5-2021

An Empirical Study of Offensive Language in Online Interactions

Diptanu Sarkar
ds9297@rit.edu

An Empirical Study of Offensive Language in Online Interactions

by

Diptanu Sarkar

A thesis submitted
in partial fulfillment of the
requirements for the degree of
Master of Science
in
Computer Science

Department of Computer Science
B. Thomas Golisano College of Computing and Information Sciences
Rochester Institute of Technology

Rochester, New York
May 2021

An Empirical Study of Offensive Language in Online Interactions

by

Diptanu Sarkar

APPROVED BY

SUPERVISING COMMITTEE:

Dr. Marcos Zampieri, Chairperson

Dr. Alexander G. Ororbia, Reader

Dr. Christopher Homan, Observer

Date

An Empirical Study of Offensive Language in Online Interactions

by

Diptanu Sarkar

Abstract

In the past decade, usage of social media platforms has increased significantly. People use these platforms to connect with friends and family, share information, news and opinions. Platforms such as Facebook, Twitter are often used to propagate offensive and hateful content online. The open nature and anonymity of the internet fuels aggressive and inflamed conversations. The companies and federal institutions are striving to make social media cleaner, welcoming and unbiased. In this study, we first explore the underlying topics in popular offensive language datasets using statistical and neural topic modeling. The current state-of-the-art models for aggression detection only present a toxicity score based on the entire post. Content moderators often have to deal with lengthy texts without any word-level indicators. We propose a neural transformer approach for detecting the tokens that make a particular post aggressive. The pre-trained BERT model has achieved state-of-the-art results in various natural language processing tasks. However, the model is trained on general-purpose corpora and lacks aggressive social media linguistic features. We propose fBERT, a retrained BERT model with over 1.4 million offensive tweets from the SOLID dataset. We demonstrate the effectiveness and portability of fBERT over BERT in various shared offensive language detection tasks. We further propose a new multi-task aggression detection (MAD) framework for post and token-level aggression detection using neural transformers. The experiments confirm the effectiveness of the multi-task learning model over individual models; particularly when the number of training data is limited.

WARNING: Due to the nature of the work, this report contains words and texts that are offensive or hateful.

Acknowledgments

I would like to express my gratitude towards my advisor Dr. Marcos Zampieri for his supervision throughout this work. His teachings have been significant for building a strong foundation in Natural Language Processing. It has been a long journey, and I am thankful for his support and guidance throughout.

I am extremely thankful to my co-advisor Dr. Alexander G. Ororbia for his invaluable feedback and guidance. His suggestions have helped me build robust and performant models. I would like to extend my thanks to Dr. Christopher Homan for serving on my committee and giving constructive feedback on my work.

I am deeply grateful to Tharindu Ranasinghe from the University of Wolverhampton, who collaborated and recommended improvements to strengthen my work. Finally, I also want to acknowledge RIT Research Computing for providing high computing resources during the work.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the last decade, social media has become a significant component of human life. Social media platforms, blogs provide a stage for free speech and express opinions openly. Social media platforms such as Facebook, Twitter, Reddit has billions of organic users. These platforms enable users to be connected to friends and family and share news as well. Information and opinions on these platforms can spread rapidly and can reach millions of users in minutes. Platforms such as LinkedIn help professionals to grow their network and find new job opportunities. While some people use social media to make their living, social media has become an integral part of our life.

The open nature of platforms provides users the voice to convey their opinions. However, often those discussions turn toxic, damaging the sanity of social media platforms. The concept of anonymity on the internet further fuels aggressive and hateful behavior. These platforms are used to drive targeted hatred, propagate biases and polarize a large number of users. Online abusive and inflamed content, posts, comments can be against gender, race, religion, sexual orientation, nationality. There's an ongoing debate on free speech and social media company-designated regulations, where the latter can be discriminatory. Facebook has recently admitted that the platform was actively used to incite violence against the Rohingya minority in Myanmar [6]. Sri Lanka banned social media due to an increase in anti-Muslim sentiments after a terrorist attack [102]. Social media companies are now being scrutinized by regulators around the world [49]. According to a study, targets of hate speech on Twitter are primarily people's race and behavior [96].

In the last decade, the safety and sanity issues of social media platforms have increased. The companies and government institutes are trying to keep online interactions safe, secure, and welcoming to everyone. Human moderation on these social media platforms does not scale as there are over millions of posts every day. Lexicon-based approaches to flag offensive contents are inadequate to solve the problem at scale. These approaches can not discriminate between profanity, hate speech, cyberbullying, sarcasm. All profane content need not be offensive or aggressive, rather used to express concern [63]. There are state-of-the-art models based on discriminatory approaches to detect and flag offensive posts on online content [27, 63, 69].

Data exploration aids the understanding of any possible relations, patterns, anomalies of the data. We choose three commonly used datasets in hate speech and offensive language domain – offensive language identification dataset [111], hate speech against immigrants and women [5], hate speech and offensive language twitter dataset [27]. Topic modeling is an unsupervised statistical method to discover general topics that occur in a set of documents. In the first exploratory study, we use statistical and neural topic modeling to extract underlying topics that appear in the datasets. Furthermore, we compare the dataset features derived by unsupervised and supervised techniques.

1.1 Motivation

Past studies have presented various state-of-the-art approaches for detecting offensive language in online interactions. In studies [63, 105], the researchers note that n-gram based features are beneficial for building reliable automated hate speech detection models. Lately, pre-trained bidirectional transformer models outperformed previous state-of-the-art Long Short-Term Memory (LSTM) [42] and Gated Recurrent Unit (GRU) [22] models on hate speech and offensive content detection problems [59, 83]. While we have various state-of-the-art post or document-level aggression detection models, the problem of word or token-level toxicity detection is understudied. Human moderators can benefit from word-level indicators while reviewing flagged social media content. Fine-grained aggression detection will provide a more interpretable recourse to the moderators and save time. However, the problem

NOTE: We use the terms offensive and aggressive interchangeably.

is challenging as profanity does not always indicate toxicity, and understanding context is crucial to detect token-level aggression. Furthermore, attaining consensus among annotators is difficult due to the subjective nature of the task, and varies across background, culture, nationality, political motivations of the annotators. We propose an efficient token-level aggression detection model using neural transformers using the toxic Spans Detection dataset [73].

Pre-trained language models have performed quite efficiently in several Natural Language Processing (NLP) tasks. The Bidirectional Encoder Representations from Transformers (BERT) [28] network has achieved state-of-the-art results in language understanding, question answering, named entity recognition, text classification. Bidirectional transformer models have outperformed other deep learning models in recent shared offensive language detection challenges [110, 112]. However, the models are generally pre-trained on general-purpose corpora and lack social media or aggressive language cues. Retraining language models on task or domain-specific data before fine-tuning have improved performance over vanilla models [80]. Recently, various studies proposed domain-specific BERT models – finance [3], legal [18], biomedical [57], social media [71]. Nonetheless, large-scale pre-trained language models explicitly for social media offensive language detection are underexplored. In this study, we present a retrained BERT model, fBERT, adapted to aggression detection tasks.

Document-level hate speech and offensiveness detection is a well-researched domain. Mathew et al. [66] noted the lack of explainable hate speech detection study and released a new benchmark dataset HateXplain, incorporating various features of hate speech in English. To the best of our knowledge, no previous studies have explored the problem of both post and token-level aggression detection at the same time. In multi-task learning, the model learns multiple related tasks simultaneously and shares information among the objectives. Multi-task learning can also serve as a regularizer and minimizing the chances of overfitting. In this study, we explore the multi-task learning approach to predict both document and token-level aggression using transformers. We show that the multi-task learning model achieves better performance over individual models when the labeled dataset is scarce. We further introduce a new robust multi-tasking aggression detection framework (MAD) for both document and token-level toxicity in social media posts.

To summarize, the major contributions of the study are –

- A highly performant token-level aggression detection transformer model to aid human moderators while reviewing flagged content on social media.
- We introduce fBERT, a new effective retrained BERT adapted to social media aggression detection.
- The first publicly available multi-task aggression detection (MAD) framework for both post and token-level aggression detection using neural transformers.

The rest of the paper is organized as follows. Chapter 2 presents previous work on offensive language and hate speech datasets and classification models. The dataset exploration findings are discussed in Chapter 3. Chapter 4 presents the token-level aggression detection challenge. The fBERT pre-trained model is discussed in Chapter 5. Chapter 6 comprises the architecture of the MAD framework. Finally, Chapter 7 summarises the conclusion and direction for future work.

Chapter 2

Related Work

Hate speech and offensive language detection in social media is a well-researched field of NLP. Previously, there have been several studies examining semantic and lexical features for detecting hate speech [30, 72]. Some extensive studies majorly focused on topical issues such as racism [20], cyberbullying [21]. Silva et al. [96] proposed a large-scale hate target detection methodology using sentence structure characteristics. They also note that most hate speech found on social media targeted to race, behavior, physical appearance, and sexual orientation. While unsupervised lexicon-based approaches provide standard baselines, it completely loses the context of the post or sentence. Hate speech is specific instances intended to degrade or insult a specific group of people. An offensive or profane post may not be hate speech. To overcome the problem, researchers often use manual or crowd-sourced annotated datasets applying various machine learning techniques [2, 13, 104]. The basic framework consists of training models on the annotated datasets using various linguistic features to detect offensive language and hate speech.

2.1 Offensive and Hate Speech Datasets

In the past years, several benchmark offensive and hate speech datasets have been released. Davidson et al. [27] presented a fine-grained English tweets dataset containing hate speech, offensive, and neither labels. In [111], the authors manually annotated an offensive language identification dataset identifying the type and target. The study introduced a new three-layer hierarchi-

cal annotation scheme – class, categorization, and target of offensive language. However, the size of the dataset was limited to train large deep learning models. Rosenthal et al. [91] further extended the dataset in a semi-supervised manner that contains over nine million annotated English tweets. Alongside the extended English dataset, OffensEval 2020 [112] presented offensive language detection datasets in Arabic, Danish, Greek, and Turkish using the same hierarchical annotation scheme. A multilingual hate speech dataset specifically targeted against immigrants and women was released on SemEval 2019 Task 5 (HatEval) [5]. Furthermore, offensive language datasets have been annotated in other languages such as Arabic [70], Dutch [100], Hindi [65], Bengali [89], Portuguese [32]. However, most of the existing datasets contain a sentence or post-level annotation of hate speech or offensiveness.

The toxic spans detection [73] challenge introduced a word-level annotated dataset collected from the toxic and severely toxic instances of the Civil Comments Dataset [12]. Toxic span was defined as the sequence of words that make a particular instance toxic. Recently, Mathew et al. [66] released a first benchmark dataset for explainable hate speech detection covering three primary areas – classification, targeted community, and the spans that make the text hateful or offensive. To the best of our knowledge, this is the first dataset that includes both document and token-level annotation for hate speech detection.

2.2 Models

In an early work, Waseem and Hovy [105] used character n-grams along with other metadata such as gender, location for hate speech detection on tweets. Even though combined features perform better, n-gram constitutes a reliable base to achieve good results. In [27], authors introduced a multi-class logistic regression classifier to detect hate speech and offensive language using n-grams weighted by term frequency-inverse document frequency. Malmasi and Zampieri [63] presented the challenge of detecting hate speech and profane non-hate speech on social media. They applied character and word n-grams, and word skip-grams as features in a linear SVM classifier. N-gram has proven to be a robust linguistic feature for automatic hate speech and offensive post detection.

Recently, various studies have employed deep learning methods along-

side NLP techniques to achieve good results. Convolutional Neural Networks (CNN) and LSTM networks are highly used in the newer models [4, 33]. Various sub-types of offensive language – aggression [52, 53], cyberbullying [90], and hate speech [87] detection model were also proposed. The focus has also shifted towards offensive language detection in various languages such as Arabic [1, 70], Danish [95], Greek [75], Turkish [17], and multi-lingual systems [48, 79, 81, 83]. However, these models can not generalize accurately due to the limited availability of annotated datasets and the highly subjective nature of the task. Shared tasks such as OffensEval [110, 112], HatEval 2019 [5], TRAC [53] presented various challenging tasks related to offensive language and hate speech detection. However, the tasks are limited to post or document-level aggression detection and do not provide any token or word-level indication.

While the detection of hate speech and offensive language was studied extensively, detection of tokens or words that make posts offensive is so far underexplored. Often human moderators have to go through lengthy posts to find out offensive or hateful contents. The task of finding toxic part(s) in social media posts presented in SemEval 2021 Task 5: Toxic Spans Detection [73]. Recently, Ranasinghe and Zampieri [82] noted that neural transformer models achieve state-of-the-art performance in offensive spans detection. They further introduced a multilingual framework for token-level offense detection. Nonetheless, the problem is overlooked majorly due to the lack of fine-grain annotated datasets.

The BERT language model was pre-trained on a large amount of English Wikipedia and BookCorpus [115] datasets using unsupervised masked language modeling (MLM) and next sentence prediction objectives. Inspired by the recent success of BERT in various NLP tasks, neural transformer models have outperformed traditional deep learning models in post-level offensive and hate speech detection tasks [83, 84]. In OffensEval 2019 [110], Liu et al. [59] used a BERT-based model that achieved the highest F_1 score in the competition. However, BERT was pre-trained on general corpora and often needs domain-specific language features. Various studies proposed specialized variation of BERT – financial domain FinBERT [3], LEGAL-BERT [18] for legal specialty, BerTweet [71] for tweet-specific tasks. Lately, HateBERT [16] was released for abusive language detection using the Reddit abusive language dataset. However, the model lacks tweet-specific aggression detection cues that can achieve better performance in a wide variety of similar tasks.

Multi-task learning (MTL) [15] is a training paradigm where the model learns multiple related tasks simultaneously from the data. MTL can generalize the model better and reduce the chances of overfitting. Multi-task learning models can also avoid sub-optimal solutions compared to a single objective model [7]. By sharing features learned during similar tasks, MTL overcomes the requirement of a large amount of data. Recognizing the advantages, multi-task architectures are used in several machine learning fields – computer vision [35, 114], various sub-tasks of NLP [23, 58, 60] and achieved exceptional results.

Lately, multi-task learning is also widely used in enhancing the performance of various hate speech and offensive language detection tasks. The majority of past research acknowledge one task as the "main" task supported by other auxiliary tasks. In [50], MTL is used in deep neural network architecture to improve the performance of hate speech detecting leveraging information from multiple comparable classification tasks. Abu Farha and Magdy [1] utilized MTL with a CNN-Bi-LSTM model to detect hate speech using sentiment prediction as an auxiliary task. Similarly, BERT-based multi-task approaches are also used for detecting the offensive language [26] and multilingual aggression and misogyny detection [94]. Despite recent success, to the best of our knowledge, no study previously explored a multi-task learning approach for both post and token-level aggression detection.

