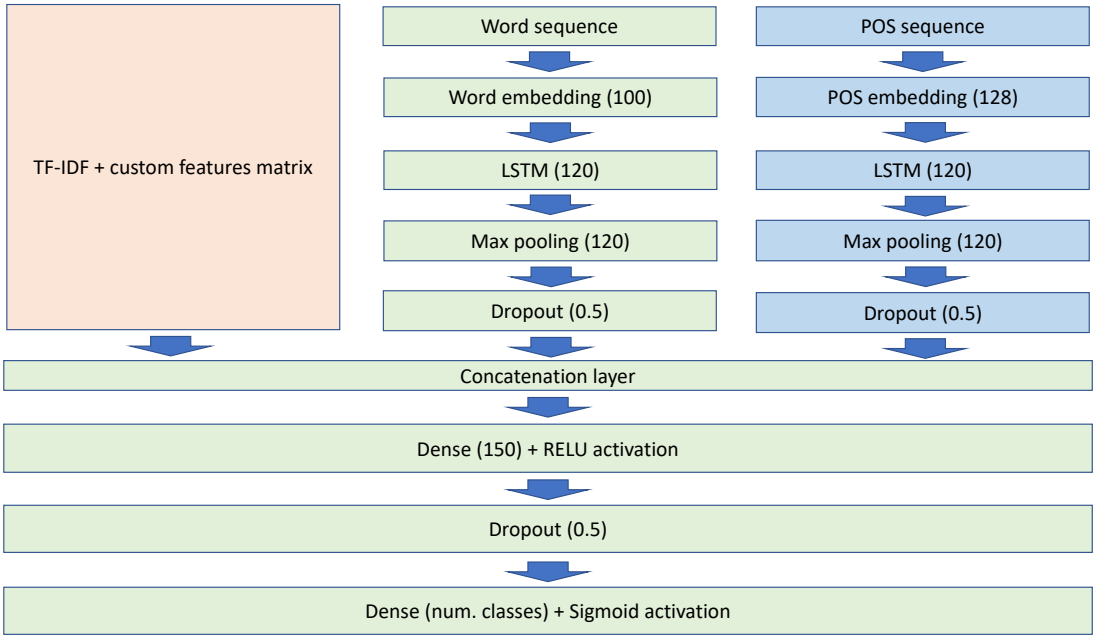


Figure 2: System architecture used in Sub-tasks B and C. Parts of the infrastructure depicted in blue were only used in Sub-task C.

max pooling. Next, dropout was applied, and the resulting vector matrix was then concatenated with the matrices from other inputs and sent to the fully-connected layer (see Figure 2).

4 Results

The results on the official test sets for all three tasks are presented in Table 1. In the Sub-task A, our BERT model, fine-tuned on the provided dataset, achieved a macro-averaged F1 score of 0.808. When we compare this result to other teams participating in the SemEval-2019 OffensEval Sub-task A, we rank fourth.

As the dataset was filtered and the class imbalances became more prominent in the subsequent tasks, the performance of our models started to deteriorate. Even though the undersampling of the dataset to offset class imbalances further reduced the available data, it proved to be the best way to ensure somewhat reliable predictions. The models for Sub-task B and C had macro-averaged F1 scores of 0.663 and 0.613 respectively and placed eighteenth and fifth overall in the SemEval-2019 OffensEval official ranking.

A closer look at the confusion matrices further confirms our claim about the impact of class imbalances on our systems' performance. While the predictions for both classes were fairly accurate in the Sub-task A (Figure 3a), we can see a dwindling performance on the untargeted insults (UNT) class in Sub-task B (Figure 3b) where approximately two thirds of the instances were misclassified as targeted insults (TIN) class on the test set.

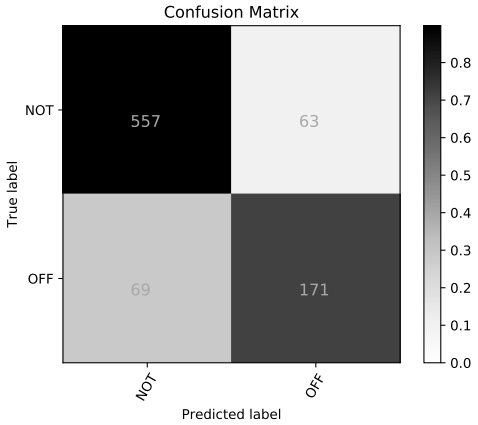
The confusion matrix for Sub-task C (Figure 3c) paints a very similar picture. Even though the majority individual (IND) and middle group (GRP) classes were heavily imbalanced in the original dataset, our model was still able to successfully discriminate between them. However, it again performed subpar on the minority other entity (OTH) class, which was heavily underrepresented compared to the other two. Of the 35 instances in the test set, three out of four were misclassified.

