Natural Language Engineering (2023), **29**, pp. 1247–1274
doi:10.1017/S1351324922000262

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Automated hate speech detection and span extraction in underground hacking and extremist forums

Linda Zhou*, Andrew Caines, Ildiko Pete and Alice Hutchings

Department of Computer Science and Technology, University of Cambridge, Cambridge CB2 1TN, UK
*Corresponding author. Email: lz423@cantab.ac.uk

(Received 30 August 2021; revised 19 January 2022; accepted 19 May 2022; first published online 20 June 2022)

Abstract

Hate speech is any kind of communication that attacks a person or a group based on their characteristics, such as gender, religion and race. Due to the availability of online platforms where people can express their (hateful) opinions, the amount of hate speech is steadily increasing that often leads to offline hate crimes. This paper focuses on understanding and detecting hate speech in underground hacking and extremist forums where cybercriminals and extremists, respectively, communicate with each other, and some of them are associated with criminal activity. Moreover, due to the lengthy posts, it would be beneficial to identify the specific span of text containing hateful content in order to assist site moderators with the removal of hate speech. This paper describes a hate speech dataset composed of posts extracted from HackForums, an online hacking forum, and Stormfront and Incels.co, two extremist forums. We combined our dataset with a Twitter hate speech dataset to train a multi-platform classifier. Our evaluation shows that a classifier trained on multiple sources of data does not always improve the performance compared to a mono-platform classifier. Finally, this is the first work on extracting hate speech spans from longer texts. The paper fine-tunes BERT (Bidirectional Encoder Representations from Transformers) and adopts two approaches – span prediction and sequence labelling. Both approaches successfully extract hateful spans and achieve an F1-score of at least 69%.

Keywords: Corpus annotation; Text classification; Span extraction

1. Introduction

Hate speech is any kind of communication, writing or behaviour, that attacks a person or a group based on their identity factors, for example, religion, race and gender (UN 2020). The rise in popularity of social media has, unintentionally, promoted the spread of hate. People have the ability to freely publish hate speech on blogs and social media and through their words influence or harm millions of people all over the world. Even if the site is moderated, some damage may have already been done by the time the hateful content is removed. A study has revealed that posts from hateful users spread faster than those from normal users (Mathew *et al.* 2019). Research also showed that online hate speech leads to hate crime in the physical world. For instance, Williams *et al.* (2020) reported a positive correlation between Twitter hate speech and offline crimes in London.

Given the importance of the problem, tackling hate speech has become the main target of many studies. Various studies have focused on introducing datasets (Waseem and Hovy 2016; Basile *et al.* 2019) and automatically detecting hateful content. The latter relies on machine learning techniques that range from logistic regression (Davidson *et al.* 2017), support vector machines (de Gibert *et al.* 2018), naive Bayes (Kwok and Wang 2013) to deep neural networks (Badjatiya *et al.* 2017). The majority of the work investigated hate speech on mainstream social media, such as

Twitter (Davidson *et al.* 2017; Basile *et al.* 2019) and Facebook (Mandl *et al.* 2019; Vu *et al.* 2020), whereas little has so far been done for underground and extremist forums.

Underground hacking forums enable cybercriminals to share their cybercriminal interest and knowledge and trade illicit materials (Pastrana *et al.* 2018a). Extremist forums serve as a place for people to spread hateful and extremist ideologies with little hindrance to the general public (Schafer 2002). Despite the communities of the two types of forums have different characteristics, we believe they are interesting communities with commonalities (e.g., relatively light moderation by administrators Caines *et al.* 2018a; Jaki *et al.* 2019) and differences (a focus on hacking vs. a focus on political-social issues) for the study of hate speech.

Some members of these forums are associated with criminal activity. An active member of an underground forum called HackForums was arrested for being the alleged author of malware designed to steal online banking credentials (Krebs 2017). In addition to hacking-related activities, there is online aggressive behaviour among underground forum members (Caines *et al.* 2018a). In contrast, members of extremist forums are more likely to be involved in real-world violence (Holpuch 2014; Jasser, Kelly, and Rothermel 2020) that may be driven by the potential prevalence of hate speech in the forums. Because of this, understanding the content of hate speech in these forums and automatically detecting them would help to design early intervention techniques.

Despite a large number of studies on automatic hate speech detection, there are still many limitations. Classifiers do not generalise well on unseen data (Bruwaene, Huang, and Inkpen 2020) and may exhibit bias. For instance, tweets written in African-American English are likely to be classified as more abusive compared to those written in standard American English (Davidson, Bhattacharya, and Weber 2019). Because of these, real-world applications (e.g., Twitter) rely on human moderators to review posts (Harrison 2019). Identifying the hateful text span (i.e., the text fragment containing hate speech) would be a crucial step towards semi-automated moderation because it can assist human moderators who deal with lengthy texts. Moreover, it would benefit researchers who want to analyse certain aspects of hate speech (e.g., target analysis) and are less interested in other parts of the text. To the best of our knowledge, there has not yet been any research focusing on hate speech span extraction (SE).

1.1 Contributions

Our research mainly focuses on understanding hate speech in underground and extremist forums and two tasks, namely hate speech detection and SE. Our contributions are as follows.

(1) **Release of hate speech annotations for a sample of posts from underground and extremist forums composed of posts from underground and extremist forums.**
The paper describes hate speech annotations for posts extracted from HackForums, an underground forum and Stormfront and Incels.co, two extremist forums. These posts come from the pre-existing CrimeBB and ExtremeBB databases (Pastrana *et al.* 2018a; Vu *et al.* 2021). Based on the sampled data, the paper also analyses the frequency and the content of hate speech across these forums. The results show a lower occurrence of hate speech in underground forums compared to the two extremist forums. While Hackforums users do not have a specific target group, the main targets on Stormfront and Incels.co, respectively, are Jews and women.

(2) **Exploration of multi-platform classifiers trained on the combined data from the underground, extremist forums and Twitter.**
We experimented with different classifiers, including CNN-GRU, BERT (Bidirectional Encoder Representations from Transformers), support vector classifier (SVC), trained on the combined data from the underground, extremist forums and Twitter. The multi-platform classifiers are evaluated against the classifiers trained on the Twitter dataset only.

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

The results showed that training classifiers on multiple sources of data may not always outperform mono-platform classifiers.

(3) **Extraction of hate speech spans from longer texts.**
A novel contribution of this work is to attempt to automatically extract hateful spans. We propose to fine-tune BERT and adopt two approaches. One is based on span prediction and the other on sequence labelling. The first predicts the start and end indices of the hateful spans and the second identifies whether each word token is part of the hateful spans. Both models achieve an F1-score of at least 69%.

1.2 Outline

This paper is organised as follows. In the next section, we review prior works related to this area of research. Section 3 provides an overview of the datasets that we use. Section 4 describes the annotation process and analyses the frequency and the content of hate speech in underground and extremist forums. Section 5 focuses on the design choice of the models for hate speech detection and SE. We evaluate the performance of the systems in Section 6 and perform a qualitative analysis and discuss the broader implications of the systems in Section 7. In our conclusion, we outline ideas for future work.

2. Related work

2.1 Hate speech

According to the UN (2020), hate speech is 'any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor'. However, this is not a universal legal definition for hate speech (UN 2020).

Previous work used a similar definition to annotated datasets for automated hate speech detection. However, because the characterisation of what is hateful is still fairly generic and definitions vary, there may be inconsistencies among the annotated datasets in terms of how labels are applied to texts (Waseem *et al.* 2017). One of the key controversies is the confusion between hate speech and offensive language (e.g., disparaging terms and racial epithets). Waseem and Hovy (2016) considered offensive language as a subset of hate speech because they believed the speakers intentionally insult a (member of a) minority group. Contrarily, Assimakopoulos *et al.* (2020) stated: 'online hate speech might often contain offensive language, but not all offensive language can be considered hate speech'.