Chapter 3

Dataset Exploration

3.1 Introduction

Data exploration aids the understanding of any possible relations, patterns, anomalies and nature of the data. Traditionally, various statistical techniques are employed to gain insights into the data. In this exploratory work, we have chosen three popular datasets in the offensive language identification domain – offensive language identification dataset [111], hate speech and offensive language twitter dataset [27] and hate speech against immigrants and women [5]. The goal of this study is to explore underlying topic modeling techniques to extract topics that appear in the datasets. Furthermore, we are interested to know the features that supervised learning techniques may learn in document classification.

In the experiments, we applied Latent Dirichlet Allocation (LDA) [10], a widely used topic modeling technique to find the optimal number of topics and topic-word distribution. We further applied a simple neural topic model to compare the results with LDA. We found that the topics mostly portray the keywords that were originally used to compile the datasets. Furthermore, we note that supervised algorithms may discriminate two discordant datasets using the prominent features determined by the topic modeling algorithms.

3.2 Datasets

3.2.1 Offensive Language Identification Dataset

In past years various social media offensive language datasets have been released. Zampieri et al. [111] compiled a new fine-grained hierarchically annotated offensive language identification dataset (OLID) classifying the type and target of the offensive contents. The OLID was the official dataset for a popular shared task OffensEval 2019: OffensEval: Identifying and Categorizing Offensive Language in Social Media [110]. The three-layer dataset comprises three labels: Offensive Language Detection (offensive and not offensive), Categorization of Offensive Language (targeted insult and untargeted), and Offensive Language Target Identification (individual, group, and other). The training dataset consists of $13,240$ instances, out of which $4,400$ instances are positive – i.e., offensive. The test set contains 860 tweets that include 240 positive examples.

3.2.2 Davidson Dataset

In finer-grain aggression detection, classifying among offensive language and hate speech is challenging. Hate speech is a specific type of offensive language that incites violence, and attacks targets based on religion, nationality, ethnicity, gender, sexual orientation, appearance, and others [31]. Davidson et al. [27] compiled a $24,783$ English tweets dataset annotated in three labels – hate speech, only offensive, and neither. Hereafter, we refer to the dataset as the Davidson dataset. The dataset contains $1,430$ hate speech, $19,190$ only offensive, and $4,163$ instances that are neither.

3.2.3 HatEval Dataset

In SemEval-2019 Task 5 [5], a new multilingual hate speech dataset (HatEval) against women and immigrants was introduced. The dataset comprised $13,000$ English and $6,600$ Spanish tweets. The dataset incorporates three categories: HS (hateful or not), Target Range (individual or generic), Aggressiveness (aggressive or not). Note that the dataset identifies only hateful tweets against women and immigrants as a positive class – i.e., hate speech; and all other instances (including offensive and hateful) are labeled as negative or not hate

speech. In this study, we only focused on English tweets. The official English training set contains $9,000$ tweets that incorporate $4,177$ hateful instances. The dev and test set contains $1,000$ and $3,000$ examples, out of which 123 and $1,380$ instances are positive.

3.3 Topic Modeling

With the increase in data generation, it has become extremely challenging to organize, understand, and retrieve information from a vast amount of textual data. Topic modeling is an unsupervised machine learning technique that can statistically discover abstract topics in a collection of documents. Topic modeling methods cluster similar words together, and a cluster of words is represented as a topic. It is used to extract hidden semantic features in a large collection of textual documents. Topic modeling is used for document classification, sentiment analysis, semantic structure discovery as well as in various disciplines such as healthcare [97], bioinformatics [9], computer vision [14].

3.3.1 Latent Dirichlet Allocation

In information retrieval, the term frequency-inverse document frequency (tf-idf) evaluates how relevant a word is to a document in a given document collection. However, the inter and intra statistical structure of the document remain unexplored. To overcome that latent semantic analysis (LSA) [56], probabilistic LSA (pLSA) [43] are proposed for topic extraction. Nonetheless, pLSA does not provide document-level probabilistic models and can not generalize unseen documents well. Latent Dirichlet Allocation is a generative statistical model widely used for topic modeling. In LDA, the documents are interpreted as a finite random mixture of latent topics, and the topics are represented as a Dirichlet distribution over infinite words [10]. It assumes that every word in a text can be assigned a probability of belonging to a topic.

In LDA, a document (w) is defined as a sequence of words and a corpus (D) is a collection of documents. The topic distribution for the document is represented as a multinomial distribution θ and α is defined as a Dirichlet hyperparameter to generate θ. The hyperparameter β matrix represents the distribution of words per topic. If k is dimension of Dirichlet distribution and

V is the size of vocabulary, then β represents the word probability in $k * V$ matrix, where β_{ij} represents the probability of word w_j in topic i. A set of topics z and words w is produced by –

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta) \tag{3.1}$$

The marginal distribution of the document is obtained by –

$$p(w|\alpha, \beta) = \int p(\theta|\alpha)(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta))d\theta \tag{3.2}$$

The probability of the corpus can be obtained by the product of individual documents marginal probability –

$$p(D|\alpha, \beta) = \int p(\theta_d|\alpha)(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta))d\theta_d \tag{3.3}$$

In the probabilistic graphical model Figure 3.1, the outer box represents the collection of documents and the inner box refers to the topics and words in a document. LDA improves the previous pLSI model by adding Dirichlet prior β and α parameters. In LDA, unseen documents are classified as the same as training and the words in a document are randomly generated from topics using α. The generative process is presented in Algorithm 1. In this study, we used the Gensim [86] library LDA implementation to extract topics from the offensive language datasets.

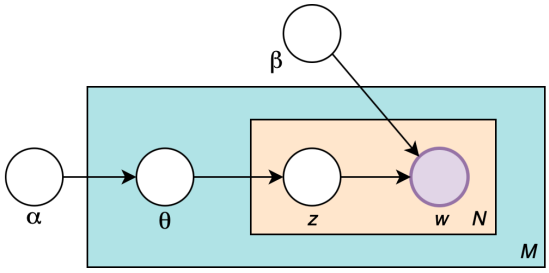


Figure 3.1: The graphical model representation of LDA [10].

Algorithm 1: The generative process for LDA [10].

for *each document in the corpus* **do**
  Choose total words $N \sim \text{Poisson}(\xi)$.
  Choose multinomial topic distribution $\theta \sim \text{Dirichlet}(\alpha)$.
  for *each word in the document* **do**
    Pick a topic z_n using multinomial distribution over θ.
    Pick a word w_n using multinomial probability conditioned on the topic z_n.
  end
end

3.3.2 Neural Topic Modeling: doc2topic

A neural take on LDA may provide better topic and word distribution in the latent semantic space. The doc2topic [93] neural topic model (NTM) computes co-occurrences between words and documents in two separate embedding spaces whose dimensions are the number of topics. The embedding layer is a linear transformation with rectified linear unit (ReLU) activation.

$$E = \text{ReLU}(xW^T + b) \tag{3.4}$$

$$\text{ReLU}(x) = max(0, x)$$

Where x is the input, W represents the weight vector and bias b. For topics to documents sparsity, the document embeddings are highly regularized compared to the word embeddings. The network is trained by feeding two words (one present in the document and another random word) at once through the embedding layers and calculating the dot product of the embeddings followed by sigmoid activation.

$$E_i = E_{doc} \odot E_{word} \tag{3.5}$$

$$\hat{y_i} = \sigma(E_i) \tag{3.6}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Where σ denotes the sigmoid function, $\odot$ is the dot product. The true output y_i is 0 for the random words and 1 for the co-occurring samples. We maximize

the probability of $\hat{y}$ to predict the true labels and the loss is calculated using binary cross-entropy.

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(\hat{y}_i) + (1 - y_i) \cdot log(1 - \hat{y}_i) \tag{3.7}$$

Where N is the output size. The final weights of two embedding layers represent the document to topic and topic to word distributions. The graphical representation of the model is shown in Figure 3.2.

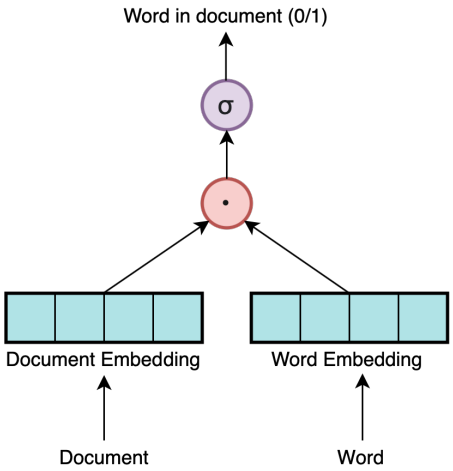


Figure 3.2: The graphical representation of doc2topic model.

3.3.3 Methodology

Data Pre-processing Data cleaning and preparation is a crucial step in data analysis. It helps to reduce the complexity and noise of the dataset and deliver good results [34].

Tokenization We used spaCy [44] software package to tokenize the documents in the dataset. The URLs, emojis, usernames, extra whitespace, punctuations are also removed.

Generalization All the texts are converted to lower-case. Stop words and words less than three characters are removed.

Lemmatization It is the process of converting infected words to their dictionary form. We used spaCy Lemmatizer for lemmatizing the tokens.

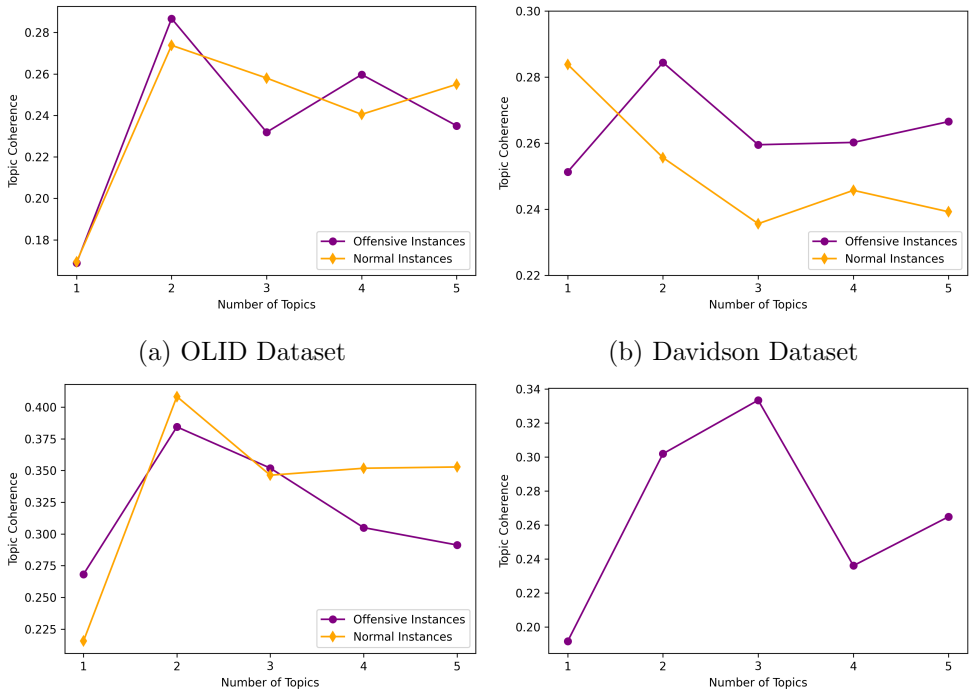
We further divided all the datasets into two subsets of offensive or hateful instances and normal instances. The subsets are denoted by offensive and normal suffixes.

Experiments We applied topic modeling to extract the topics in the three popular hate speech and offensive language datasets. For LDA we used bag-of-words (BOW) as features by varying the number of topics. We also fine-tuned model hyperparameters document-topic density (alpha) and word-topic density (beta). Besides, we combined the offensive or hateful instances of OLID and HatEval datasets to verify if topic modeling can distinguish the topics in two distinct datasets. The combined dataset is referred to as the **OLID-HatEval** dataset. As per our expectations, we found three prominent topics in the combined dataset experiment. We further repeated the same steps with the doc2topic NTM to analyze the variations in topics and words compared to LDA.

3.3.4 Evaluation and Results

Topic modeling is an unsupervised technique for finding latent topics in a vast collection of textual information. Evaluating topic models is not apparent due to the lack of ground truth annotation in datasets. Also, topic models are required to provide the number of topics beforehand. Traditionally, eyeballing methods such as looking at top n-words in topics are used to evaluate topic models. Perplexity is a statistical measure of how well the distribution predicts a sample and is used to compare language models in NLP. However, better predictive perplexity does not correlate to human interpretable topics [19]. Topic coherence is a measure of semantic similarity in top occurring words in topics. We used c_v topic coherence measure that uses normalized pointwise mutual information and the cosine vector similarity [99]. We further interpret the topic models using visualization and top n-words in topics across all the data subsets.

OLID Dataset In the OLID dataset, both offensive and non-offensive instances achieve the best topic coherence score when the number of topics is



(a) OLID Dataset

(b) Davidson Dataset

(c) HatEval Dataset

(d) OLID-HatEval Dataset

Figure 3.3: Topic coherence scores for the datasets with varying number of topics.

2 as shown in Figure 3.3a. The top 10 word distribution by topic is shown in Appendix A.1. Eyeballing over the words in topics, we can determine that one topic portrays political cues and another swear words with some outliers. The data collection procedure of the OLID dataset states the use of politically motivated keywords that we also found in topic modeling [111]. Neural topic modeling also found similar results presented in Appendix A.2; nonetheless, the two topics are not highly discrete.

Davidson Dataset The Davidson dataset contains offensive and hate speech samples; we observe the same topic distribution in the results, one concerning hate speech and another general profanity. The normal instances depict more benign words best represented in a single topic. The topic coherence score with a varying number of topics is shown in Figure 3.3b. The distribution of words per topic using LDA and NTM is presented in Tables A.5 and A.6.

HatEval Dataset Topic modeling on the HatEval dataset prominently portrays two discrete topics – profanity against women and keywords related to immigrants. The topic coherence score is highest with 2 topics as shown in Figure 3.3c. The keywords associated with the topics are presented in Tables A.3 and A.4.