5 Conclusion

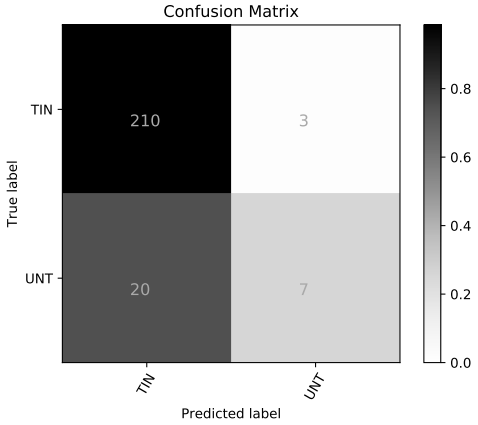
In this paper, we presented the results of the Embeddia team on the SemEval-2019 Task 6: OffensEval: Identifying and Categorizing Offensive Language in Social Media using the dataset provided by the organizers of the task. The task was further divided into three sub-tasks, namely offensive language identification (Sub-task A), automatic categorization of offense types (Sub-task B) and offense target identification (Sub-task C). We trained three models, one for each sub-task. For Sub-task A, we used a BERT model fine-tuned on the OLID dataset, while for the second and third tasks we developed a neural network architecture

Sub-task	System	F1 (macro)	Accuracy
A	BERT	0.8078	0.8465
B	BOW+GloVeLSTM	0.6632	0.9042
C	BOW+GloVeLSTM+POS_LSTM	0.6133	0.7042

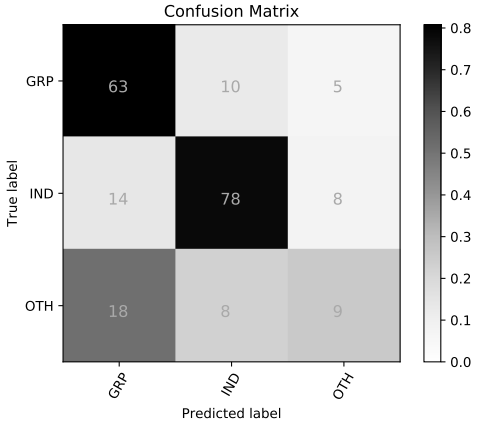
Table 1: Results of the submitted systems for each sub-task.



(a) Confusion matrix for the BERT system, fine-tuned on the provided dataset for Sub-task A.



(b) Confuaion matrix of the two-input neural network with a LSTM based on word sequences and a bag-of-words matrix for Sub-task B.



(c) Confusion matrix of the three-input neural network with an LSTM based on word sequences, LSTM based on part-of-speech tags sequences and a bag-of-words matrix for Sub-task C.

which combines bag-of-words features and automatically generated sequence-based features. Our models ranked fourth, eighteenth and fifth in Subtasks A, B and C, respectively.

We noticed that the class imbalances in the datasets had a significant impact on the performance of our systems and were especially deteriorating for the performance of the BERT system. To counteract the impact of class imbalances we used various techniques to resample the original datasets. While randomly removing instances from the majority classes proved to be the most consistent approach to improve the predictive power of our systems, the effect of the class imbalance persisted.

Our aim for the future is to make the systems more robust to imbalanced data to better generalize over all the classes. Since we already have several models that perform adequately, a good next step would be to implement an ensemble model using a plurality voting or a gradient boosting scheme. We will also conduct an ablation study to identify which features work particularly well for offensive content and hate speech detection.

Acknowledgments

This paper is supported also by the European Unions Horizon 2020 research and innovation programme under Grant No. 825153, EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this publication reflect only the authors views and the Commission is not responsible for any use that may be made of the information it contains. The authors acknowledge also the financial support from the Slovenian Research Agency core research programme Knowledge Technologies (P2-0103). The Titan Xp used for this research was donated by the NVIDIA Corporation.

References

Segun Taofeek Aroyehun and Alexander Gelbukh. 2018. Aggression detection in social media: Us-

ing deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 90–97.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of ICWSM*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Sara Kiesler, Jane Siegel, and Timothy W McGuire. 1984. Social psychological aspects of computer-mediated communication. *American psychologist*, 39(10):1123.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling (TRAC)*, Santa Fe, USA.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in Discriminating Profanity from Hate Speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30:1–16.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual twitter sentiment classification: The role of human annotators. *PloS one*, 11(5):e0155036.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135.

Jeffrey Pennington, Richard Socher, and Christopher 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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection Using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics*, pages 1–10, Valencia, Spain.

LaShel Shaw. 2011. Hate speech in cyberspace: bitterness without boundaries. *Notre Dame JL Ethics & Pub. Pol'y*, 25:279.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Langauge Online*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval)*.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *Lecture Notes in Computer Science*. Springer Verlag.

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.