Mandl *et al.* (2019) distinguished offensive language and hate speech based on the target. Texts targeting individuals are offensive, whereas those targeting a group of people are hate speech. However, this contradicts many hate speech definitions that consider both individuals and groups to be targets of hate speech (Waseem and Hovy 2016; Basile *et al.* 2019; UN 2020). In contrast, Vu *et al.* (2020) claimed that texts are offensive rather than hate speech if they do not target individuals or groups based on their identity factors. Nonetheless, this is not sufficient for drawing a boundary between offensive language and hate speech because one can hardly separate pejorative words from the identity factors, for example, 'bitch' is a female-referent slur, 'negro' refers to Black people (Kleinman, Ezzell, and Frost Kleinman *et al.* 2009). Other scholars, instead, took into account the context in which the pejorative words are used (Warner and Hirschberg 2012; Davidson *et al.* 2017; de Gibert *et al.* 2018). They considered the following cases to be not hateful: (1) the speaker belongs to the target group (Warner and Hirschberg 2012; Davidson *et al.* 2017), and (2) the offensive word does not contain a deliberate attack (de Gibert *et al.* 2018).

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

Other matters of debate include whether to consider hate speech the praise of an organisation associated with hate crimes, the support of hateful hashtags (Warner and Hirschberg 2012; Waseem and Hovy 2016; Basile *et al.* 2019) and the expression of excessive pride in the speaker's own race or group (Warner and Hirschberg 2012; Basile *et al.* 2019).

Hate speech has been positioned as a subtype of online toxicity more generally (Salminen *et al.* 2020), where other types can include offensive language such as swearwords, or arguments which contain neither hate speech or offensive language. Therefore, we are interested in detecting a kind of internet toxicity, which overlaps with offensive language but is not exactly the same as it.

The problem of hate speech annotation has been shown to be a highly demanding task due to the lack of a universal detailed definition for hate speech. For this work, we adopt the UN definition that is enriched with the following details that some researchers have used to clarify what constitutes hate speech:

- The association of stereotypes to a race or ethnicity (Warner and Hirschberg 2012; Waseem and Hovy 2016; Basile *et al.* 2019).
- Language that seeks to silence a minority (Waseem and Hovy 2016).
- Language that dehumanises and degrades an individual based on their belonging to a group (Davidson *et al.* 2017).
- Language that incites or promotes hate or violence (Davidson *et al.* 2017).
- Defence of hateful content, for instance, xenophobia or sexism (Waseem and Hovy 2016).

2.2 Hate speech detection

The automatic detection of hate speech has been addressed in many different ways. Studies have attempted to identify hate speech by proposing a binary classification problem which is an approach we also adopt. The task aims at detecting whether a given text is hateful (de Gibert *et al.* 2018; Jaki *et al.* 2019). Some studies focused on subtypes of hate speech, for example, Waseem and Hovy (2016) studied the classification of racist and sexist tweets. Others, instead, proposed a multi-label classification task: for instance, Davidson *et al.* (2017) detected hate speech, offensive language and non-hateful posts.

To automatically classify hate speech, previous research focused on traditional machine learning classifiers, such as logistic regression (Davidson *et al.* 2017), support vector machines (de Gibert *et al.* 2018) and naive Bayes (Kwok and Wang 2013). The input features of these classifiers are lexical and syntactic features, including bag-of-words (BoW) and term frequency-inverted document frequency (TF-IDF) (Sparck Jones 1972). Both calculate the occurrence of word or character n-grams (i.e., contiguous sequence of n items) in the text, but TF-IDF assigns higher weights to more informative words. Kwok and Wang (2013) examined the effect of BoW features as input and showed the ineffectiveness of a single linguistic feature due to insufficient information of the text. More sophisticated work includes combining multiple features. This is the case of Davidson *et al.* (2017) who achieved an F1-score of 90% in their proposed dataset. They used n-grams that range from 1 to 3 weighted by their TF-IDF. They added part-of-speech tags that are categories of the words (e.g., noun and verb) to capture some morpho-syntactic information and used Twitter metadata, such as tweet sentiment score and quality.

Recent work showed deep neural networks to outperform traditional machine learning systems in many cases because they better capture the complex relationships of the data. The most frequently used deep neural networks include CNNs (Badjatiya *et al.* 2017), recurrent neural networks (RNNs) and its variants – long short-term memory (LSTM) (Badjatiya *et al.* 2017) and gated recurrent unit (GRU) (Alshalan and Al-Khalifa 2020) – and transformers, such as BERT (Devlin *et al.* 2019). CNN captures the local features of the text, RNN extracts sequence information and

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

Table 1. Examples of texts where propaganda (Da San Martino *et al.* 2020) and toxic spans[a] are highlighted in bold

Task	Text
Propaganda	Coronavirus **'risk to the American people remains very low'**, Trump said.
Toxic	What if his opinion is that most other commenters are **idiots**?

transformers adopt an attention mechanism that learns the relationship among all the words of the input text based on their importance (Vaswani *et al.* 2017).

Some researchers have explored the combination of these deep neural networks. Zhang, Robinson, and Tepper (2018) proposed a CNN-GRU model that improved the F1-score of a single CNN by 1% on multiple hate speech datasets. Mozafari, Farahbakhsh and Noël (2019) fine-tuned pre-trained BERT learned on the BookCorpus (Zhu *et al.* 2015) and English Wikipedia. They examined the effect of different layers on top of BERT including a CNN and a bidirectional LSTM. Both models outperformed traditional machine learning classifiers on the datasets introduced by Waseem and Hovy (2016) and Davidson *et al.* (2017). The majority of these classifiers use word embeddings as input features. Word embeddings are real-valued word representations such that words with similar semantics are closer in the vector space Bengio *et al.* (2003). The most commonly used pre-trained word embeddings are Global Word Vectors (GloVe) (Pennington, Socher, and Manning 2014), FastText (Bojanowski *et al.* 2017) and Word2Vec (Mikolov *et al.* 2013).

Despite the improvement in classifier performance, it is unclear how the systems generalise because they are trained and tested on a single dataset representing a single data source, such as Twitter (Davidson *et al.* 2017). One way to achieve generalisability is to train a classifier with data from multiple platforms (Bruwaene *et al.* 2020). Recent studies investigated multi-platform classifiers for cyberbullying (Bruwaene *et al.* 2020) and hate speech detection (Corazza *et al.* 2019) and showed their effectiveness compared to mono-platform classifiers. While they combined data from mainstream social media (e.g., Twitter, Facebook, WhatsApp and Instagram), we aim at investigating multi-platform classifiers trained on Twitter, underground hacking and extremist forums that have more diverse discussion topics (e.g., cybercrime and white supremacy).

2.3 Span extraction

SE aims at identifying the fragment of the text of interest. Previous work have focused on two shared tasks from SemEval-2020 (Task 11) (Da San Martino *et al.* 2020) and SemEval-2021 (Task 5).[a] The first involves extracting the propaganda (i.e., expression that influences other people's opinion or actions) spans, and the second toxic or abusive spans from a text. Table 1 shows an example of propaganda and toxic spans.

The solutions for the two tasks can be categorised into span prediction and sequence labelling. Span prediction identifies the start and end offsets of the span (Chhablani *et al.* 2021). Sequence labelling classifies each member of a sequence, for example, identify whether each token is toxic (Chhablani *et al.* 2021) or use BIO encoding (i.e., mark the token as (B) if it is at the beginning, (I) if it is inside or (O) if it is outside of the span) (Morio *et al.* 2020).

Since SE tasks require highly nuanced semantic understanding, most solutions leveraged large language models pre-trained using transformers, including BERT (Devlin *et al.* 2019) and other types of transformers (Morio *et al.* 2020; Chhablani *et al.* 2021). These models are pre-trained on billions of words of English text data and can be easily fine-tuned to adapt to new tasks.

For the propaganda SE task, Jurkiewicz *et al.* (2020) treated it as a sequence labelling problem and used a conditional random field (CRF) (Lafferty, McCallum, and Pereira 2001) which is

[a]The toxic span extraction shared task and the examples can be found at: https://competitions.codalab.org/ competitions/25623#learn_the_details-overview.

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

a conditional probabilistic graphical model for labelling or parsing sequential data. The authors inserted a CRF layer on top of RoBERTa (Liu *et al.* 2019), a variant of BERT. They experimented with an ensemble of a model trained on the shared task dataset and one on the combined dataset of the original and silver data produced using self-learning. Though they achieved promising results, their performance was poorer compared to the solution of Morio *et al.* (2020). Although they also used an ensemble of pre-trained transformers, their success relied on a complex heterogeneous multi-layer neural network. The network uses the representation for each input token generated by the pre-trained language models, part-of-speech and named entity embeddings. These are fed into three bidirectional LSTMs, each of them is responsible for a different task: namely BIO sequence tagging, token-level and sentence-level classification. Their system achieved an F1-score of 51.5% ranking in first place for SemEval-2020 task 11.