	LDA
Topic-1:	bitch, woman, fuck, fucking, like, skank, hoe, get, ass, cunt
Topic-2:	immigration, maga, illegal, buildthatwall, migrant, country, buildthewall, refugee, alien, trump
Topic-3:	gun, antifa, liberal, people, control, maga, conservative, know, hysterical, american
	NTM: doc2topic
Topic-1:	bitch, whore, woman, fucking, people, stupid, hoe, cunt, want, rape
Topic-2:	illegal, buildthatwall, maga, look, make, refugee, work, alien, stop, think
Topic-3:	gun, fuck, liberal, control, shit, trump, people, illegal, maga, fucking

Table 3.1: Top 10 words distribution by topic for the combined OLID-HatEval dataset using LDA and NTM.

OLID-HatEval Dataset As presented in Table 3.1, both statistical and neural topic modeling on the OLID-HatEval dataset reveals three distinct topics in the dataset. The profanity against women and keywords on immigration are extended from the HatEval dataset and political keywords are infused from the OLID dataset. Topic coherence score is also the highest with 3 topics, shown in Figure 3.3d.

In our experiments, we found that LDA provides good insights into the topics present in offensive language datasets. A simple neural topic model also provides similar human-comprehensible results. From the results, we observe that the OLID dataset portrays more politically inflected keywords. The Davidson dataset exhibits words that correspond to hate speech and profan-

ity and the HatEval dataset topics centered on profanity against women and immigrants-related keywords. We found that topics in the datasets denote the type of keywords employed to compile the dataset and different datasets focused on separate themes. To support our claim further, we perform supervised dataset classification using the combined OLID-HatEval dataset.

3.4 Dataset Classification

Document classification is the problem of assigning documents to classes or categories based on the document contents. In supervised document classification, the model maps the input to output by learning hidden patterns in the training data to predict the class for unseen data. Motivated by the distinction of topics in the OLID-HatEval dataset, we applied supervised learning to discriminate the features in the dataset. The goal of the experiment is to explore supervised learning features and compare them to unsupervised topic modeling results.

3.4.1 Dataset

For the task, we combined only the offensive or hate speech instances of OLID and HatEval datasets into a new OLID-HatEval dataset. The HatEval dataset lacks the fine-grained annotation – i.e. if a particular hateful instance is against women or immigrants. Therefore in the combined dataset, we labeled the OLID instances as '0' and HatEval instances as '1'. The final dataset comprised 4639 OLID and 4031 HatEval tweets. Table 3.2 shows a sample of the combined dataset. We further used 20% of the consolidated dataset as a test set.

3.4.2 Experiments and Results

We normalized the dataset by removing extra whitespaces, stop words, and punctuations and replaced the usernames and URLs with placeholders. We further lemmatized the words to remove inflectional endings. We used the term frequency-inverse document frequency to convert texts into vectors with maximum features of 200 words. We trained logistic regression, naive bayes, and support vector machine models with *10*-fold cross-validation. We used the macro F_1 score to compare the results of different classifiers.

Text:	Crisis in Germany: women start arming themselves due to Islamic immigration Crisis - Free Speech Time - https://t.co/Xrt0RQk3k0 @ISupportIsrael
Dataset:	1
Text:	@liberalparty I was in Toronto last year. It no longer looked like Canada. Very sad situation. The Canadians were good people. https://t.co/7bHGpS1cH0 Stop immigration. Start deportations. We have the right to our homelands.
Dataset:	1
Text:	@USER @USER Yes mothers now this is what the deranged left will make you do to your little boys! Are we really going to let this happen? Pure EVIL! #VoteDemsOut #MAGA #2020 #ConfirmJudgeKavanaugh @USER URL
Dataset:	0
Text:	@charliekirk11 We need to improve our country, schools, neighborhoods, hospitals, prisons, by kicking out the illegals. #BuildTheDamnWall #BuildThatWall
Dataset:	1
Text:	@USER How about just F*!k her @USER she is an instigating big mouth bitc*; so is her evil friend.@USER MY OPINION!!! #teamtani
Dataset:	0

Table 3.2: A sample of the consolidated OLID-HatEval dataset. Dataset labels 0 represents OLID and 1 denotes HatEval instances.

Model	**Test F_1 Score**
Support Vector Machine	**0.93**
Logistic Regression	0.90
Naive Bayes	0.89

Table 3.3: Results of OLID-HatEval dataset classification ordered by the test set macro F_1 score.

From the results shown in Table 3.3, we can observe that all the classifiers achieved good performance, and the SVM classifier with linear kernel achieves the highest macro F_1 score of 0.93. We further plot the top 10 features and their importance in discriminating the datasets in Figure 3.4. We can observe

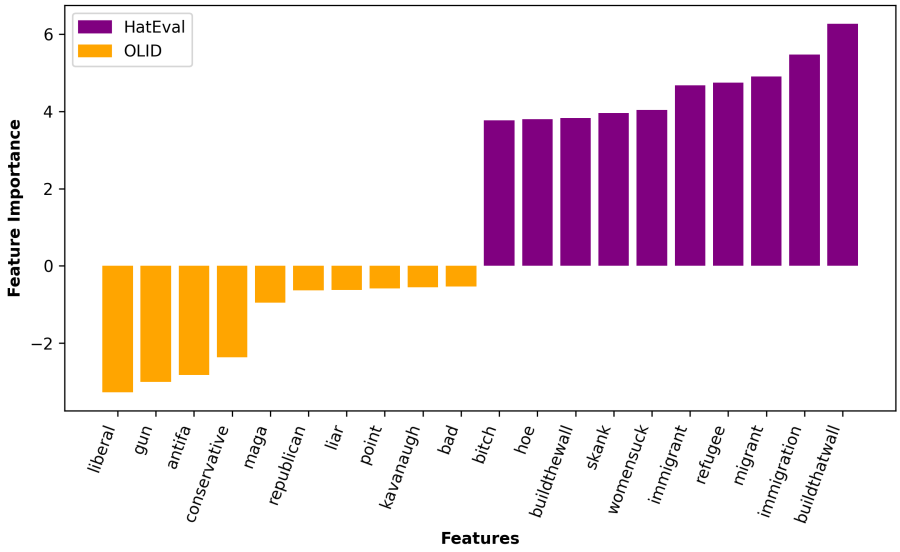


Figure 3.4: Top 10 feature words of the SVM classifier and their importance in predicting OLID and HatEval instances.

that the OLID dataset attributes political words, whereas the HatEval dataset presents obscene terms against women and immigration-oriented keywords. Furthermore, close observation reveals that over 70% of the top 10 words in the supervised features matches the unsupervised topic modeling results.

3.5 Discussion

Dataset exploration helps to understand datasets better by finding hidden patterns, relations, and even anomalies. Topic modeling is a widely used unsupervised topic extraction method from a vast collection of documents. In this study, we applied Latent Dirichlet Allocation and neural topic modeling to extract topics from three highly popular offensive and hate speech detection datasets. We discovered the optimal number of topics in the datasets by topic coherence measure. We presented the top 10 words by topics that appear in the dataset. We found that LDA and a simple neural topic model discover human-interpretable comparable topics. It is also noted that the topic and word distribution mostly resembles the keywords used to compile the dataset and different dataset focuses on distinct themes.

To further support our claim, we extended the dataset comparison in a

supervised manner. We used three machine learning models to discriminate among the OLID and HatEval dataset instances. The top features from the best performing SVM model exhibit alike results as unsupervised learning. To summarize, we derived topics from three popular datasets in the offensive language detection domain and found that topics mostly resemble the nature of compiling datasets.

Chapter 4

Token-level Aggression Detection

4.1 Introduction

The use of offensive and aggressive language in social media has increased over the years. The offensive language detection problem has interested many researchers in academia and industrial institutions. An automatic aggression detection system can keep social media clean, welcoming by flagging inappropriate content. Previously, studies have focused on n-gram, skip-gram features for automatic hate speech detection [63, 64, 105]. Some studies focused on detecting sub-types of offensive language in social media such as aggression [52, 53], cyberbullying [90], and hate speech [87]. Shared tasks such as TRAC[53], OffensEval 2019 [110], OffensEval 2020 [112], HatEval [5] mainly focused on post-level aggression detection. Recently, the focus has shifted towards abuse detection in various languages such as Arabic [1, 70], Danish [95], Greek [75], Turkish [17] and also multi-lingual systems [48, 79, 81, 83].

Over the years, researchers released various hate speech and offensive language detection datasets [5, 27, 91, 111]. All the datasets focus on the post-level classification of the social media posts, comments, or conversations. Identification of tokens or words that inherently make a post offensive or hateful is mostly overlooked [66]. Human moderators often have to read through lengthy texts without any word-level indications, making the task difficult. Token annotations can aid human moderators while reviewing the flagged contents.

The SemEval 2021 Task 5: Toxic Spans Detection [73] presents a new dataset (TSD) with token-level toxic spans annotation. Recently to the best of our knowledge, Mathew et al. [66] recently released the first benchmark dataset for explainable hate speech detection with both post and token-level annotations.

Since the introduction of BERT [28], neural transformer models have become prevalent in offensive and hateful language identification. BERT models have outperformed previous deep learning models in post-level aggression detection [4, 33, 41]. Transformer models have achieved excellent results in offensive language detection in resource-limited languages such as Bengali [81], Malayalam [84]. However, the current models are limited to post-level aggression detection and do not predict word-level aggression. In this study, we use the TSD dataset to propose a robust token-level aggression detection system. We experimented with a lexicon-based word matching algorithm, a recurrent neural network model, and neural transformer models. We found RoBERTa with language modeling and ensembling outperforms all other approaches and achieves a 0.68 F_1 score in the test set. We have also submitted our best system to the SemEval 2021 Task 5.

4.2 Dataset

The dataset for token-level aggression detection is introduced in SemEval 2021 Task 5: Toxic Spans Detection [73]. The sequence of tokens or words that make a particular instance aggressive is defined as toxic spans. The dataset is compiled from the toxic and severely toxic instances of the Civil Comments dataset [12]. Each instance is annotated by three annotators using the crowd annotation platform Appen[1]. The toxic spans detection dataset contains the text and spans that are toxic. The training data consists of $7,939$ instances, out of which 690 instances do not contain any toxic spans. The trial and test sets include 690 and $2,000$ instances, that includes 43 and 394 instances without toxic spans. Table 4.1 shows four instances from the training dataset. The text presents the Text and the spans are indicated by the character index (starting from zero) position in the Text. The whitespace between two consecutive aggressive tokens is also marked as toxic.

[1]Appen annotation platform: `https://appen.com`

Text:	Propaganda of homosexuality should be forbidden.
Spans:	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26]
Text:	How fucking stupid are you?
Spans:	[4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17]
Text:	That is not a friggen hat, It s called a cover you idiots...
Spans:	[14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 51, 52, 53, 54, 55, 56]
Text:	What an inciting pile of trash including the website. You're dangerous and need to be monitored.
Spans:	[]
Text:	No, Victimitis is an a-s-s h-o-l-e
Spans:	[21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33]
Text:	You must be ecstatic with your "F****** Moron" who doesn't care enough about the military people
Spans:	[32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44]

Table 4.1: Five instances from the training dataset along with their annotations. The aggressive tokens are displayed in red and the spans are indicated by index.

4.3 Methodology

4.3.1 Lexicon-based Word Match

Word matching algorithms often achieve balanced results and provide a good baseline in similar problems. We compiled an offensive word lexicon using publicly available profanity word resources[2,3]. We further added the aggressive words from the training dataset that are not present in the lexicon. We run a substring matching algorithm using the trie data structure and identify the indices of the tokens. As anticipated, this method does not consider words in context and misses words that are not present in the lexicon. Moreover, partially censored words such as *f**k* are also not detected by the algorithm. Nonetheless, this method presents a baseline performance for the task.

[2] `https://www.cs.cmu.edu/~biglou/resources/bad-words.txt`
[3] `https://github.com/RobertJGabriel/Google-profanity-words`

4.3.2 Recurrent Networks: Long Short-Term Memory

Long short-term memory (LSTM) [42] networks is a type of recurrent neural network (RNN) capable of learning long-term dependencies. The hidden layer of LSTM is replaced with a memory cell and proved better at retaining and leveraging long-term dependencies compared to RNNs. An LSTM memory cell consists of an input gate (i), forget gate (f), cell state (c), and output gate (o). Regulated by the gates, LSTM can add or remove information to the cell state. It uses feedback connections to learn order dependencies (previous to present) in sequential data. LSTM networks are highly effective in sequential problems such as time series prediction, language translation, speech recognition [37, 39]. The input layer of the network takes one-hot-encoded features at time t. The input layer size n is the same as the maximum sequence length. For each input x_t the context representation vector h_t is computed as follows –

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o)$$

$$h_t = o_t \odot tanh(c_t) \tag{4.1}$$

Where σ denotes the sigmoid function, $\odot$ is the element-wise product. Also, W represents the weight matrix of the subscripts, and b values are the biases. Nonetheless, the classic LSTM network is only able to use previous contexts in the sequence. Bidirectional LSTM (Bi-LSTM) [38] is capable of learning both forward and backward contextual information using two hidden LSTM layers. The input is provided to both forward and backward networks to provide bidirectional contextual information. The final context representation of a word $\hat{h}_t$ is calculated as concatenation of forward $\overrightarrow{h_t}$ and backward $\overleftarrow{h_t}$ vectors –

$$\hat{h}_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{4.2}$$

The bidirectional architecture has outperformed unidirectional LSTMs in speech processing and sequence tagging problems [25, 38]. Conditional random fields (CRF) [54] are a discriminative model capable of incorporating contextual information to predict the current label. A CRF layer embedded on top of the Bi-LSTM network can model past and future contextual information to predict the current tag. For an input sequence $x = (x_1, x_2, ..., x_n)$ and the output matrix of Bi-LSTM network P with dimension $n * l$, where l is the number of output labels. The output is represented as $y = (y_1, y_2, ..., y_n)$. The label probability matrix is computer as –

$$s(x, y) = \sum_{i=0}^{n} A_{yi,yi+1} + \sum_{i=1}^{n} P_{i,yi}$$

The probability of sequence y is calculated using softmax function –

$$p(y|x) = \frac{e^{s(x,y)}}{\sum e^{s(x,y)}}$$

The objective function is the maximal log probability of correct labels and defined as –

$$log(p(y|x)) = s(x, y) - log(\sum e^{s(x,y)}) \tag{4.3}$$

During training, we maximize the probability of $\hat{y}$ to predict the true labels, and the maximum score is calculated as –

$$\hat{y} = arg\ max\ s(x, y) \tag{4.4}$$

For token-level aggression detection, we combined the Bi-LSTM with CRF to create a Bi-LSTM-CRF architecture, similar to the previous sequence tagging state-of-the-art model [47]. The model architecture is presented in Figure 4.1, where the forward and backward LSTM networks are used to learn past and future input features and the top CRF layers outputs the final tag by using the sentence level tag information. The model is trained using backpropagation through time (BPTT) [11]. The final model has 4.2 million trainable parameters. We experimented by varying hyperparameters and achieved the best test set result while training the model for 5 epochs with a 0.005 learning rate, mini-batch size of 16, and maximum sequence length of 200.

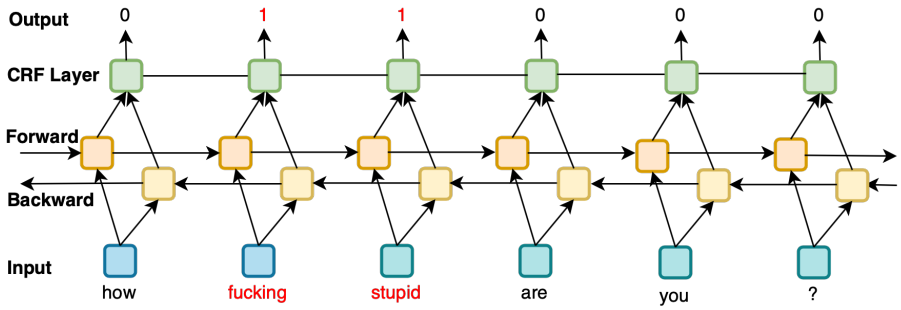


Figure 4.1: The Bi-LSTM-CRF architecture for token-level aggression detection. Non-aggressive and aggressive tokens are shown as 0 and 1, respectively.

4.3.3 Neural Transformers

In recent years, neural transformers have achieved state-of-the-art performance in sequence modeling and language translation tasks [101]. It overcomes parallelization and long-term dependency bottlenecks in RNNs by modeling global dependencies between the input and output. The transformer architecture consists of six stacked identical encoder and decoder layers with a self-attention mechanism. The encoder is composed of a multi-head attention mechanism and position-wise fully connected feed-forward network. The sub-layers are connected using residual connection succeeded by layer normalization. If the sub-layer function is denoted as $S(x)$, the output of each sub-layer is computed as –

$$\text{Output} = \text{LayerNorm}(x + S(x))$$

The decoder additionally includes a multi-head attention sub-layer over the output from the encoder block. An attention function maps a query (Q), key (K), and value (V) vectors to an output vector. The scaled dot-product attention is calculated as –

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{4.5}$$

Where d_k is the dimension of the key and $\frac{1}{\sqrt{d_k}}$ is the scaling factor. Multi-head attention performs better over a single attention function and is applied by projecting Q, K, V vectors h times with different linear projections. It allows the model to acquire information from several representations at various positions.

$$\text{MH Attention}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, ..., \text{head}_h)W^O \tag{4.6}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Where the projection parameters are matrices $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, $W_i^O \in \mathbb{R}^{hd_v \times d_{model}}$. The various use of multi-head attention in transformers is described in [101]. The feed-forward network in encoder and decoder blocks consists of two liner transformations with ReLU activation.

$$\text{F}(x) = \max(0,\, W_1 x + b_1)W_2 + b_2 \tag{4.7}$$

The model further introduces positional encoding of dimension d_{model} to retain positional information of the tokens. The sinusoidal positional embedding is added to the input embedding at the bottom of the encode and decoder blocks. The positional embeddings are defined as –

$$\text{PE}_{pos,2i} = \sin(\frac{pos}{10000^{2i/d_{model}}}) \tag{4.8}$$

$$\text{PE}_{pos,2i+1} = \cos(\frac{pos}{10000^{2i/d_{model}}}) \tag{4.9}$$

Pre-trained bidirectional neural transformers models have achieved excellent results in various NLP tasks. Bidirectional encoder representation from transformers (BERT) [28] is pre-trained on a large amount of English Wikipedia and BookCorpus [115] datasets using unsupervised masked language modeling (MLM) and next sentence prediction tasks (NSP) objectives. The BERT base architecture consists of 12 bidirectional transformers encoders with 768 hidden layers and 12 bidirectional self-attention heads. Along with previous token and positional embeddings, BERT has introduced segmentation embeddings to denote the sentence number of every token. It has outperformed previous deep learning models in many NLP sub-task such as question answering, text classification. Bidirectional transformer models are also highly effective for sequence labeling tasks such as named entity recognition [62].

We integrated a token-level classifier with the uncased BERT model. The classifier takes the last hidden state of the sequence as input and predicts a

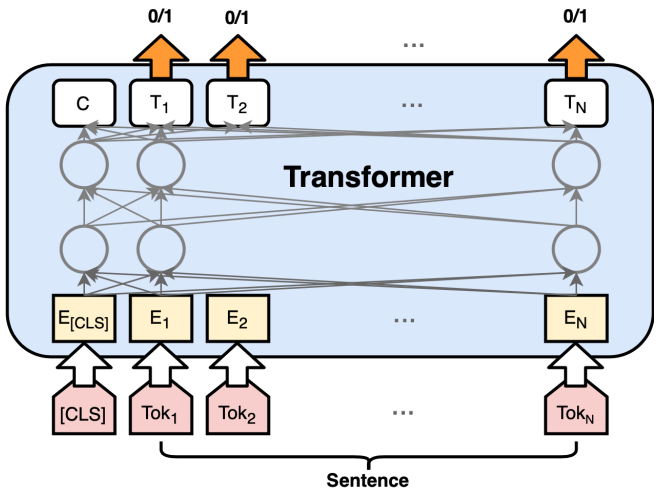


Figure 4.2: The transformer model for token-level aggression detection. Non-aggressive and aggressive tokens are shown as 0 and 1, respectively.

toxic or non-toxic label for each token as output. A general transformer model for token-level aggression detection is shown in Figure 4.2. We fine-tuned the model using Adam optimizer with a maximum sequence length of 400. The hyperparameters are presented in Appendix B.1. Previously, BERT-CRF architecture has shown improvements over BERT models in similar tasks [46, 98]. We further added a CRF layer between the BERT and the token-level classifier and fine-tuned the model for the problem. We introduced dropout with a 0.2 probability of randomly dropping a neuron in the CRT layer to add some regularization. Unfortunately, the BERT-CRT model did not improve the performance compared to the BERT model. Furthermore, we experimented with transfer learning to improve the performance by training BERT on HateXplain [66] dataset and later fine-tune on the TSD dataset. However, transfer learning also did not improve the results any further. Fine-tuning language models on MLM objectives before performing downstream tasks deliver better results [80]. We re-trained the BERT model on the MLM objective with the TSD training dataset. We use the same previously mentioned parameters along with early stopping and models ensembling. Early stopping is executed when the validation loss did not improve over 10 steps. Retraining with ensembling has improved the performance of the BERT model, we denote the model as BERT-Ensemble.

RoBERTa [61] is an optimized BERT variant that is pre-trained on 160

GB English data without NSP loss. RoBERTa outperforms BERT's previous state-of-the-art results in downstream tasks such as GLUE [103], RACE [55], SQuAD [77, 78] tasks. Inspired by the performance improvements of the BERT-Ensemble model, we followed the same re-training and ensembling strategy with a RoBERTa base model. The training hyperparameters are reported in Appendix B.1. The RoBERTa-Ensemble model outperformed the BERT-Ensemble model and achieved the highest performance on the test set.

4.4 Evaluation and Results

For evaluation we used the same F_1 measure presented by the SemEval 2021 Task 5 organizers [73]. Let model A_i return a set $S^t_{A_i}$ of character indices for parts of a text that are toxic. Let G_t be the character indices of the ground truth annotations of post t. We compute the F_1 score of system A_i with respect to the ground truth G for t as mentioned in Equation 4.10 where $|.|$ denotes set cardinality. P^t and R^t measure the precision and recall, respectively.

$$F_1^t\left(A_i, G\right) = \frac{2 \cdot P^t\left(A_i, G\right) \cdot R^t\left(A_i, G\right)}{P^t\left(A_i, G\right) + R^t\left(A_i, G\right)} \tag{4.10}$$

$$P^t\left(A_i, G\right) = \frac{|S^t_{A_i} \cap S^t_G|}{|S^t_{A_i}|} \tag{4.11}$$

$$R^t\left(A_i, G\right) = \frac{|S^t_{A_i} \cap S^t_G|}{|S^t_G|} \tag{4.12}$$

The results of the experiments are presented in Table 4.2. From the results, we can observe that the Lexicon-based word match achieved a balanced result despite being a rule-based method. All the neural transformer models have outperformed the Bi-LSTM-CRF model. The BERT model performs better than the BERT with a CRF layer and transfer learning BERT HateXplain model. Language modeling and model ensembling improve BERTs performance. The ensembled RoBERTa achieves the highest trial and test set F_1 scores. Our RoBERTa-Ensemble achieves a 0.68 test set F_1 score, which is very comparable to 0.70, the best test set score in the SemEval 2021 Task 5 competition. In this study, we presented an efficient token-level aggression detection model using RoBERTa. We also found that language modeling and ensembling further improves the result.

Model	Trial F_1	Test F_1
RoBERTa-Ensemble	0.6886	0.6801
BERT-Ensemble	0.6771	0.6698
BERT	0.6738	0.6538
BERT-CRF	0.6643	0.6517
BERT HateXplain	0.6387	0.6326
Bi-LSTM-CRF	0.5631	0.5398
Lexicon word match	0.3378	0.4086

Table 4.2: Token-level aggression detection results ordered by Test F_1 score. The Trial and Test F_1 score shows the F_1 score on the trial and test set.

4.5 Discussion

In the past years, researchers on social media aggression detection have been mainly focused on instance-level. Shared tasks such as OffensEval 2019 [110], 2020 [112], HatEval [5] presented post-level toxicity detection challenges. However, the problem of word or token-level aggression detection is understudied [66]. A finer-grain detection of toxicity may aid human moderators while reviewing flagged contents in social media. Recently, SemEval 2021 Task 5: Toxic Spans Detection [73] presents the problem of token-level toxicity detection. We believe that token-level toxicity detection is an important step towards explainable aggression detection.

We started our experiments with a simple lexicon-based word matching by compiling an offensive words lexicon from online resources. The result of the rule-based algorithm is balanced and serves as a baseline for the supervised models. The neural transformer models outperformed the recurrent network Bi-LSTM-CRF model. The RoBERTa transformers with language modeling and ensembling achieve the highest trial and test set F_1 scores. In the study, we experimented with various deep learning models for detecting token-level aggression using the TSD dataset. We proposed a robust and efficient transformer model for word-level toxicity detection that can benefit human moderators while reviewing flagged posts. We also submitted our best-performing model to SemEval 2021 Task 5 [85].

Chapter 5

fBERT: Adapting BERT to Aggression Detection

5.1 Introduction

To curtail the pervasiveness of offensive and hateful posts on social media researchers over the years have developed various automatic abusive language detection systems. Early studies utilized linguistic features with linear classifiers for the task [63, 105]. Later on, deep neural network models [51, 68, 83], transfer learning architectures [1, 88, 107], and pre-trained language models have achieved excellent results [59, 82, 107]. One general observation is that the performance of these models varies with datasets and architecture; a support vector machine (SVM) model has outweighed complex neural transformers in hate speech detection against immigrants and women [5]. On the contrary, pre-trained language models BERT [61] have outperformed other neural architectures in general offensive language detection tasks [110, 112].

The introduction of bidirectional encoder representation from transformers (BERT) [28] language model has been a pivotal point in various NLP sub-fields – language understanding, question answering, named entity recognition, text classification. The model is pre-trained on a large amount of English Wikipedia and BookCorpus [115] datasets using unsupervised masked language modeling and next sentence prediction objectives. Subsequently, multiple variations of language models, e.g., RoBERTa [61], XLNet [109], XLM-R [24] are introduced. These language models are trained on a large

amount of general-purpose data for better language understanding and lack domain-specific knowledge. For that reason, recently, domain-specific pre-trained language models are proliferating – financial domain FinBERT [3], LEGAL-BERT [18], BerTweet [71] for tweet-specific tasks.

In chapter 4, we observed that retraining the language models before downstream tasks improves the performance over vanilla pre-trained models and also learns dataset-specific features. Recently, Caselli et al. [16] proposed HateBERT, BERT retrained on Reddit English abusive language dataset for abusive language detection. However, the model lacks tweet-specific aggression detection cues that can achieve better performance in a wide variety of similar tasks. In this study, we present a pre-trained BERT model, fBERT specific to aggression and hate speech detection in English tweets. The fBERT is trained on over 1.4 million offensive instances of English tweets. We further show the effectiveness and portability of fBERT over BERT and HateBERT on various aggression and hate speech detection tasks.

5.2 Retraining Dataset

The limited size of datasets has been a bottleneck for aggression detection tasks. Lately, Rosenthal et al. [91] released a large-scale offensive language identification dataset SOLID with over 9 million English tweets. The dataset follows the same annotation taxonomy as previously discussed OLID [111] dataset. The data is collected using Twitter streaming API and annotated using semi-supervised methods. All the usernames and URLs are replaced with placeholders and tweets less than two words or 18 characters were discarded. For retraining, we have chosen over 1.4 million offensive instances from SOLID. We considered an instance of the SOLID dataset to be offensive if the average score is more than 0.5. We choose the threshold to maximize the training data while presenting offensive language cues to the model. We did not pre-process the data before training as cleaning may negatively impact the model's understanding of tweets and aggressive language features.

5.3 Development of fBERT

In this section, we will provide a detailed overview of the retraining procedure of fBERT.

5.3.1 Input Representation

We take the sentence input and tokenize it using WordPiece embeddings [108] with $30,000$ token vocabulary as described in [28]. The tokenized input can be presented as –

$$X = (x_{[CLS]}, x_1, x_2, ..., x_n, x_{[SEP]}) \tag{5.1}$$

We further process the tokenized input through $Bert(X)$ in Equation 5.2 to generate contextualized embeddings.

$$H = Bert(X) \tag{5.2}$$

$$X' = (h_{[CLS]}, h_1, h_2, ..., h_n, h_{[SEP]}) \tag{5.3}$$

5.3.2 Retraining Procedure

The goal of the study is to adapt the BERT model for social media aggression detection tasks. We utilized a BERT base uncased model that consists of 12 bidirectional transformers encoders with 768 hidden layers and 12 self-attention heads. To use the general understanding of the English language and context, we initialize the BERT with pre-trained weights[1]. We used over 1.4 million offensive texts from the SOLID dataset to retrain the model. No cleaning was applied to preserve the incoherent composition of social media posts, such as excessive use of mentions, emojis, hashtags. We retrained the model on masked language modeling objective to adapt deep bidirectional representation of social media offensive language.