For the toxic SE, Chhablani *et al.* (2021) experimented with span prediction, binary sequence labelling and a hybrid of the two approaches. Their best performing model was the second approach achieving an F1-score of 68.5%.

The performance of these systems may reflect the difficulty of the two tasks. Extracting toxic spans is easier than the propaganda spans, because toxic spans are often characterised by the use of offensive words which might be repeated and easier to recognise. The propaganda spans, instead, tend to be more heterogeneous with 14 different styles identified by Da San Martino *et al.* (2020). Therefore, the system must better capture the meaning of the text.

In terms of hate speech spans, the most relevant work is from Binny *et al.* (2021) who introduced a hate speech dataset called HateXplain. Each instance of HateXplain contains the class (i.e., hate speech and offensive or normal), the target community and the rationales which are parts or spans of the texts that annotators justify the labelling decision for being hate speech of offensive post. The rationales would be generated by the machine learning models to explain their classification results. For the rationales, Binny *et al.* (2021) generated a ground truth attention vector where they assigned 1 to each token in the rationale and normalised the attention vector so that the sum of the tokens equals 1. In addition, they normalised the attention vector using a softmax function with a temperature parameter to prevent the difference between the values of rationale and non-rationale tokens from being low. They tuned the parameter using a validation set. They experimented with Bi-RNN with an attention layer and BERT that need to output an attention vector that should assign higher weights to tokens in the rationale. They achieved an F1 score of 50.6% and 41.1% respectively.

Although Binny *et al.* (2021) provided the data containing hate speech rationales, we will not use their dataset because their rationales are mainly part of the text that are often disconnected. In contrast, our objective is to extract connected spans from the text such that by only reading at them, it is sufficient to assist human annotators to understand the text and make the classification. Therefore, we will be creating a specific data corpus for the hate speech SE task.

Among the three approaches that we described, we decided to explore span prediction and sequence labelling approaches to automatically extract hateful spans. The latter approach was not considered due to the unavailability of a large data corpus to tune the temperature parameter in the softmax function.

We believe that the difficulty of the hate speech SE task falls in between these two tasks. It is likely to be more straightforward than the propaganda SE because a hateful span may contain offensive words which make it easier to identify. But similarly, it also requires rich semantic understanding because a hateful span should contain the target and the attribute that makes the text hateful. This makes it more challenging than the SE of toxicity alone.

2.4 Hateful and aggressive content in underground and extremist forums

Most of the work in underground hacking forums has so far focused on understanding discussion topics (Caines *et al.* 2018b) or the cybercrime marketplace, for example, identifying key actors (Pastrana *et al.* 2018b) or supply chains (Bhalerao *et al.* 2018). In terms of hateful and

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

aggressive content, the most relevant work is from Caines *et al.* (2018a). They reported a lower level of aggressive language on HackForums, the largest English hacking forum (Pastrana *et al.* 2018a), compared to the Wiki Comments Corpus (Wulczyn, Thain, and Dixon 2016), a dataset that consists of 115,864 comments about Wikipedia page edits. They found that the interaction among HackForums members tends to be more positive as many posts are mainly instructive and educational.

In contrast, there has been more research relating to hateful content on extremist forums. Gerstenfeld, Grant, and Chiang (2003) analysed 157 extremist websites and observed racist content in almost half of these websites. Although the findings suggest a high proportion of hateful content, it is unknown how hate speech is distributed across individual websites. de Gibert *et al.* (2018) investigated the frequency on Stormfront, one of the longest-running white supremacist platforms (Bowman-Grieve 2009). They randomly sampled content from several subforums and split them into sentences. They manually labelled 9916 sentences among which 11% were hateful. However, because they analysed hate speech at the sentence level, the actual distribution of hate speech over texts is unclear.

Besides white supremacist forums, extremist forums include 'incel' forums. Involuntary celibates or incels are single men who adhere to an extreme misogynistic, anti-feminist ideology (Jaki *et al.* 2019). Previous work analysed the content on Incels.me (Jaki *et al.* 2019) and in an incel Reddit subforum (Tranchese and Sugiura 2021). According to their findings, these forums are full of abusive language. Incels express their hate towards women because they attribute their lack of sexual activity and their misfortunes in life to women (Jaki *et al.* 2019; Tranchese and Sugiura 2021). They objectify and dehumanise women and encourage violence by providing instructions to rape or murder women (Jaki *et al.* 2019). Jaki *et al.* (2019) also used a deep neural network to detect hate speech on Incels.me. They defined hate speech as posts containing offensive words. They, initially, chose 10 offensive words related to misogyny, homophobia and racism. They reported a distribution of 5% of hate speech in 50,000 posts. This is likely an underestimate given they only selected 10 offensive words and three subsets of hate speech. Moreover, as Assimakopoulos *et al.* (2020) stated that not all hate speech contains offensive words. Therefore, the picture of hate speech in incel forums is incomplete.

3. Data

This section provides an overview of the data that we extracted from the HatEval, CrimeBB and ExtremeBB corpora. The latter two contain hacking- and extremist-related posts, respectively. The first corpus includes posts from Twitter and helps to augment the dataset for the training of the multi-platform classifiers. We also discuss the ethical considerations that play an important role in this work.

3.1 HatEval

The HatEval dataset was introduced by Basile *et al.* (2019) for the SemEval-2019 hate speech detection shared task. The dataset was extracted from Twitter by identifying potential victims of hate speech and hate accounts and using the keyword approach (i.e., selecting potentially hateful posts based on offensive words). The dataset contains tweets targeting women and immigrants in English and Spanish. The data were labelled by two expert annotators who are experienced in the annotation of this task and also using crowdsourcing. Because only 10% of the labelled data contained hate speech, Basile *et al.* (2019) altered the natural distribution to have a more balanced distribution of hateful and non-hateful tweets.

We used the English portion of the dataset that consists of 9000 training and 3000 test data instances. Each dataset contains 42% hateful and 58% non-hateful tweets.

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

Table 2. List of keywords for the search of potential hate speech

Category	Keywords
Gender	whore, cunt, slut, women
Religion	musla, islam, jewish, religious
Nationality	arab, chinese, japanese, spanish, chink
Sexual orientation	fag, gay, lesbian, queer
Race	nigga, nigger, white, black, negro
Class	ghetto, rich
Disability	douchebag, moron, retard

3.2 CrimeBB and ExtremeBB

We used existing datasets called CrimeBB (Pastrana *et al.* 2018a) and ExtremeBB (Vu *et al.* 2021) that contain posts from underground and extremist forums, respectively. CrimeBB and ExtremeBB are collected and maintained by the Cambridge Cybercrime Centre.[b] Both datasets continue to expand, but at the time of writing CrimeBB contained more than 90 million posts from underground forums and ExtremeBB contained more than 38 million posts from extremist forums. We extracted a sample of posts from HackForums, the largest forum in CrimeBB and Stormfront and Incels.co from ExtremeBB to analyse.

HackForums is the most well-known hacking community that gained recent attention in 2017 due to the arrest of a suspicious author of banking malware (Krebs 2017). While members of HackForums might be engaged in cybercriminal activities, people driven by extreme ideologies such as those expressed on Stormfront and incel forums have been involved in serious offline illegal harms. Many crimes committed by murderers have self-identified as incels, including the Hanau shooter who killed nine people in 2020 (Jasser *et al.* 2020). Furthermore, many massacres are associated with Stormfront including the mass murder of 77 people in Norway in 2011 (Holpuch 2014).

We worked with posts extracted from HackForums that were posted by users between 2007 and 2020 and from the two extremist forums that were posted between 2001 and 2021.

Initially, we randomly selected 500 posts from two subforums of HackForums, called *religion-philosophy-science* and *news and happenings*, that are likely to contain hate speech because they are not technical- or game-related. We also randomly sampled data from the two extremist forums. Many posts from Stormfront were non-English and contained news articles related to non-white people committing a crime. We filtered out these as we were seeking user-generated texts in English leaving 500 posts from the two forums. We further extracted 2200 posts, half of which come from the entire HackForums and the other half from the two extremist forums, using a keyword approach in case of a low frequency of hate speech.