Masked Language Modeling In MLM, we randomly mask some percentage of tokens and predict the masked tokens. As described in the original BERT implementation, we randomly select 15% of the total tokens for replacement, and 80% of the selected tokens are replaced with $[MASK]$, 10% are substituted with a random token from the vocabulary and 10% remain unchanged. The hidden vectors with masked tokens are fed into softmax activation over the vocabulary to generate the probability of masked tokens (P_{MLM}).

[1]BERT Pre-trained weights: `https://github.com/google-research/bert`

$$P_{MLM} = softmax(h * W + b) \tag{5.4}$$

The model is trained to predict the original token by minimizing cross-entropy loss. A schematic representation of the BERT masked language model is shown in Figure 5.1.

$$Loss_{MLM} = -\sum\sum y \ * \ log(P_{MLM}) \tag{5.5}$$

Retraining Setup We retrained the BERT for 25 epochs on MLM objective with 0.15 probability to randomly mask tokens in the input. The language model is trained with a batch size of 32, $5e-5$ learning rate with Adam optimizer, and 512 maximum token length. The complete training took 5 days on an Nvidia V100 GPU. The shifted BERT variant, fBERT, now has a better comprehension of social media linguistic cues and offensiveness.

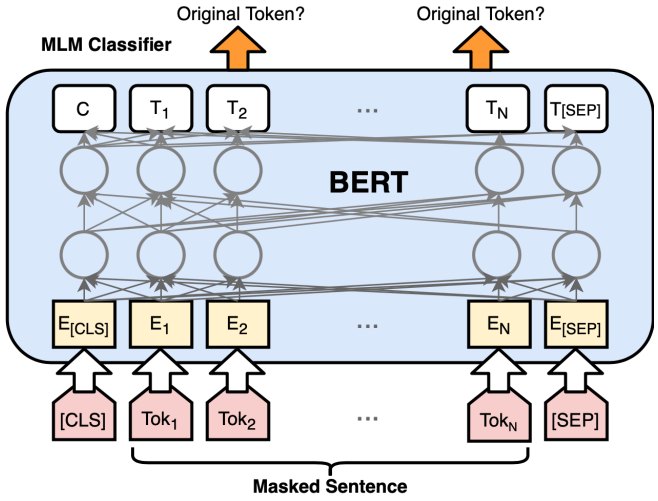


Figure 5.1: A schematic representation of BERT masked language model for retraining.

5.4 Experiments

To determine the effectiveness and portability of the retrained fBERT, we conducted a series of experiments and compared it with a general-purpose BERT model.

HatEval 2019 In the SemEval 2019, HatEval [5] introduced the challenge of detecting multilingual hate speech against women and immigrants. The dataset for the task is described earlier in Section 3.2.3. As pre-processing, we removed extra whitespaces, usernames and URLs are replaced with placeholders, used the Emoji[2] package to convert the emojis to text, and the Word Segmentation[3] package to segment the words in hashtags. We applied the same pre-processing steps and fine-tuned the models to compare the test set macro F_1 score.

OffensEval 2019 In one of the most popular offensive language detection tasks of SemEval 2019, Zampieri et al. [110] presented OffensEval. For the experiment, we have chosen sub-task A, which is a binary classification task between offensive and non-offensive posts. The dataset for the competition is described earlier in Section 3.2.1. We used 10% of the training data as development data. We performed the pre-processing and cleaning steps mentioned by Liu et al. [59]. We trained fBERT for the offensive language detection task and compared the performance with other language models using the macro F_1 score.

Hate Speech and Offensive Language Detection In fine-grain aggression detection, classifying offensive language and hate speech is challenging. Hate speech is explicit instances targeted towards a specific group of people intended to degrade or insult. We used the Davidson dataset described in Section 3.2.2 for the task. The dataset contains three categories – hate speech, offensive, or neither. We further split the dataset into training, dev, and test sets in a 3:1:1 ratio. We applied the same preprocessing steps mentioned in the previous HatEval 2019 Section 5.4. We applied the same preprocessing and fine-tuning steps to compare the effectiveness of fBERT.

[2]Emoji Package: `https://pypi.org/project/emoji/`
[3]Word Segmentation Package: `https://pypi.org/project/wordsegmentation/`

Toxic Spans Detection The toxic spans detection challenge [73] was presented in SemEval 2021. The sequence of words that make a particular post offensive or toxic is defined as toxic spans. The dataset for the task is previously discussed in Section 4.2. We compared the performance of various BERT-based models using the F_1 measure defined in Equation 4.10.

The hyperparameters for the above experiments are presented in Appendix B.2.

5.5 Results

In the HatEval Sub-task A, fBERT has outperformed BERT by increasing the test macro F_1 score by over 23%. That proves the advantage and generalizability of the domain-specific retrained BERT model. The best model [48] in the task used an SVM model with RBF kernel, exploiting sentence embeddings from Google's Universal Sentence Encoder as features. The results are shown in Table 5.1. The fBERT also performs better than the generic BERT and abusive language HateBERT in OffensEval Sub-task A, pushing the test set macro F_1 score to 0.8132 shown in Table 5.2. From the results presented in Table 5.3, we observe that the fBERT is also highly effective in fine-grain offensive and hate speech detection and obtained a 10% increase in F_1 score. Unfortunately, fBERT underperformed compared to the vanilla BERT model in the toxic spans detection; however, it still outperforms HateBERT. We believe that the reason for performance degradation is majorly due to the dataset source. The source of the Toxic Spans Detection dataset is news site Civil Comments platform, which deviates from general social media language. Also, the annotation of the dataset is found to be inconsistent in some cases.

HatEval 2019 Sub-task A	
Model	**Macro F1 Score**
fBERT	0.5962
HateBERT	0.5245
BERT	0.4832

Table 5.1: The test set macro F_1 scores for HatEval 2019 Sub-task A.

OffensEval 2019 Sub-task A	
Model	**Macro F1 Score**
fBERT	0.8132
HateBERT	0.8011
BERT	0.7943

Table 5.2: The test set macro F_1 scores for OffensEval 2019 Sub-task A.

Hate Speech and Offensive Language Detection	
Model	**Macro F1 Score**
fBERT	0.8781
HateBERT	0.8456
BERT	0.8063

Table 5.3: The test set macro F_1 scores for Hate Speech and Offensive Language Detection.

Toxic Spans Detection	
Model	**F1 Score**
BERT	0.6538
fBERT	0.5303
HateBERT	0.4109

Table 5.4: The test set F_1 scores for Toxic Spans Detection.

From the above experiments, we can observe that fBERT has outperformed the abusive language HateBERT model in all the experiments. The proposed fBERT has also performed efficiently in all the post-level aggression detection tasks. This validates the effectiveness of the proposed shifted BERT model for offensive and hateful language classification tasks. The proposed fBERT model is also effective across different datasets and objectives, demonstrating its generalizability.

5.6 Discussion

Over the years, neural transformer models have outperformed previous state-of-the-art deep learning models in various NLP tasks. Neural transformers have obtained highly competitive results in earlier offensive and hate speech detection tasks. In Chapters 4 and 6, we also present transformer-based mod-

els for aggression detection tasks. Nevertheless, these transformers are trained on general corpora, and lack tweet and offensive language-specific cues. Previous studies have shown pre-training the language model before performing downstream tasks achieves excellent results. In this study, we present a large-scale retrained BERT model specifically for social media offensive and hate speech detection tasks.

We proposed fBERT, a retrained *bert-base-uncased* model with over 1.4 million offensive instances from the SOLID dataset on MLM objective. The shifted fBERT model has better incorporation of toxic social media characteristics. The fBERT has achieved better results in OffensEval and HatEval tasks over BERT and HateBERT; additionally, in the next chapter, we further showed that fBERT outperforms BERT in both post and token-level aggression detection. The retrained model is publicly available on GitHub[4].

[4]fBERT model: `https://github.com/imdiptanu/fbert`

Chapter 6

A Multi-task Aggression Detection Framework using Transformers

6.1 Introduction

The users in popular social media platforms can easily be exposed to aggressive and hateful posts or comments. The openness of the platforms and the sense of anonymity fuel the aggressive behaviors among the users. In the past, aggression detection tasks were mainly dominated by machine learning models aided by NLP techniques [27, 52]. In studies [63, 64, 105], researchers noted that n-gram and skip-gram features capture a deeper understanding of the text in automatic hate speech detection. Various deep learning models – convolutional neural networks (CNN) and long short-term memory (LSTM) networks are also developed for the task [4, 8, 36]. Inspired by the recent success of BERT [28] in various NLP tasks, neural transformer models have outperformed traditional deep learning models in post-level offensive and hate speech detection tasks [83]. In SemEval 2019 Task 6 [110] identifying offensive language in social media, Liu et al. [59] used a BERT-based model that achieved the highest F_1 score. Neutral transformer models also yield excellent results in multilingual aggression detection where annotated datasets are limited [81, 84, 112].

Despite the recent success of transformer models for post-level offensive and hate speech identification, the problem of word or token-level aggression detection is understudied [66]. As discussed in the previous chapter, finer-grain detection of toxicity may aid human moderators while reviewing flagged contents in social media. Nonetheless, SemEval-2021 Task 5: Toxic Spans Detection [73] introduces the task of token-level aggression detection. In this study, we have proposed a highly efficient transformer model for token-level toxicity detection. Furthermore, Ranasinghe and Zampieri [82] recently developed an open-source multilingual framework, MUDES, for offensive spans detection using transformers.

Although we have separate state-of-the-art models for post and token-level aggression detection, to the best of our knowledge, there are no models for the detection of both post and token-level aggression at the same time. In this study, we propose a robust multi-task learning framework, MAD, for both post and token-level aggression detection using neural transformers. We further hypothesize that a multi-task learning framework can share information learned between the two objectives and apply that to achieve a similar or better overall performance compared to a single-task setup.

6.1.1 Transfer Learning

Transfer learning has lead to various breakthroughs and improvements in computer vision. In the past years, transfer learning is also used in NLP tasks using pre-trained language models Universal Language Model Fine-Tuning (ULMFit) [45], Embedding from Language Models [74], and has seen good improvements over previous state-of-the-arts. The introduction of BERT [28], a pre-trained language model that can be fine-tuned on downstream tasks, has transformed many NLP tasks such as question answering, named entity recognition, sentence classification. BERT is pre-trained on a large amount of English Wikipedia and BookCorpus [115] datasets using unsupervised masked language modeling (MLM) and next sentence prediction tasks (NSP) objectives. The BERT architecture consists of 12 bidirectional transformers encoders with 768 hidden layers and 12 self-attention heads. Inspired by BERT's success, another optimized BERT version, RoBERTa [61], is proposed that is pre-trained on 160 GB English data without NSP loss. RoBERTa outperforms BERT's previous state-of-the-art results in downstream tasks such as GLUE

[103], RACE [55], SQuAD [77, 78] tasks.

Over the past years, the application of transfer learning proliferated in the offensive language detection domain. Rizoiu et al. [88] used transfer learning for detecting hate speech using Bi-LSTM and also proposed a single representation embedding for hateful contents. Transfer learning with transformers has achieved state-of-the-art results in different offensive and hate speech detection tasks where the datasets are limited [59, 69]. In [81], the researchers used cross-lingual contextual word embeddings and transfer learning to detect offensive occurrences in resource-scarce languages such as Bengali, Hindi, Spanish. Transfer learning is also found to be effective in detecting hateful statements in code-switched languages [79].

6.1.2 Multi-task Learning

Multi-task learning (MTL) [15] is a training paradigm where the model learns multiple related tasks simultaneously from the data compared to the traditional single-task setup. Our choice of MTL is inspired by the notion that MTL allows learning low-level representation of the data that can be shared across different tasks. This can help information learned in one task to apply to the other tasks. In multi-task learning, several parameters are optimized concurrently for several objectives which may serve as a regularizer to generalize unseen data better. Additional information learned in simultaneous tasks helps the model to avoid overfitting. Multi-task learning models have also been proven to avoid sub-optimal solutions compared to a single objective model [7]. Furthermore, MTL overcomes the requirement of a large amount of data by sharing information among the related tasks and learning missing information for related objectives [15, 76]. However, MTL is not always beneficial and may lead to detrimental performance if the relation between the tasks is obscure. Moreover, it increases the training complexity and time. Waseem et al. [106] note that the main task in MTL might not gain performance increase if other auxiliary tasks are highly predictive. Nonetheless, in our framework, we do not recognize any task as a "main task" and give equal weightage to both the objectives. Furthermore, the class label of an instance is highly dependent on the tokens that makes it toxic.

Multi-task architectures are employed in several machine learning fields – computer vision [35, 114], various sub-tasks of NLP [23, 58, 60] and achieved

exceptional results. The MTL architecture is also studied in hate speech and abusive language detection. In [106] researchers found MTL vastly improves the performance of post-level hate speech detection and the model strongly generalizes unseen datasets. Recently, Kapil et al. [50] proposed a deep neural network multi-task learning framework for hate speech detection. In another study, Farha et al. [1] used sentiment prediction as an auxiliary task to detect offensive and hate speech in an MTL setup using a CNN-Bi-LSTM model. Past studies also exhibit the use of neural transformer multi-task learning models that achieve competitive results in shared tasks [26, 29].

To the best of our knowledge, this is the first work exploring multi-task learning for detection of both post and word-level aggression using neural transformers. Furthermore, we propose a robust, open-source **M**ulti-task **A**ggression **D**etection (MAD) framework, that can classify post-level offensiveness and flag the tokens that makes the post toxic with high accuracy.

6.2 Multi-task Aggression Detection Model

Multi-task learning endows transferring information between related tasks that increases performance when the annotated dataset is limited. Sharing information between unrelated tasks may degrade the performance of the model, also known as negative transfer [92]. The two tasks in this research are highly correlated as token-level aggression determines the post-level annotation of the text. Our multi-task learning approach is based on bidirectional neural transformer networks. We described the general transformer network in Section 4.3.3. Hard parameter sharing is implemented by sharing the hidden layers between both post and token-level tasks. The shared part includes a pre-trained transformer language model that learns shared information among the tasks by minimizing combined loss. Furthermore, finding the representation that captures all the sub-tasks considerably reduces the chances of overfitting [92]. The task-specific classifiers receive input from the last hidden layer of the bidirectional transformer language model and predict the output for the tasks.

Post-level Aggression Detection By utilizing the hidden representation of the classification token presented in Equation 5.2 we predict the target

labels (offensive/hate speech/normal) by applying linear transformation with softmax activation.

$$\hat{y}_{post} = softmax(h_{[CLS]} * W_{[CLS]} + b_{[CLS]}) \tag{6.1}$$

Token-level Aggression Detection We predict the token labels (toxic/non-toxic) by applying a similar linear transformation over every input token from the last hidden layer of the model.

$$\hat{y}_{token} = softmax(h * W_{token} + b_{token}) \tag{6.2}$$

Optimization We train the model by minimizing the cross-entropy loss for both tasks as defined in Equation 6.5, where y_{post} and y_{token} represents the true labels. We also introduced α and β parameters to calibrate the importance of the tasks. As we give equal important to both the tasks and in our experiments we considered $\alpha = \beta = 1$.

$$Loss_{post} = -\sum y_{post} * log(\hat{y}_{post}) \tag{6.3}$$

$$Loss_{token} = -\sum\sum y_{token} * log(\hat{y}_{token}) \tag{6.4}$$

$$Loss_{overall} = \frac{\alpha * Loss_{post} + \beta * Loss_{token}}{\alpha + \beta} \tag{6.5}$$

6.3 MAD: Multi-task Aggression Detection Framework

Derived by the success of transformers in the post and token-level aggression detection, we developed an open-source multi-task neural transformer framework for both objectives. The core architecture of the framework is inspired by MUDES [82] and built on top of a transfer learning framework, FARM[1]. The framework has four main components – language modeler, transformer collection, model ensembling, and model tuning. In our experiments, we found

[1]FARM transfer learning framework: `https://farm.deepset.ai`

providing both the objectives equal weightage produces good results. Hence, the loss function of the framework is defined as a mean of individual prediction head losses. The post label classifier predicts the aggression on sentence level in one of the following classes – offensive, hate speech, and normal. Conversely, the token-level classifier takes sequential input from the last hidden state and predicts output for each token. Toxic and non-toxic are two possible outputs of the token-level classifier.

6.3.1 Language Modeler

Re-training language models on masked language modeling objective often help models to capture nuances of unseen datasets and achieve good performance in downstream tasks. In the MAD framework, we include a language modeler that can run MLM on the given dataset and language model. By default, the modeler masks 15% of the tokens randomly in the dataset and considers a maximum sequence length of 512. The model weights can be further stored and loaded in the transformer layer.

6.3.2 Transformer Model

There are multiple variations of neural transformers, e.g., BERT [28], RoBERTa [61], XLNet [109], XLM-R [24]. Our framework is well generalized to support different variations of transformers to analyze the performance of various language models. Furthermore, locally stored models can also be loaded using the framework.

6.3.3 Model Ensembling

Ensembling methods in machine learning is used to create an optimal predictive model by using multiple base models. Model ensembling can be used to make a robust and performant system over a single model [113]. The MAD framework supports model ensembling using both random seeds and changing model hyperparameters. A majority vote strategy is used to get the final predictions. Also, the framework can store individual model results for analysis.

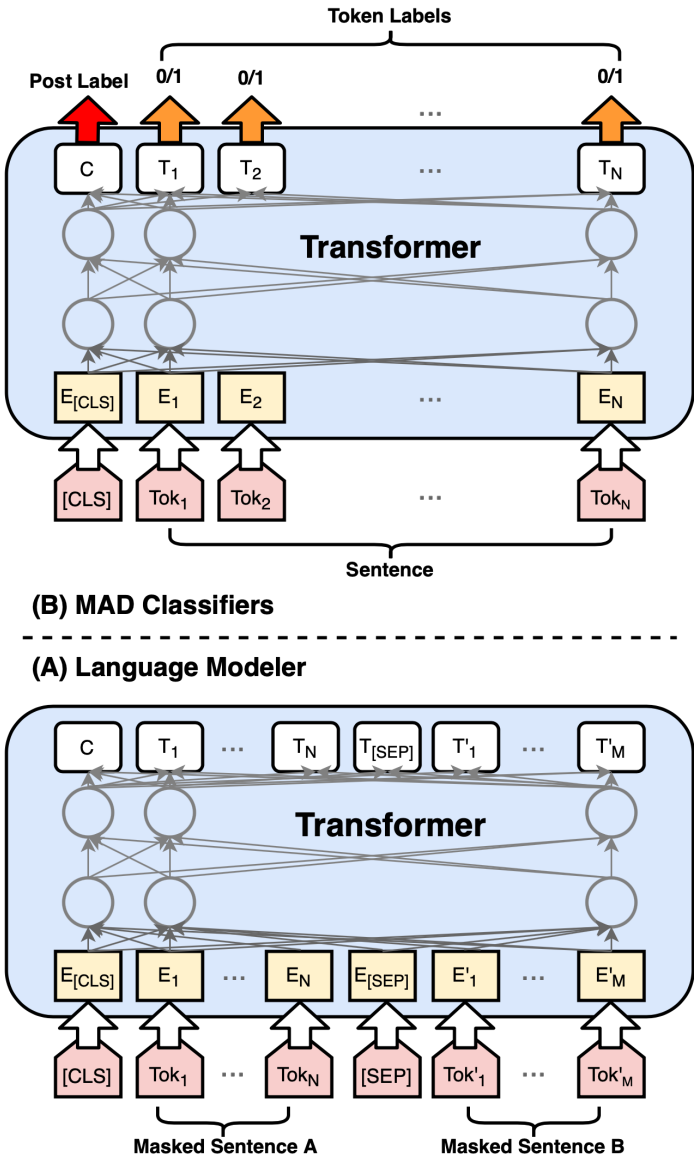


Figure 6.1: The two components of the MAD framework. Section A represents the language modeling part. Section B shows the multi-task aggression detection classifier – the post label predicts post-level aggression (offensive, hate speech, and normal); token label 0 and 1 denotes non-toxic and toxic tokens respectively.

6.3.4 Model Tuning

Tuning deep learning models is significant to achieve good performance. In the framework, we expose a simple JSON-based API for defining model hyperparameters effortlessly. Furthermore, early stopping can be easily combined with the desired number of steps and criteria.

The language modeler and MAD classifier of the framework are shown in Figure 6.1. The code and framework are also publicly available for use on GitHub[2].

6.4 Dataset

Over the years, several offensive and hate speech datasets have been released with post-level annotations [5, 27, 91, 111]. Toxic span is defined as the sequence of words that makes a post or post offensive. The SemEval 2021: Toxic Spans Detection [73] task introduces a new dataset that has toxic span annotations. The recently released HateXplain dataset [66], to the best of our knowledge, is the first benchmark dataset with both post and token-level annotations of hate speech and offensiveness. The dataset is collected from Twitter and Gab, and annotated using Amazon Mechanical Turk[3]. Each instance in the dataset is annotated by three annotators in three categories - *label* (offensive, hate speech, and normal), *rationales* (tokens based on which labeling decision was made), and *target communities* (the group of people that are denounced in the post). The dataset contains $20,148$ posts ($9,055$ from Twitter and $11,093$ from Gab), out of which $5,935$ instances are hateful, $5,480$ are offensive, and $7,814$ are normal. The dataset also contains 919 undecided posts, where all three annotators annotated the label differently.

Dataset Preparation For the task, we used labels and rationales from the HateXplain dataset. A majority vote strategy, where half or more annotators agree on an annotation, is used to determine the final annotation of the label and individual tokens in the rationales. We removed the 919 undecided annotations from the final dataset. The dataset is further split into $11,535$ train, $3,844$ dev, and $3,844$ test sets. The distribution of labels and tokens in the

[2]MAD Framework code: `https://github.com/imdiptanu/MAD`
[3]Amazon Mechanical Turk: `https://www.mturk.com`

Class	Train	Dev	Test
Offensive	3,325	1,061	1,093
Hate speech	3,547	1,185	1,202
Normal	4,663	1,598	1,550
Total	11,535	3,844	3,845

Table 6.1: The distribution of hate speech, offensive, and normal instances in train, dev, and test sets.

Token	Train	Dev	Test
Toxic	22,224	7,493	7,561
Non toxic	248,054	82,622	82,432
Total	270,278	90,115	89,993

Table 6.2: Number of toxic and non toxic tokens in train, dev, and test sets.

Post:	[<user>, keep, running, to, russia, you, nazi, sympathizer]
Rationales:	[0, 0, 0, 0, 0, 0, 1, 0]
Label:	Offensive
Post:	[iron, fist, is, a, nigger, lover]
Rationales:	[0, 0, 0, 0, 1, 1]
Label:	Hate speech
Post:	[expectations, are, a, bitch]
Rationales:	[0, 0, 0, 0]
Label:	Normal
Post:	[yep, communist, nigger, fag]
Rationales:	[0, 1, 1, 1]
Label:	Hate speech

Table 6.3: Four instances from the dataset along with their annotations.

final processed dataset is shown in Tables 6.1 and 6.2. From the figure, we can observe that the train, dev, and test sets follow a similar class imbalanced distribution. Instances from the processed dataset are shown in Table 6.3. The final processed dataset is available on GitHub[4].

[4]Dataset used in MAD framework: `https://github.com/imdiptanu/MAD/Data`

6.5 Methodology

To evaluate our proposed multi-task learning model, we setup two individual tasks – post-level and token-level aggression detection, as baselines. The post-level transformer model takes the complete sentence as an input and predicts the aggression label – normal, offensive, or hate speech; the token-level model predicts each token in the sentence whether the word is toxic or not.

We fine-tuned BERT and RoBERTa transformer models for the two baseline models and a multi-task learning model using a maximum sequence length of 128 and batch size of 16 with ensembling. Early stopping is also executed if the validation loss did not increase over 10 steps. The models are trained using a 16 GB Tesla P100 GPU over three epochs.

It is often beneficial to re-train the language model to better capture the features of the new dataset before performing down-stream tasks. We re-trained a BERT language model for 30 epochs on the training and dev instances of the HateXplain dataset with a maximum sequence length of 512 and 0.15 MLM probability. We refer to the re-trained BERT model as BERT-HTX. Derived by the performance increase across all the objectives, we further re-trained BERT and RoBERTa models using HateXplain [66], HatEval [5], and OLID [111] datasets; the shifted models are denoted by the H_2O suffix. The hyperparameters used for the models are reported in Appendix B.3.

6.6 Evaluation and Results

Given the imbalance in the number of instances in different classes and tokens, we have chosen the macro F_1 score as an evaluation measure across all the objectives. For the post-level evaluation, we used a macro F_1 score that is computed as a mean of per-class F_1 scores, shown in Equation 6.6. If the total number of instances is n, the final aggregated F_1 score A for the token-level task is shown in Equation 6.7.

$$F_1\ Score = \frac{F_1(Offensive) + F_1(Hate\ speech) + F_1(Normal)}{3} \tag{6.6}$$

$$A = \frac{1}{n}\sum_{i=1}^{n} F_1(Per\ Instance) \tag{6.7}$$

Models	Individual		MAD	
	Post-level	*Token-level*	*Post-level*	*Token-level*
BERT	0.6809	0.8106	0.6858	0.8104
BERT-HTX	0.6907	0.8113	0.6913	0.8146
BERT-H_20	0.6923	0.8133	0.6933	0.8167
fBERT	0.6930	0.8134	0.6932	**0.8183**
RoBERTa	0.6847	0.8110	0.6906	0.8120
RoBERTa-H_20	**0.6935**	**0.8139**	**0.6949**	0.8145

Table 6.4: The macro F_1 scores of different transformer models on the test set.

In the result presented in Table 6.4, we observe that only the BERT transformer performs better in individual tasks. The re-trained language models achieve better results than the vanilla model across all the objectives. The fBERT model achieves the overall highest macro F_1 score for the token-level aggression detection using the multi-task framework. Furthermore, the RoBERTa-H_2O model has achieved a macro F_1 score of 0.6949 and 0.8145 in post-level and token-level tasking using the proposed multi-task learning framework. That proves the robustness and efficiency of the proposed multi-task learning framework. From the results, we can conclude that MTL can achieve highly comparable performance to individual setups by sharing information across tasks.

Metric (Avg.)	Individual Models		MAD Model
	Post-level	*Token-level*	
RAM usage (GB)	2.21	3.39	3.21
GPU usage (GB)	6.18	6.55	8.33
Training time per epoch (Sec)	193.68	178.37	184.73
*Inferencing w/ GPU (Sec)	3.64	3.56	4.76
*Inferencing w/ CPU (Sec)	4.24	4.91	5.49

Table 6.5: Performance comparison of two individual models and MAD framework model. *Inferencing 100 instances.

One advantage of multi-task learning is that it learns with fewer data by sharing information across related tasks; hence it reduces the requirements for a large labeled dataset [15, 76]. The two tasks in this study are highly correlated as the rationales motivate the final label of the posts, and the MTL models can learn good relational representations among the tasks. We further

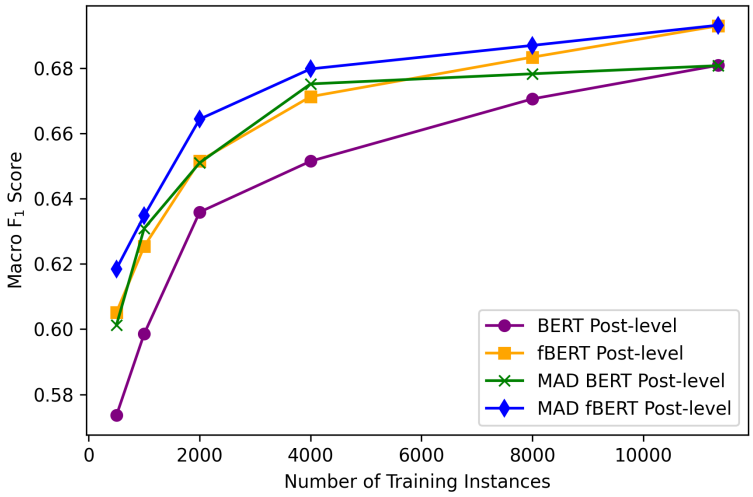


Figure 6.2: The test set F_1 score with an increasing number of training instances using BERT and fBERT models in post-level individual and MAD framework multi-task setup.