Because this work considers hate speech in general, we used the keywords listed in Table 2 from seven categories defined by Hatebase Inc. (2020), the world's largest online repository of multilingual hate speech, that can cover a broad amount of communities. These categories are gender, religion, nationality, sexual orientation, race, class and disability.

[b]Both datasets can be found at: https://www.cambridgecybercrime.uk/process.html and can be accessed by application to the Cambridge Cybercrime Centre director.

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

Fleiss Kappa	Interpretation
<0.00	Poor agreement
0.00 to 0.20	Slight agreement
0.21 to 0.40	Fair agreement
0.41 to 0.60	Moderate agreement
0.61 to 0.80	Substantial agreement
0.81 to 1.00	Almost perfect

Figure 1. Framework for interpreting Fleiss' kappa.

3.3 Ethical issues

This work uses and analyses data that may contain harmful content from CrimeBB and ExtremeBB. Because of this, this research has undergone ethical review to consider potential harms, ensure safeguards to protect the researchers and participants and report potential criminal materials, such as terrorist and child sexual abuse material, to the authorities. We received ethics approval from the Department of Computer Science's ethics committee and complied with the Cambridge Cybercrime Centre's data-sharing agreements. Researchers and annotators were warned about the hateful content before they read the data, and they were made aware of the counselling services available to them. Finally, we took care in storing data securely, not revealing user identities (e.g., usernames) and reporting the results objectively.

4. Data annotation

Data annotation is an integral part of supervised and semi-supervised machine learning tasks. It is required when the labelled data for the specific task is not available or to augment existing datasets, since machine learning systems require a large amount of data to learn patterns and make accurate predictions. This work requires the annotation of the unlabelled data that we extracted from HackForums, Stormfront and Incels.co for training hate speech detection and SE models. The annotators for both hate speech detection and SE tasks are three researchers from the Department of Computer Science of the University of Cambridge.

4.1 Annotation for hate speech classification

We randomly sampled the extracted dataset and divided it into training and test sets. The first consists of 2200 posts and was annotated using active learning as described below, whereas the test set contains 1000 posts that were manually labelled by 3 annotators.

The task consists of labelling 1 if the post is hateful, otherwise, 0. It is based on the definition provided by the UN (2020) that is enriched with other aspects of hate speech that we listed in Section 2.1. The annotation process is evaluated using Fleiss' kappa (Fleiss 1971) inter-annotator agreement, and we used the framework from Landis and Koch (1977) as shown in Figure 1.

4.1.1 Active learning

Active learning is a machine learning algorithm that interactively queries the user to label the data that the system is uncertain (Cohn 2010). We used this algorithm because training a multi-platform classifier requires a large amount of training data from the hacking and extremist forums, and manual annotation is time-consuming. Algorithm 1 outlines our procedure for active learning.

We used SGDClassifier, more specifically a SVC, with smoothed hinge loss, from Python ScikitLearn (Pedregosa *et al.* 2011) which supports incremental learning – a method of machine learning in which the model's knowledge is continuously updated with new data without being

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

Table 3. Distribution of posts over categories in the multi-platform training data

Assigned labels	#posts	Percentage
Hateful	4089	38
Non-hateful	6711	62

Algorithm 1. Active learning (Uncertainty sampling strategy) algorithm.

(1) **U** ← unlabelled data.

(2) **L** ← labelled data.

(3) (Incremental) Train a classifier on **L**.

(4) Empty **L**.

(5) Predict the probabilities for a subset of unlabelled data **C**.

(6) If the predicted probability is above the confidence threshold, the instance is given its predicted class, added to **L** and removed from **U**.

(7) If the predicted probability is below the confidence threshold, manually label the data, add them to **L** and remove them from **U**.

(8) Repeat Steps 3–7 until **U** is empty or a stopping criterion is met.

retrained from scratch (Geng and Smith-Miles 2009). We used text embeddings as input features. To generate these, we used the Universal Sentence Encoder from Tensorflow Hub (Abadi *et al.* 2016) because it shows strength in capturing the semantics of and the similarities among texts (Cer *et al.* 2018). Because the Universal Sentence Encoder was trained for greater-than-word length text (e.g., sentences, phrases and short paragraphs), the English input text can be of variable length, but the output is a fixed 512-dimensional vector (Cer *et al.* 2018).

The confidence threshold that we set was initially 0.9 for both hateful and non-hateful predictions. Texts for which model predictions had a confidence lower than the threshold were passed to the human annotators for labelling. However, the classifier ended up being confident in predicting the latter and not so much in predicting hate speech. Therefore, we lowered the threshold to 0.5 for potential hate speech.

We started by training the classifier on the HatEval training data. The initial performance tested on the HatEval test set achieved an F1-score and an accuracy of 58% and 59%, respectively. We used this classifier to label posts from CrimeBB and ExtremeBB some of which could not be assigned a label because the human annotator was unable to confidently determine whether the post is hateful or not. There were 400 discarded posts leaving 1800 labelled data instances. Among these, there were 306 hateful and 1494 non-hateful posts. These data were combined with the HatEval training data to construct a single multi-platform training dataset. Table 3 shows the distribution of the data.

4.1.2 Manual annotation

The annotators were asked to manually label the test sets from ExtremeBB and CrimeBB. After removing 29 posts that could not be clearly labelled, the final set consisted of 971 posts. We assigned the final label for each instance based on the majority voting from the three annotators. We divided the data into two test sets, one from HackForums and one from the two extremist forums. Table 4 shows the distribution of the hateful and non-hateful posts across test sets from HatEval, underground and extremist forums.

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

Table 4. Distribution of posts over categories in the test sets from HatEval, HackForums and the two extremist forums

	HatEval		HackForums		Extremist forums	
Assigned labels	#posts	Percentage	#posts	Percentage	#posts	Percentage
Hateful	1260	42	82	17	231	47
Non-hateful	1740	58	396	83	262	53

Table 5. Distribution of the data instances for span extraction across different platforms

Platform	Total posts	Percentage
HatEval	213	31
HackForums	80	11
Extremist forums	407	58

To measure inter-rater agreement, we again turned to Fleiss' kappa and obtained a score of 0.73, indicating that there was substantial agreement among the three annotators.

4.1.3 Discussion

Despite the available hate speech definition, the labelling process was admittedly challenging. The main source of disagreements was due to the difficulty of determining the dividing line between offensive language and hate speech. Besides, due to the unavailability of demographic information of the speaker and the addressee, annotators were unable to determine whether the offensive word is referring to the addressee or whether the speaker belongs to the target group.

Other disagreements include the expression of excessive pride of a particular race, for example, the claim of *white power* on Stormfront, and sarcastic posts because contextual information to make an informed decision about the real intention behind the posts was not available.

Furthermore, as mentioned in Section 4.1.2, we discarded 29 posts because at least one annotator could not provide a label. They are, probably, part of a dialogue or discussion thread that cannot be understood without considering the wider context.

Finally, we noticed that in active learning annotation, the machine was mainly confident in predicting non-hateful posts resulting in these being the majority even though the initial model was trained on the approximately balanced two-class HatEval training set. For hate speech, the classifier correctly classified hateful posts against women. This is probably because the initial training data for the active learning classifier was from HatEval that only contains posts related to women and immigrants. Therefore, the system may be biased towards the two targets and not recognise hate speech against other targets which were present in the unlabelled data.

4.2 Annotation for hate speech SE

For hate speech SE, we randomly extracted 700 posts labelled as hateful from HatEval, HackForums and Extremist forums. Table 5 shows the distribution of the data instances across different platforms.

The purpose of hate speech SE is to determine which section(s) of a text are the source of hateful content. A possible application for such a system is to enable moderators to quickly review

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

Table 6. Hate speech frequency across different platforms

Platform	#hate speech	Total posts	Percentage
Twitter	5043	12,000	42
HackForums	122	1257	9
Stormfront	208	716	29
Incels.co	289	819	35

potentially harmful texts by bringing key sections to their attention quickly. The span must contain the fragment that incites hate as we described in Section 2.1. The annotators were asked to extract the hateful span from the posts.