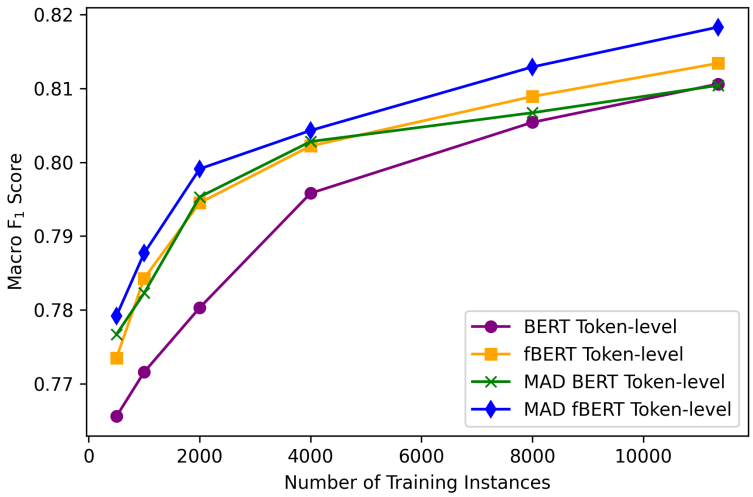


Figure 6.3: The test set F_1 score with an increasing number of training instances using BERT and fBERT models in token-level individual and MAD framework multi-task setup.

compared the MAD framework model performance with individual baseline models when the number of training instances are limited. The plots of the

test set F_1 score and the increasing number of training instances presented in Figures 6.2 and 6.3 show that fBERT constantly performs better than BERT while varying the number of training instances in both post and token-level aggression detection tasks. The result demonstrates the generalizability of multi-task learning even when the labeled dataset is scarce. Furthermore, we can observe that using fBERT in the MAD Framework outperforms all other BERT-based approaches in both tasks. From the figures, we can conclude that the multi-task learning setup performs much better when the training instances are limited. The architecture can be highly effective in aggression detection in low resource languages such as – Bengali, Hindi, Danish.

Furthermore, we compared the performance of individual baseline models with our MAD framework models. The results of the comparison are shown in Table 6.5. We can observe that the MAD framework model outperformed combined individual models in every metric shown. A closer observation shows that the multi-task model uses less RAM than the token-level model and training time per epoch is also less compared to the post-level model. This proves the robustness and efficiency of the MAD framework.

6.7 Discussion

Nowadays, offensive and aggressive content have proliferated on social media platforms. In the past studies, highly performant automatic offensive and hate speech detection models are proposed. In some recent research, more fine-grained token-wise aggression detection is studied. We argue that we can efficiently perform both post and token-level offensive and hate speech detection using a multi-task learning architecture. The high correlation among the label and rationales makes the task suitable for MTL, where the model can learn and share information from related tasks.

In this study, firstly, we have shown that using multi-task neural architecture we can achieve better performance in both sentence and word level aggression detection compared to individual models. We thoroughly examined the effects of two different transformer language models and the effect of re-training the language models. We report that re-training transformers on domain-specific datasets help the language model learn and generalize better. Secondly, our experiments show the extensibility of the proposed MTL architecture with fewer training instances. Third, we propose an open-source

multi-task aggression detection (MAD) framework using neural transformers. The framework supports language modeling and saving the weights to train robust models for the purpose. We further compare the performance of the proposed MAD framework in various metrics and determine that the MAD framework uses fewer resources and time compared to the combined individual models.

Chapter 7

Conclusion

7.1 Conclusion

Over the years, social media platforms have gained popularity across a large number of users. People use these platforms to connect to friends and family, share opinions, and news. However, the spread of offensive and hateful content on social media is a longstanding problem. The sense of anonymity on the internet further drives aggressive behavior. In some countries, there is evidence of social media sparked violence against certain communities [6, 102]. Government institutions and companies are trying to keep the platform clean and welcoming to everyone. Past researches proposed automatic aggression detection using rule-based techniques to current state-of-the-art pre-trained transformer models. However, supervised detection of offensive language is dependent on human-annotated datasets. Due to the subjective nature of the annotation achieving high consensus among annotators is difficult and varies by annotator's culture, geographies, background.

Dataset exploration aids the understanding of data by finding patterns, relations, and anomalies. In the first exploratory work, we applied statistical and neural topic modeling to extract topic-word distributions from three highly popular offensive language and hate speech datasets. We found that the most prominent topics in the dataset denote the keywords applied to compile the dataset. Also, all three datasets inherently concentrate on three different themes. We further showed that the top features from a supervised learning model exhibit the same results as unsupervised learning.

In the past, several benchmark datasets for offensive and hate speech detection were released. However, the primary focus on these datasets is post or document-level aggression detection. Current state-of-the-art models are limited to post-level toxicity detection and lack token or word-level indications. The problem of fine-grain aggression detection is also presented in SemEval 2021 Task 5: Toxic Spans Detection [73]. We experimented with lexicon-based word match, recurrent neural networks, and neural transformers. Word matching algorithm achieved balanced results and served as a baseline performance for evaluation. The RoBERTa language model outperformed other approaches and achieved the highest F_1 score of 0.68. We also submitted our system in the SemEval 2021 competition [85]. The model can aid human annotators while reviewing lengthy flagged posts or comments on social media.

Neural transformers have outperformed previous state-of-the-art models in various NLP sub-tasks. The transformer models are trained on general corpora such as Wikipedia, BookCorpurs and lack domain-specific information. Recently, retrained transformer models are proliferating to solve domain specific problems. In this study, we propose a shifted BERT model fBERT adapted to social media aggression detection tasks. The fBERT is trained on over 1.4 million offensive tweets collected from the SOLID dataset. We proved the model's effectiveness and portability across various shared tasks and offensive language identification. We firmly believe that the model will boost research on social media offensive language and hate speech detection tasks.

Even though there are individual models for automatic post and token-level aggression detection using deep learning models. To best our knowledge, we proposed the first multi-task learning approach for both document and token-level aggression detection at the same time using neural transformers. We evaluated our model using the recently released HateXplain benchmark dataset. We concluded that the MTL approach could achieve similar or better performance over individual models, specifically when the training resources are limited. We further proposed the first publicly available robust multi-task aggression detection (MAD) framework with transformers. The performance of our MTL approach will serve as a baseline for future works in the domain.

In this study, we first explored various topics and features of popular datasets in the offensive language identification domain. We proposed a robust token or word-level aggression detection model using transformers to help human moderators while reviewing flagged social media posts. We adapted a

BERT model for social media aggression detection tasks and demonstrated its effectiveness across various datasets. We publicly released the fBERT model to encourage further research on social media aggression detection. We took a multi-task learning approach for both post and token-level aggression detection using neural transformers that may serve as a baseline for future study in the field. Furthermore, we open-sourced a highly robust multi-task aggression detection framework for public use.

7.2 Future Work

In terms of future work, we would like to experiment with neural topic models – neural variational inference [67], BERTopic [40] models. Even though the HateXplain dataset presents both post and token-level aggression annotations, the dataset does not incorporate emojis. As emoji are highly used in social media posts to express emotions, we would like to observe the performance variations with emojis. The hierarchical annotation scheme presented in OLID [111] is highly effective in explainable hate speech detection. We would like to see the incorporation of the annotation scheme in other datasets as well.

Inspired by the performance gain achieved by offensive language-specific fBERT, we further like to extend the work by proposing other shifted transformer models. We would like to evaluate the proposed multi-task architectures on multi-domain and multilingual settings. Furthermore, we are interested in experimentation with an added layer of CNN or RNN on top transformer models proposed in the study.

Bibliography

[1] Ibrahim Abu Farha and Walid Magdy. Multitask learning for Arabic offensive language and hate-speech detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 86–90, Marseille, France, May 2020. European Language Resource Association. ISBN 979-10-95546-51-1. URL https://www.aclweb.org/anthology/2020.osact-1.14.

[2] Swati Agarwal and Ashish Sureka. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In Raja Natarajan, Gautam Barua, and Manas Ranjan Patra, editors, *Distributed Computing and Internet Technology*, pages 431–442, Cham, 2015. Springer International Publishing. ISBN 978-3-319-14977-6.

[3] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *CoRR*, abs/1908.10063, 2019. URL http://arxiv.org/abs/1908.10063.

[4] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 759–760, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349147. doi: 10.1145/3041021.3054223. URL https://doi.org/10.1145/3041021.3054223.

[5] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela

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*, pages 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2007. URL `https://www.aclweb.org/anthology/S19-2007`.

[6] Nick Beake. Facebook admits it was used to 'incite offline violence' in myanmar, Nov 2018. URL `https://www.bbc.com/news/world-asia-46105934`. BBC News, Accessed: 25-April-2021.

[7] Joachim Bingel and Anders Søgaard. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169, Valencia, Spain, April 2017. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/E17-2026`.

[8] Akanksha Bisht, Annapurna Singh, H. S. Bhadauria, Jitendra Virmani, and Kriti. *Detection of Hate Speech and Offensive Language in Twitter Data Using LSTM Model*, pages 243–264. Springer Singapore, Singapore, 2020. ISBN 978-981-15-2740-1. doi: 10.1007/978-981-15-2740-1_17. URL `https://doi.org/10.1007/978-981-15-2740-1_17`.

[9] David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4): 77–84, April 2012. ISSN 0001-0782. doi: 10.1145/2133806.2133826. URL `https://doi.org/10.1145/2133806.2133826`.

[10] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, March 2003. ISSN 1532-4435.

[11] Mikael Boden. A guide to recurrent neural networks and backpropagation. *the Dallas project*, 2002.

[12] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561, 2019. URL `http://arxiv.org/abs/1903.04561`.

[13] Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2):223–242, 2015.

[14] L. Cao and Li Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007. doi: 10.1109/ICCV.2007.4408965.

[15] Rich Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, July 1997. ISSN 1573-0565. doi: 10.1023/A:1007379606734. URL `https://doi.org/10.1023/A:1007379606734`.

[16] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. Hatebert: Retraining bert for abusive language detection in english, 2021.

[17] Çağrı Çöltekin. A Corpus of Turkish Offensive Language on Social Media. In *Proceedings of LREC*, 2020.

[18] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school, 2020.

[19] Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*, 2009. URL `http://umiacs.umd.edu/~jbg//docs/nips2009-rtl.pdf`.

[20] Irfan Chaudhry. #hashtagging hate: Using twitter to track racism online. *First Monday*, 20(2), Feb. 2015. doi: 10.5210/fm.v20i2.5450. URL `https://journals.uic.edu/ojs/index.php/fm/article/view/5450`.

[21] Lu Cheng, Jundong Li, Yasin N. Silva, Deborah L. Hall, and Huan Liu. Xbully: Cyberbullying detection within a multi-modal context. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 339–347, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359405. doi:

10.1145/3289600.3291037. URL https://doi.org/10.1145/3289600.3291037.

[22] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014. URL http://arxiv.org/abs/1406.1078.

[23] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 160–167, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390177. URL https://doi.org/10.1145/1390156.1390177.

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

[25] Savelie Cornegruta, Robert Bakewell, Samuel Withey, and Giovanni Montana. Modelling radiological language with bidirectional long short-term memory networks. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 17–27, Auxtin, TX, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-6103. URL https://www.aclweb.org/anthology/W16-6103.

[26] Wenliang Dai, Tiezheng Yu, Zihan Liu, and Pascale Fung. Kungfupanda at SemEval-2020 task 12: BERT-based multi-TaskLearning for offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2060–2066, Barcelona (online), December 2020. International Committee for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.semeval-1.272.

[27] Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive

language. *CoRR*, abs/1703.04009, 2017. URL `http://arxiv.org/abs/1703.04009`.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://www.aclweb.org/anthology/N19-1423`.

[29] Marc Djandji, Fady Baly, Wissam Antoun, and Hazem Hajj. Multi-task learning using AraBert for offensive language detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 97–101, Marseille, France, May 2020. European Language Resource Association. ISBN 979-10-95546-51-1. URL `https://www.aclweb.org/anthology/2020.osact-1.16`.

[30] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 29–30, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450334730. doi: 10.1145/2740908.2742760. URL `https://doi.org/10.1145/2740908.2742760`.

[31] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4), July 2018. ISSN 0360-0300. doi: 10.1145/3232676. URL `https://doi.org/10.1145/3232676`.

[32] Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. A Hierarchically-labeled Portuguese Hate Speech Dataset. In *Proceedings of ALW*, 2019.

[33] Björn Gambäck and Utpal Kumar Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop*

on Abusive Language Online, pages 85–90, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3013. URL https://www.aclweb.org/anthology/W17-3013.

[34] Norjihan Abdul Ghani, Suraya Hamid, Ibrahim Abaker Targio Hashem, and Ejaz Ahmed. Social media big data analytics: A survey. *Computers in Human Behavior*, 101:417 – 428, 2019. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2018.08.039. URL http://www.sciencedirect.com/science/article/pii/S074756321830414X.

[35] Ross Girshick. Fast r-cnn. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 1440–1448, USA, 2015. IEEE Computer Society. ISBN 9781467383912. doi: 10.1109/ICCV.2015.169. URL https://doi.org/10.1109/ICCV.2015.169.

[36] Bharti Goel and Ravi Sharma. USF at SemEval-2019 task 6: Offensive language detection using LSTM with word embeddings. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 796–800, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2139. URL https://www.aclweb.org/anthology/S19-2139.

[37] Alex Graves and Jürgen Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2008/file/66368270ffd51418ec58bd793f2d9b1b-Paper.pdf.

[38] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2005.06.042. URL https://www.sciencedirect.com/science/article/pii/S0893608005001206. IJCNN 2005.

[39] Alex Graves, Abdel-rahman Mohamed, and Geoffrey E. Hinton. Speech recognition with deep recurrent neural networks. *CoRR*, abs/1303.5778, 2013. URL http://arxiv.org/abs/1303.5778.

[40] Maarten Grootendorst. Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics., 2020. URL `https://doi.org/10.5281/zenodo.4381785`.

[41] Hansi Hettiarachchi and Tharindu Ranasinghe. Emoji powered capsule network to detect type and target of offensive posts in social media. In *Proceedings of RANLP*, 2019.

[42] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[43] Thomas Hoffman. Probabilistic latent semantic analysis. In *proc. of the 15th Conference on Uncertainty in AI, 1999*, 1999.

[44] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020. URL `https://doi.org/10.5281/zenodo.1212303`.

[45] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1031. URL `https://www.aclweb.org/anthology/P18-1031`.

[46] Weipeng Huang, Xingyi Cheng, Taifeng Wang, and Wei Chu. BERT-Based Multi-Head Selection for Joint Entity-Relation Extraction. In *Proceedings of NLPCC*, 2019.