The annotation process is evaluated in terms of γ agreement (Mathet *et al.* 2015) that measures the overlap of the spans that the annotators extracted. Normally, γ is a score between 0 and 1, where 1 means the annotators extracted the same spans. When the annotated spans are completely different from each other, γ can be less than 0 (Mathet *et al.* 2015).

All the data were labelled by only one annotator, except for 50 instances that were labelled by 3 annotators to calculate the inter-annotator agreement. The average γ agreement was 0.87.

The annotated spans are all longer than a few words as may be found in, for example, toxic spans. This is expected because hate speech needs to include a target and the attribute that makes the post hateful.

We also observed some posts containing multiple disconnected spans. We updated the annotations to include these, and there are in total 44 posts with multiple spans.

4.3 Findings

4.3.1 Hate speech frequency

We combined the test and training data of each platform. Table 6 shows the hate speech frequency on different platforms. The statistics only act as an indicator and do not reflect the real-life distribution because hate speech may be over-sampled as they have been extracted in a targeted fashion using the keyword approach.

HackForums has the lowest occurrence of hate speech. This is in line with the findings of Caines *et al.* (2018a) who reported a relatively low level of abusive and aggressive behaviour. The positive behaviour of HackForums members is probably due to the strict rules imposed on the website. According to Caines *et al.* (2018a), there are administrators and a reputation scoring system that constrain user behaviour.

The hate speech frequency of the two extremist forums, Stormfront and Incels.co, reaches 29% and 35%, respectively. This is not surprising because these forums are driven by extreme ideologies, such as anti-Semitism (de Gibert *et al.* 2018) and misogyny (Jaki *et al.* 2019). Besides, unlike HackForums which constrains users behaviour, in these extremist forums, it is unlikely to enforce restrictions on hate speech because it would deviate from the ideologies of these forums. Most of the rules are probably concerned with cyberbullying as Jaki *et al.* (2019) found on Incels.me.

Twitter has the largest amount of hate speech. However, as we described in Section 3.1, the HatEval dataset distribution of hateful and non-hateful content was updated to have a more balanced distribution of the two categories. The original data only contained 10% of hate speech which was less than the two extremist forums. This is probably due to site moderators who actively ban illegal content (Harrison 2019). Despite the lower frequency, the total number of texts containing hate speech would be greater than other platforms (Stricker 2014) because Twitter users post 500 million tweets per day (Stricker 2014), whereas HackForums, Stormfront and

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

Table 7. Examples of vocabulary that members on Incels.co use

Words	Explanation
Femoid, foid, stacey	Woman
Chad, tyron	Good looking man
Bluepill	The preference of remaining ignorant to be happy and opposition to the belief that physical attraction plays a key role in society (Incels Wiki 2017)
Redpill	A belief that women have many suitors, thus they develop restrictive standards in dating and incels would not have any chance (Incels Wiki 2017)
Blackpill	A belief that women only date good-looking men (Incels Wiki 2017)

Incels.co contain in total around 42 million, 10 million and 6 million (counted from CrimeBB and ExtremeBB) posts, respectively.

4.3.2 Analysis of forums content

We inspected the data to better understand the content on HackForums, Stormfront and Incels.co. The use of derogatory terms is frequent in all these forums. In line with previous findings (Pastrana *et al.* 2018a; Caines *et al.* 2018a), we found that the primary content on HackForums is technical and commercial. Forum members share hacking knowledge and focus on earning fast money from selling illicit materials. In terms of the few hateful posts that we found, some involved conflict among forum members, where members insult individuals using discriminatory words related to certain group identity factors. HackForums members do not tend to have a particular group they attack.

In contrast, the target of the two extremist forums is more evident. On Stormfront, users show hatred against a particular group, Jews. This can take the form of demonising the group, denying the Holocaust and spreading the conspiracy that the Jews created COVID-19. Stormfront also contains racist content, for example, demonstrated by white nationalists who claim the supremacy of *white power*. These forum members attribute social problems, such as high crime rates and poverty, to non-white groups. Other hateful posts include expressions of disgrace and hate towards women who date non-white people by calling them *race traitors*. Nonetheless, the discussion in Stormfront is not always hateful. They also discuss topics such as politics and use the forum as a place to socialise. We found posts in which they introduce themselves and invite other members to meet offline.

Unlike HackForums and Stormfront, incels use specific vocabulary, from which we list a few examples in Table 7. The primary targets on Incels.co are women who are portrayed as being immoral, corrupted, promiscuous and superficial. Although we found incels incite violence towards this target group, we did not observe any detailed instructions on how to rape and kill women that Jaki *et al.* (2019) found on Incels.me. In addition to women, incels also hate *chads* because they believe *chads* raised women's standards in finding a partner and, thus, women would not 'date down' with incels. But, at the same time, they admire *chads* for their physical appearance. Other hateful posts include racist posts. Some of them are related to women, for example, they discuss women of which ethnicity are the easiest to date and compare women's physical appearance of different races. Most of these hateful topics are consistent with previous findings (Jaki *et al.* 2019; Tranchese and Sugiura 2021) suggesting that the discussion of incels is similar across different platforms.

In terms of non-hateful content, similar to what Jaki *et al.* (2019) found, incels suggest ideas to become attractive and to date women. Furthermore, we observed that on Incels.co, there are

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

many complaints. Incels complain about their physical appearance, being maltreated or bullied. Some of them even considered suicide because of their miserable life and fellow incels comforted them to prevent the tragedy from happening.

5. Methods

5.1 Hate speech detection

This work defines hate speech detection as a binary classification problem, where the classifier outputs 1 if hate speech is detected and outputs 0 if the post does not contain hateful content.

5.1.1 Data pre-processing

We lowercased each post to avoid word types such as 'Hello' and 'hello' being treated differently (Pradha, Halgamuge, and Tran Quoc Vinh 2019). We also removed punctuation, URLs, @ mentions and hashtags because they may not provide much information to the text and may be noisy (Gurusamy and Kannan 2015). Finally, we used text embeddings as input features which are generated using the Universal Sentence Encoder (Cer *et al.* 2018).

5.1.2 Models

We explored different machine learning models. Each model has two versions, one where the models are trained using only the HatEval training data and the other where the models are trained on a dataset from multiple sources (i.e., HatEval, Hackforums and Extremist forums).

Support Vector Classifier (SVC) A SVC is a linear model that works for classification problems (Boser, Guyon, and Vapnik 1992). It separates the data into classes by finding the optimal hyperplane that maximises the margin – the distance between the hyperplane and the support vectors, which are the points closest to the hyperplane. The maximisation reduces the generalisation error of the system (Boser *et al.* 1992).

We used the model from Indurthi *et al.* (2019) that ranked first in the SemEval-2019 hate speech detection subtask. It is an SVC with Radial basis function kernel, which maps data to a higher-dimensional space and Universal Sentence Encoder sentence embeddings.

CNN-GRU We also experimented with CNN-GRU that Zhang *et al.* (2018) used to achieve better performance on multiple hate speech datasets compared to a single CNN. The first layer of the model is an embedding layer that loads weights of pre-trained word embeddings. We use word embeddings with 300 dimensions pre-trained on 3 billion words from Google News (Mikolov *et al.* 2013). The output feeds into a dropout layer with a rate of 0.2 to avoid overfitting (Hinton *et al.* 2012). Then, the output feeds into a convolutional neural network (CNN) that consists of a 1D convolutional layer (Conv1D) and a max-pooling layer (Max Pooling 1D). The first uses 100 filters, a window size of 4 and a rectified linear unit as activation function. Max Pooling 1D with a pool size of 4 down-samples the input feature by taking the maximum value, which can be considered as the most salient information in the text (Goldberg 2015). The information is then fed into the GRU layer, which captures the sequence information, that is flattened out after being passed to a global max-pooling layer. Finally, we used a fully connected layer to output a prediction.

BERT BERT stands for Bidirectional Encoder from Transformers and was introduced by Google in 2019 (Devlin *et al.* 2019). As the name suggests, it adds bidirectionality to the standard transformer which is a network architecture solely based on attention mechanisms (Vaswani *et al.* 2017). The bidirectionality allows the machine to read the entire text at once and is achieved by using the Masked Language Model (MLM). MLM randomly masks some input tokens and

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

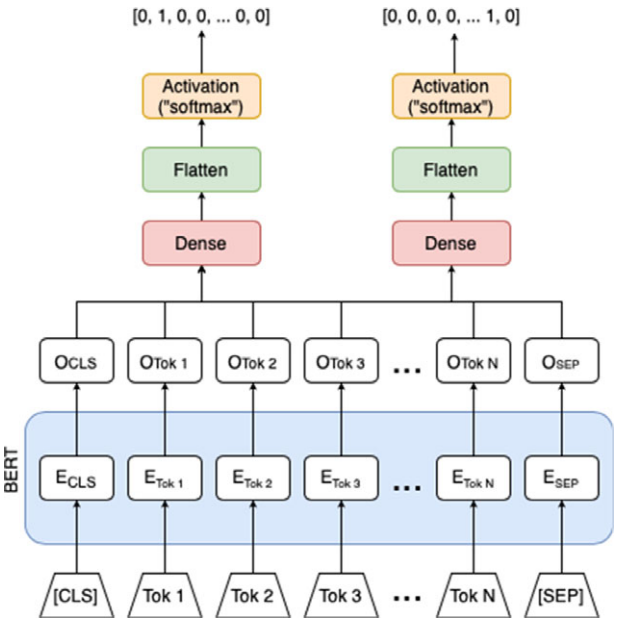

Figure 2. Architecture of BERT+span.

aims at predicting the masked words by considering the left and right contexts. In addition to the MLM, BERT is trained on another unsupervised task, namely Next Sentence Prediction (NSP) which learns sentence relationships. It predicts whether a sentence is subsequent to the first input sentence.

BERT is pre-trained on 800 million words from the BookCorpus and 2500 million words from English Wikipedia. We decided to fine-tune pre-trained BERT models from HuggingFace (Wolf *et al.* 2020) because training BERT from scratch is computationally expensive. Fine-tuning updates the parameters of pre-trained BERT based on our dataset and task.

We used $\text{BERT}_{\text{base\_uncased}}$ from HuggingFace's (Wolf *et al.* 2020) transformers. On top of BERT, we added a fully connected layer. We used a batch size of 32 and the Adam optimiser (Kingma and Ba 2015) with a learning rate of 1e-6.

5.2 Hate speech SE

To extract hateful spans from posts, we fine-tuned BERT and performed span prediction and sequence labelling. Both of these models are token-level based. Tokens are small units (e.g., words, phrases, symbols or other meaningful elements Gurusamy and Kannan 2015) into which a text is split.

5.2.1 Models

We implemented two models to extract hate speech spans.

BERT+span is based on span prediction and outputs a single span. It predicts the start and end indices of the first and last token of the ground truth spans. Figure 2 shows the architecture of BERT+span. On top of the BERT model, we inserted two fully connected dense layers to predict the start and end indices. The softmax activation function generates a probability distribution over the indices being the start or end indices. The index with the highest probability from each prediction is selected.

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

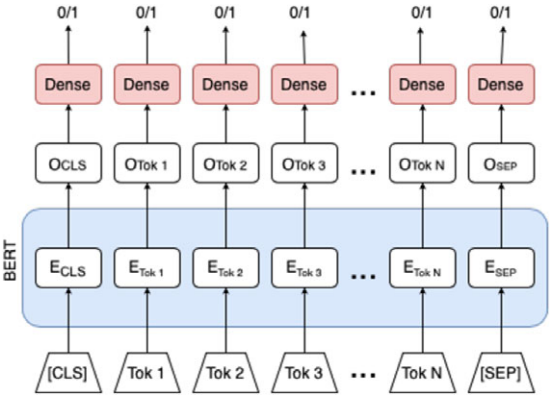

Figure 3. Architecture of BERT+token.

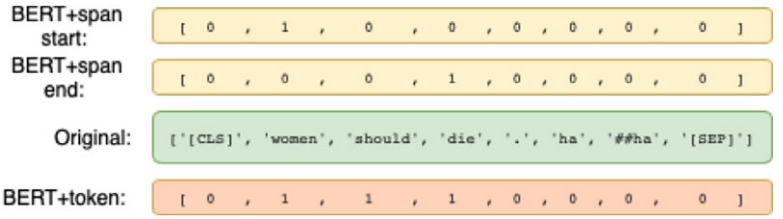

Figure 4. Example of how single ground truth spans are pre-processed for BERT+span and BERT+token. The original text has been encoded into sub-tokens with Bert WordPiece tokenizer.

BERT+token relies on sequence labelling. It labels each token to be either hateful (1) or not hateful (0). Figure 3 shows the architecture of BERT+token. On top of the BERT model, there is a fully connected dense layer for each token.

Both models fine-tuned $\text{BERT}_{\text{base\_uncased}}$. We used a batch size of 32 and the Adam optimiser with a learning rate of 3e-4.

5.2.2 Data pre-processing

We used BERT WordPiece tokenizer (Wolf *et al.* 2020) that lowercases and splits the text into a list of tokens. It always inserts two special tokens. These are [CLS] which is at the beginning and [SEP] which separates sentence pairs. The tokenizer handles out-of-vocabulary (OOV) by breaking down unseen words into subwords (Wolf *et al.* 2020).

The input of the models is the original text and is pre-processed in the same way for the two models. The input consists of two vectors, called input ids and attention mask. The input ids are numerical representations of tokens of the original text. The attention mask sets all tokens of the original text to 1 to which the model should pay attention.

Contrarily, the ground truth spans are pre-processed differently for the two models. Figures 4 and 5 show how a single hateful span and multiple spans, respectively, are processed for BERT+span and BERT+token.

For BERT+span, we created two vectors of zeros, called *start* and *end* where the start and end offsets of the output span are set to 1. When the text has a single hateful span, the start and end offsets are the indices of the first and last tokens of the ground truth span. When there are multiple spans, the system should connect all the spans to not lose any information. In this case, the start offset is the index of the first token of the first ground truth span and the end index is the last

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

Table 8. Performance of the classifiers, where [Mono] and [Multi] mean mono- and multi-platform classifier, in percentage in terms of accuracy, precision, recall and F1-score. Values in bold are the best scores

	HatEval		HackForums		Extremist forums	
Model	Acc	F1	Acc	F1	Acc	F1
SVC [Mono]	64.5	66.4	77.6	28.1	64.5	52.5
SVC [Multi]	**65.8**	**67.1**	**82.6**	33.6	**69.7**	58.2
CNN-GRU [Mono]	56.7	36.1	58.9	20.3	48.6	36.2
CNN-GRU [Multi]	53.7	44.5	51.4	27.5	52.1	51.6
BERT [Mono]	57.6	60.5	44.3	33.8	52.7	**59.4**
BERT [Multi]	57.3	58.2	71.7	**41.0**	59.0	48.7

Figure 5. Example of how multiple ground truth spans are pre-processed for BERT+span and BERT+token. The original text has been encoded into sub-tokens with Bert WordPiece tokenizer.

index of the last span token. For BERT+token, instead, we created one zeroed vector and marked as 1 the indices of all the tokens belonging to the spans.

Finally, all the input and the ground truth vectors have their first and last indices set to zero for the [CLS] and [SEP] tokens. They were padded by zeros to have the same length. We set a maximum sequence length of 168 tokens.

5.2.3 Post-processing

We performed post-processing for the predictions of the two models. In terms of BERT+span, we connected all the tokens of the original text starting and ending at the predicted indices.

For BERT+token, because it relies on sequence labelling, the system may not consider some tokens to be hateful and disconnect a hateful span. These tokens are the subwords generated by BERT WordPiece tokenizer and English stop words. We, first, connected all the subwords to avoid the case in which subwords of the same word are classified differently. Then, we kept merging two spans if the tokens between them are stop words.

6. Evaluation

6.1 Hate speech detection results

The models, described in Section 5.1.2, are evaluated on the HatEval, HackForums and Extremist forums test sets using accuracy and F1-score. The results are reported in Table 8.