[47] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[48] Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 70–74, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2009. URL `https://www.aclweb.org/anthology/S19-2009`.

[49] Cecilia Kang and Adam Satariano. Regulators around the world are circling facebook, Apr 2019. URL `https://www.nytimes.com/2019/04/25/technology/facebook-regulation-ftc-fine.html`. The New York Times, Accessed: 25-April-2021.

[50] Prashant Kapil and Asif Ekbal. A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems*, 210:106458, 2020. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2020.106458. URL `https://www.sciencedirect.com/science/article/pii/S0950705120305876`.

[51] Rohan Kshirsagar, Tyus Cukuvac, Kathleen R. McKeown, and Susan McGregor. Predictive embeddings for hate speech detection on twitter. *CoRR*, abs/1809.10644, 2018. URL `http://arxiv.org/abs/1809.10644`.

[52] Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W18-4401`.

[53] Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. Evaluating Aggression Identification in Social Media. In *Proceedings of TRAC*, 2020.

[54] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, 2001.

[55] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1082. URL `https://www.aclweb.org/anthology/D17-1082`.

[56] Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

[57] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL `https://doi.org/10.1093/bioinformatics/btz682`.

[58] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1001. URL `https://www.aclweb.org/anthology/P17-1001`.

[59] Ping Liu, Wen Li, and Liang Zou. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2011. URL `https://www.aclweb.org/anthology/S19-2011`.

[60] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1441. URL `https://www.aclweb.org/anthology/P19-1441`.

[61] Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

[62] Jouni Luoma and Sampo Pyysalo. Exploring Cross-sentence Contexts

for Named Entity Recognition with BERT. In *Proceedings of COLING*, 2020.

[63] Shervin Malmasi and Marcos Zampieri. Detecting hate speech in social media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 467–472, Varna, Bulgaria, September 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6_062. URL `https://doi.org/10.26615/978-954-452-049-6_062`.

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

[65] Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of FIRE*, 2019.

[66] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289*, 2020.

[67] Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2410–2419. JMLR.org, 2017.

[68] Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. Neural character-based composition models for abuse detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 1–10, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5101. URL `https://www.aclweb.org/anthology/W18-5101`.

[69] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. A BERT-based transfer learning approach for hate speech detection in online

social media. In *Complex Networks 2019: 8th International Conference on Complex Networks and their Applications*, volume Studies in Computational Intelligence book series (SCI, volume 881) of *Complex Networks and Their Applications VIII : Volume 1, Proceedings of the Eighth International Conference on Complex Networks and Their Applications*, pages 928–940, Lisbonne, Portugal, December 2019. Springer. doi: 10.1007/978-3-030-36687-2_77. URL `https://hal.archives-ouvertes.fr/hal-02344806`.

[70] Hamdy Mubarak, Kareem Darwish, and Walid Magdy. Abusive language detection on Arabic social media. In *Proceedings of ALW*, 2017.

[71] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.2. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.2`.

[72] Dennis Njagi, Z. Zuping, Damien Hanyurwimfura, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10:215–230, 04 2015. doi: 10.14257/ijmue.2015.10.4.21.

[73] John Pavlopoulos, Léo Laugier, Jeffrey Sorensen, and Ion Androutsopoulos. Semeval-2021 task 5: Toxic spans detection. In *Proceedings of SemEval*, 2021.

[74] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL `https://www.aclweb.org/anthology/N18-1202`.

[75] Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. Offensive Language Identification in Greek. In *Proceedings of LREC*, 2020.

[76] M. Qiu, P. Zhao, K. Zhang, J. Huang, X. Shi, X. Wang, and W. Chu. A short-term rainfall prediction model using multi-task convolutional neural networks. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 395–404, 2017. doi: 10.1109/ICDM.2017.49.

[77] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL `https://www.aclweb.org/anthology/D16-1264`.

[78] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124. URL `https://www.aclweb.org/anthology/P18-2124`.

[79] Kshitij Rajput, Raghav Kapoor, Puneet Mathur, Hitkul, Ponnurangam Kumaraguru, and Rajiv Ratn Shah. *Transfer Learning for Detecting Hateful Sentiments in Code Switched Language*, pages 159–192. Springer Singapore, Singapore, 2020. ISBN 978-981-15-1216-2. doi: 10.1007/978-981-15-1216-2_7. URL `https://doi.org/10.1007/978-981-15-1216-2_7`.

[80] Tharindu Ranasinghe and Hansi Hettiarachchi. BRUMS at SemEval-2020 task 12: Transformer based multilingual offensive language identification in social media. In *Proceedings of SemEval*, 2020.

[81] Tharindu Ranasinghe and Marcos Zampieri. Multilingual Offensive Language Identification with Cross-lingual Embeddings. In *Proceedings of EMNLP*, 2020.

[82] Tharindu Ranasinghe and Marcos Zampieri. Mudes: Multilingual detection of offensive spans, 2021.

[83] Tharindu Ranasinghe, Marcos Zampieri, and Hansi Hettiarachchi.

Brums at hasoc 2019: Deep learning models for multilingual hate speech and offensive language identification. In *Proceedings of FIRE*, 2019.

[84] Tharindu Ranasinghe, Sarthak Gupte, Marcos Zampieri, and Ifeoma Nwogu. Wlv-rit at hasoc-dravidian-codemix-fire2020: Offensive language identification in code-switched youtube comments. In *Proceedings of FIRE*, 2020.

[85] Tharindu Ranasinghe, Diptanu Sarkar, Marcos Zampieri, and Alex Ororbia. Wlv-rit at semeval-2021 task 5: A neural transformer framework for detecting toxic spans, 2021.

[86] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

[87] Michael Ridenhour, Arunkumar Bagavathi, Elaheh Raisi, and Siddharth Krishnan. Detecting Online Hate Speech: Approaches Using Weak Supervision and Network Embedding Models. In *Proceedings of SBP-BRiMS*, 2020.

[88] Marian-Andrei Rizoiu, Tianyu Wang, Gabriela Ferraro, and Hanna Suominen. Transfer learning for hate speech detection in social media. *CoRR*, abs/1906.03829, 2019. URL `http://arxiv.org/abs/1906.03829`.

[89] Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. Hate speech detection in the bengali language: A dataset and its baseline evaluation. *arXiv preprint arXiv:2012.09686*, 2020.

[90] Hugo Rosa, N Pereira, Ricardo Ribeiro, Paula Costa Ferreira, Joao Paulo Carvalho, S Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345, 2019.

[91] Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. A large-scale semi-supervised dataset for offensive language identification. *arXiv preprint arXiv:2004.14454*, 2020.

[92] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098, 2017. URL `http://arxiv.org/abs/1706.05098`.

[93] Samuel Rönnqvist. Doc2topic: Neural topic modeling, January 2018. URL `https://github.com/sronnqvist/doc2topic`. GitHub, Accessed: 25-April-2021.

[94] Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131, Marseille, France, May 2020. European Language Resources Association (ELRA). ISBN 979-10-95546-56-6. URL `https://www.aclweb.org/anthology/2020.trac-1.20`.

[95] Gudbjartur Ingi Sigurbergsson and Leon Derczynski. Offensive Language and Hate Speech Detection for Danish. In *Proceedings of LREC*, 2020.

[96] Leandro Araújo Silva, Mainack Mondal, D. Correa, F. Benevenuto, and Ingmar Weber. Analyzing the targets of hate in online social media. *ArXiv*, abs/1603.07709, 2016.

[97] Chang-Woo Song, Hoill Jung, and Kyungyong Chung. Development of a medical big-data mining process using topic modeling. *Cluster Computing*, 22(1):1949–1958, 2019.

[98] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Portuguese Named Entity Recognition using BERT-CRF. *arXiv preprint arXiv:1909.10649*, 2020.

[99] S. Syed and M. Spruit. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 165–174, 2017. doi: 10.1109/DSAA.2017.61.

[100] Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. A Dictionary-based Approach to Racism Detection in Dutch Social Media. In *Proceedings of TA-COS*, 2016.

[101] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL `http://arxiv.org/abs/1706.03762`.

[102] Jane Wakefield. Sri lanka attacks: The ban on social media, Apr 2019. URL `https://www.bbc.com/news/technology-48022530`. BBC News, Accessed: 25-April-2021.

[103] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL `https://www.aclweb.org/anthology/W18-5446`.

[104] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada, June 2012. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W12-2103`.

[105] Zeerak Waseem and Dirk Hovy. 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, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-2013. URL `https://www.aclweb.org/anthology/N16-2013`.

[106] Zeerak Waseem, James Thorne, and Joachim Bingel. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In Jennifer Golbeck, editor, *Online Harassment*, pages 29–55, Cham, 2018. Springer International Publishing. ISBN 978-3-319-78583-7. doi: 10.1007/978-3-319-78583-7_3. URL `https://doi.org/10.1007/978-3-319-78583-7_3`.

[107] Gregor Wiedemann, Seid Muhie Yimam, and Chris Biemann. UHH-LT at SemEval-2020 task 12: Fine-tuning of pre-trained transformer net-

works for offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1638–1644, Barcelona (online), December 2020. International Committee for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.semeval-1.213`.

[108] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL `http://arxiv.org/abs/1609.08144`.

[109] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Proceedings of NeurIPS*, 2019.

[110] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of SemEval*, 2019.

[111] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media. 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 1415–1420, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1144. URL `https://www.aclweb.org/anthology/N19-1144`.

[112] Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis,

and Çağrı Çöltekin. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*, 2020.

[113] Cha Zhang and Yunqian Ma. *Ensemble Machine Learning*. Springer, 2012.

[114] Xiangyun Zhao, Haoxiang Li, Xiaohui Shen, Xiaodan Liang, and Ying Wu. A modulation module for multi-task learning with applications in image retrieval. In Martial Hebert, Vittorio Ferrari, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018 - 15th European Conference, 2018, Proceedings*, pages 415–432. Springer Verlag, 2018. ISBN 9783030012458. doi: 10.1007/978-3-030-01246-5_25.

[115] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Appendices

Appendix A

Topic Modeling Results

OLID - LDA	
Offensive Instances	
Topic-1:	liberal, gun, control, people, conservative, trump, law, make, democrats, hate
Topic-2:	shit, antifa, maga, fuck, ass, say, bad, man, never, fucking
Normal Instances	
Topic-1:	gun, control, law, state, bill, government, school, shoot, criminal, brexit
Topic-2:	liberal, antifa, maga, conservative, right, go, people, trump, vote, think

Table A.1: LDA topic modeling topics to top 10 words distribution for OLID.

OLID - doc2topic	
Offensive Instances	
Topic-1:	gun, shit, ass, fuck, people, antifa, fucking, bitch, know, get
Topic-2:	liberal, trump, antifa, maga, like, make, ask, conservative, right, people
Normal Instances	
Topic-1:	gun, get, like, people, control, know, think, well, take, love
Topic-2:	liberal, control, maga, antifa, think, good, conservative, need, make, see

Table A.2: Neural topic modeling (doc2topic) topics to top 10 words distribution for OLID.

HatEval - LDA	
Offensive Instances	
Topic-1:	bitch, women, whore, hoe, cunt, get, fuck, ass, rape, girl
Topic-2:	illegal, refugee, buildthatwall, immigration, migrant, country, trump, maga, go, people
Normal Instances	
Topic-1:	woman, men, rape, bitch, hysterical, like, get, whore, fuck, kitchen
Topic-2:	refugee, migrant, immigrant, trump, child, work, home, country, via, new

Table A.3: LDA topic modeling topics to top 10 words distribution for HatEval dataset.

HatEval - doc2topic	
Offensive Instances	
Topic-1:	hoe, get, bitch, ass, women, whore, girl, skank, stupid, cunt
Topic-2:	illegal, buildthatwall, refugee, go, immigration, fuck, want, trump , need, like
Normal Instances	
Topic-1:	women, men, rape, say, people, hysterical, cunt, right, whore, girl
Topic-2:	immigrant, migrant, say, men, immigration, refugee, child, like, trump, cunt

Table A.4: Neural topic modeling (doc2topic) topics to top 10 words distribution for HatEval Dataset.

Davidson Dataset - LDA	
Offensive Instances	
Topic-1:	bitch, hoe, get, nigga, like, ass, shit, fuck, know, lol
Topic-2:	trash, faggot, white, right, cunt, go, say, people, love, kill
Normal Instances	
Topic-1:	bird, like, trash, yellow, lol, make, yankee, charlie, monkey, colored

Table A.5: LDA topic modeling topics to top 10 words distribution for Davidson dataset.

Davidson Dataset - doc2topic	
Offensive Instances	
Topic-1:	bitch, like, hoe, get, nigga, fuck, shit, pussy, ass, lol
Topic-2:	get, go, bad, love, say, trash, know, good, look, niggas
Normal Instances	
Topic-1:	trash, bird, charlie, like, make, brownie, ghetto, good, one, monkey

Table A.6: Neural topic modeling (doc2topic) topics to top 10 words distribution for Davidson Dataset.

Appendix B

Hyperparameters

Hyperparameter	**BERT**	**RoBERTa**
Maximum Sequence Length	400	400
Learning Rate	2e-5	2e-5
Number of Epochs	3	3
Batch Size	16	8
Embedding Dropout Probability	0.1	0.1
Early Stopping	Yes	Yes
Model Ensembling	Yes	Yes

Table B.1: Hyperparameters for BERT and RoBERTa models presented in Section 4.3.3

Hyperparameter	**HE**	**OE**	**HSO**	**TSD**
Maximum Sequence Length	128	128	128	350
Learning Rate	3e-5	2e-5	3e-5	2e-5
Number of Epochs	3	3	4	3
Batch Size	16	8	32	16
Embedding Dropout Probability	0.1	0.1	0.1	0.1
Early Stopping	Yes	Yes	Yes	Yes
Model Ensembling	Yes	Yes	Yes	Yes

Table B.2: Hyperparameters for shared task performance comparison in Section 5.4. HE: HatEval, OE: OffensEval, HSO: Hate Speech and Offensive Language Detection, TSD: Toxic Spans Detection.

Hyperparameter	**Individual**		**MAD**
	Post-level	*Token-level*	
Maximum Sequence Length	128	128	128
Learning Rate	3e-5	2e-5	2e-5
Number of Epochs	3	3	4
Batch Size	32	32	32
Embedding Dropout Probability	0.1	0.1	0.1
Early Stopping	Yes	Yes	Yes
Model Ensembling	Yes	Yes	Yes

Table B.3: Hyperparameters of the models shown in Table 6.4