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

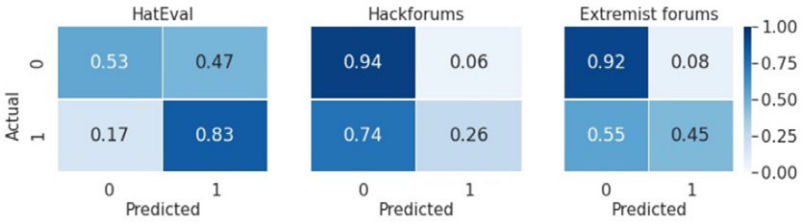

Figure 6. Confusion matrix for the SVC[Multi].

The SVC[Multi] outperformed other models in terms of accuracy on the three test sets, whereas achieved the best F1-score only on the HatEval test set. BERT[Mono] and BERT[Multi] achieved the best F1-score on the Extremist forums and Hackforums test sets.

We also observed that all the multi-platform classifiers outperformed mono-platform classifiers on HackForums test sets and SVC[Multi] outperformed SVC[Mono] in terms of accuracy and F1-score on all three test sets. This was not the case for CNN-GRU and BERT. CNN-GRU[Mono] achieved higher accuracy than CNN-GRU[Multi] on HatEval and HackForums, and BERT[Mono] accuracy on HatEval was slightly higher than BERT[Multi] and achieved a better F1-score on HatEval and Extremist forums. This suggests that multi-platform classifiers may not always outperform mono-platform classifiers which was shown by Corazza *et al.* (2019).

We inspected the confusion matrix, shown in Figure 6, of the SVC[Multi] because it has overall the best performance, and we analyse the performance of the SVC[Multi] across different test sets.

6.1.1 Performance across different test sets

The SVC[Multi] achieved the highest F1-score in HatEval. This is probably due to the HatEval training data being the largest of the multi-platform training data. However, the system achieved a false positive rate of 50%, which is the highest across all the test sets (Figure 6). A possible explanation would be the system being unable to differentiate offensive language and hate speech and classifying non-hateful posts containing pejorative words as hateful.

In terms of HackForums, the SVC[Multi] achieved a poor F1-score of 33.6%, but a high accuracy of 82.6%. The confusion matrix in Figure 6 shows that the system is slightly skewed towards non-hateful predictions. This explains the high accuracy in HackForums because, as Section 4.1.2 reports, it does not contain much hate speech. Though the system scored a low false positive rate (6%), it misclassified 74% of hate speech. This is probably the same problem that we encountered during active learning in which the system scarcely recognised hate speech categories that were not present in the HatEval training data. The HatEval training data only contain posts related to women and immigrants and account for the majority of the multi-platform training data. The unbalanced distribution of hate speech categories may affect the performance of the SVC[Multi].

In comparison, the SVC[Multi] performed better in Extremist forums because there are many misogynistic and racist posts on Stormfront and Incels.co that match the majority of hate speech in the training data. However, similar to the case of HackForums, the system scored a false negative rate of 55% (Figure 6) due to the presence of other hate speech categories.

6.2 Hate speech SE results

We evaluated our SE models described in Section 5.2.1 against two baseline models, namely Entire and Random. The first predicts the entire input text as the hateful span and the second randomly assigns 0 or 1 at the index of each token and outputs a vector. The predictions are post-processed in the same way as BERT+token, as described in Section 5.2.2. Due to the limited

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

Table 9. Performance of the span extraction models in percentage in terms of exact match, precision, recall and F1-score. Mean, standard deviation and maximum values across five runs are reported. Values in bold are the best scores

Model	EM (std,	max)	P (std,	max)	R (std,	max)	F1 (std,	max)
Entire	13.4 ± 0.0	13.4	48.1 ± 0.0	48.1	**100.0** ± 0.0	**100.0**	59.7 ± 0.0	59.7
Random	0.8 ± 0.3	1.2	48.6 ± 0.3	49.0	61.2 ± 0.9	62.8	48.0 ± 0.4	48.7
BERT+span	**32.0** ± 3.2	**37.1**	65.3 ± 2.4	68.4	79.0 ± 5.0	86.7	65.8 ± 2.6	69.4
BERT+token	20.2 ± 1.5	26.4	**68.9** ± 2.3	**72.5**	78.7 ± 3.1	84.0	**68.2** ± 1.5	**70.4**

data, we performed fivefold cross-validation to estimate how BERT+span and BERT+token generalise on unseen data.

We evaluated the performance of SE models with four metrics. These are exact match, precision, recall and F1-score introduced by Rajpurkar *et al.* (2016) to evaluate the models for the question answering (QA) task. The QA task extracts the answer from a reading passage given a question. Similarly, the hate speech SE task extracts the span from a hateful post except that it does not have a question as input. All the metrics are token-level based and ignore punctuation.

Exact match (EM) counts how many predicted spans exactly match the ground truth spans. It is calculated using Equation (1):

$$EM = \frac{\sum_{i=0}^{N} Match(pred_i, gold_i)}{N} \tag{1}$$

where *pred* is the predicted spans, *gold* is the ground truth spans, N is the data size and $Match(pred, gold)$ is

$$Match(pred, target) = \begin{cases} 1 & \text{if } pred = gold, \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Precision, recall and F1-score measure the percentage of overlap between the predicted and ground truth spans and are the same standard metrics used in classification with some differences in the notion of true positive (TP), false negative (FN) and false positive (FP).

In SE:

- TP is the number of tokens that are in both predicted and ground truth spans.
- FN is the number of tokens that are in the ground truth but not in the predicted spans.
- FP is the number of tokens that are in the predicted spans but not in the ground truth.

6.2.1 Results

For all models, except for Entire, we calculated the mean, standard deviation and maximum values for exact match, precision, recall and F1-score across five runs. The results are reported in Table 9.

Random has overall the worst performance. It failed at predicting spans exactly matching the ground truth due to its random behaviour and disconnected predictions. Entire, instead, achieved an EM score of 13.4%. However, this only indicates that there are 13.4% of instances in which the entire post is hateful. Entire also scored the highest recall (100%). This score is expected as recall is the ratio of the total number of overlapping tokens and the number of the ground truth spans. Because the prediction is the entire post, the two numbers would always be the same.

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

Compared to Entire, the two BERT-based models outperformed it except for recall. The performance of BERT+token in terms of precision and F1-score was better than BERT+span. One of the reasons is that BERT+token which is token-level based would be more precise, whereas BERT+span which outputs a single connected span may include many unwanted tokens.

Nevertheless, because BERT+span always outputs a continuous span, it achieved a higher EM score compared to BERT+token that may disconnect the spans. A pitfall of EM of BERT+span is that it relies on the number of posts in which there is a single hateful span. If all the training data consist of multiple disconnected spans, BERT+span that should connect the ground truth spans would achieve an EM of 0%.

6.2.2 Comparison with other works

To the best of our knowledge, this is the first work on extracting hate speech spans. In Section 2.3, we described the propaganda (Da San Martino *et al.* 2020) and toxic (Chhablani *et al.* 2021) SE tasks.

Because other solutions evaluated their models with F1-score, we compare BERT+token which achieved a maximum F1-score of 70.4%. Compared to our solution, the best performing system for the propaganda SE task has a more complex architecture achieving an F1-score of 51.5%. It is an ensemble of BERT and many types of transformers each of which is trained for three tasks: BIO sequence tagging, token-level and sentence-level classification. Similar to the hate speech SE, the propaganda SE shared task had limited data (446 news articles). The higher F1-score of our solution suggests that hateful spans are easier to identify because they are often similar to each other, since they express hate towards or attack individuals or groups.

Turning to toxic SE, the best system achieved an F1-score of 68.5% among the solutions proposed by Chhablani *et al.* (2021). Similar to BERT+token, it is based on sequence labelling except it fine-tunes SpanBERT (Joshi *et al.* 2019). In contrast, while BERT+token was trained on the data consisting of only hateful posts, they trained their model on 10,000 Civil Comments texts which do not always contain a toxic span. This may have increased the difficulty of their task and explain the better performance of BERT+token in identifying hateful spans.

7. Discussion

In this section, we perform an error analysis to get a deep understanding of the performance of the hate speech classifiers and SE models. The example posts that we present have been changed, while keeping the same meaning, for privacy reasons. Finally, we discuss the broader implications of the systems.

7.1 Hate speech detection

7.1.1 Classifier error analysis

We analysed the posts that were misclassified by the SVC[Multi], the best performing classifier. We considered how the system behaved, in terms of the problems that human annotators encountered during the annotation process. For example, the following post:

(1) XXX - welcome. I wish you enjoy living in such a culturally rich, diverse, vibrant, cosmopolitan area. Aren't those Muslims lovely, hardworking and peaceful people? (Extremist forums)

raised some debates among human annotators because some believed it is a sarcastic post whose real intention is the opposite. The majority of human annotators considered it to be hateful, but the machine classified it as non-hateful. This is not surprising because the post, literally, is not

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

hateful, and it would be hard for the machine to detect any sarcastic intent on the part of the author.

The other source of disagreement was the distinction between offensive language and hate speech. We selected some false positives and false negatives. Some of them contain pejorative words. The following two posts give an example of the most common posts that the system misclassified as hateful:

(2) I love you. . .bitch. I'll never stop loving you. . .bitch. You're my best friend [. . .] and I can't wait to hug you (HatEval)

(3) Just get them legitly, don't whore post. (HackForums)

Posts (2) and (3) are probably labelled as hate speech due to the presence of the offensive words *bitch* and *whore*. These posts show that the machine still cannot understand the meaning of the text and may be biased towards the use of offensive words even if they are used in a way that is offensive but arguably not hateful (oddly in the case of (2) the author is trying to express affection; in (3) 'whore posting' is a type of forum behaviour by attention-seeking 'newbies').

In terms of false negatives, the following posts show three examples:

(4) Nigga you are going to die (Extremist forums)

(5) They are gay and useless. (HackForums)

(6) It's not the Chinese virus. It's the Jew virus (Extremist forums)

Post (4) contains the offensive word *nigga*, but the system did not consider it to be hateful even if it is telling someone to die perhaps because of the non-standard orthography. In post (5), the author indicates that being gay is on a par with being useless. This may indicate that the machine struggled to understand hateful posts against other targets that constitute the minority in the training data because HatEval contains hateful posts against women and immigrants. The active learning annotation process brought in other types of hate speech, but we recognise that the initial seed set for training focused on a subset of hate speech, and therefore we emphasise that additional annotations should be collected specifically targeting other hate speech types such as ableism, homophobia and trans-exclusionary texts.

Example (6) is an interesting case. It shows a further difficulty of the hate speech detection task. Hate speech does not only manifest with sole words, but also it is often accompanied by social phenomena. This post is considered to be hateful because it is related to COVID-19. The slur 'Chinese virus' appeared in 2020 to refer to the coronavirus, which was first detected in the city of Wuhan in China, incited hate towards Asians (Reja 2021). This post claimed that the Jews created the virus trying to incite hate towards this group. Although this post would be identified as hateful by a human, it may not be by the machine because it has no knowledge about the relevant social phenomena.

Finally, most of the errors in HatEval are tweets that use hashtags as subject, verb, and object. For example:

(7) #Bulgaria doing it the way it should be done. #illegalaliens try to enter, #IllegalAliens are put in a #pinebox (HatEval)

These kinds of tweets lost some information because we opted to remove hashtags during text pre-processing. However, the removal resulted in higher performance as some non-hateful posts also use the same hashtags. How to handle hashtags requires further investigation.

7.1.2 Summary

From the evaluation in Section 6.1, we observe that the SVC[Multi] has overall better performance than the other classifiers despite worse F1-score in the Hackforums and Extremist forums

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

compared to BERT[Multi] and BERT[Mono], respectively. However, the overall performance is still far from perfect. The error analysis shows many limitations of the system.

It is biased towards posts related to women and cannot always recognise hateful posts against other targets. This shows the importance of having a similar distribution of hate speech categories in the training data. Because of this, there are many false negatives on HackForums and Extremist forums test sets that contain other categories that Table 2 lists. The false negatives may also be affected by the multi-platform training data in which non-hateful posts are the majority. A large amount of data with a balanced class distribution is ideal to train any classifier to achieve desirable performance. However, this is challenging because, as we analysed in Section 4.3.1, hate speech is not frequent across all the platforms.

Hate speech detection is a demanding task. One significant issue is the problem of defining hate speech and its boundary with language which is 'only' offensive. As reviewed in Section 2.1, the lack of a universal detailed hate speech definition raised many disputes. It was revealed to be a difficult task for humans during the annotation process since judging what is hateful and what is offensive, or not, is a highly subjective task based heavily on people's background, experiences and personal views. Then, it is even harder for the machine that might still suffer from a poor level of natural language understanding as indicated by some of our error analyses above.

Researchers have struggled to define the boundary between offensive language and hate speech. It is likely for a classifier to use pejorative words, that are frequent and repeated, as an indicator of hate speech. Kwok and Wang (2013) found 86% of tweets are classified as hate speech due to the presence of an offensive word. Classifiers biased towards offensive language would cause many issues. In real-life applications (e.g., Twitter), the machine would ban most of the posts because the use of pejorative words has become a social norm (Kleinman *et al.* 2009). This would raise debate related to topics such as free speech, tolerance and civics (de Gibert *et al.* 2018). A suggestion to the problem would be the incorporation of some contextual information (e.g., authors' demographics characteristics) in the classifiers because some offensive words that refer to the authors' target group would not be considered as hateful (Warner and Hirschberg 2012; Davidson *et al.* 2017), for instance. However, this would not solve the problem of the classifiers being scarcely able to recognise hate speech not containing any pejorative terms.

Finally, we found two factors that increased the difficulty of the task. As we reported in Section 4.1.3, some posts were discarded as their intention could not be understood due to the lack of context of the conversation. The reconstruction of the dialogue would help the annotation process and the training of a more robust classifier. However, it would require a large amount of data. Second, hate speech does not manifest with sole words. There are sarcastic posts in which the system does not comprehend the real intention and social phenomena (e.g., COVID-19) that remain unknown to the system. While sarcasm detection has been revealed to be a difficult task that requires further study (Parmar, Limbasiya and Dhamecha 2018), it is possible to incorporate the social phenomena by applying domain adaptation that allows classifiers to generalise to a specific target domain (Daumé III 2009).

Although hate speech classifiers may be used in real-life applications, their predictions would still need to be reviewed by human moderators, which was our motivation for exploring hate speech SE so that moderators may review the texts more closely. In semi-automated moderation, the classifier should aim for a low false negative rate and not depend only on the use of offensive language because there are many instances of hate speech not containing disparaging terms.

7.2 Hate speech SE error analysis

We conducted an error analysis to understand how BERT+span and BERT+token performed on the hate speech SE task. We selected the fold on which the two models had the best performance. The spans in bold are the ground truth spans.

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

7.2.1 BERT+span

We looked at the predicted posts that exactly match the ground truth spans. For example:

(1) **what a fucking cheating whore** (HackForums)

(2) LOL you [. . .]. Can be but **I hate gays**. (HackForums)

Post 1 is an example in which the ground truth is the entire post and the machine correctly extracted it. Example 2 shows that the system was trying to understand the post and successfully found the hateful span leaving out other parts of the text.

We also looked at other predictions that partly overlap the ground truth spans. These predictions either add more words or remove some words without losing their hateful intention. However, there are also a few in which the removal of some parts resulted in a non-hateful span. For example, the following is a hateful post against women in which the entire post should be extracted:

(3) **My country was one of the first to allow women to vote. I fucking hate my ancestors**. (Extremist forums)

The predicted span is *I fucking hate my ancestors* which makes the span not hateful anymore because the identity of their ancestors or the reason the speaker hates them is unknown. The latter can be found in the first sentence which has been removed. The removal of the first sentence is not surprising as it is not hateful. This example shows that the system failed at understanding the relationship among sentences.

Finally, we analysed posts that have multiple spans. Although some predictions show the ideal behaviour of the system (i.e., connecting all the spans), there are many cases similar to the following example:

(4) Good post. **Women are the devil's children tbh**. The Bible warned us about this. [. . .]. **Women are corrupt/easy to corrupt**. (Extremist forums)

in which the system output *Women are corrupt/easy to corrupt*, only one of the ground truth spans.

7.2.2 BERT+token

BERT+token is better for multi-span predictions. In the following example:

(5) Omfg **this bitch is so annoying** always talking shit [. . .] **you damn attention whore**. (HatEval)

the system correctly predicted the two hateful spans. However, there are only a few predictions that exactly match the ground truth spans.

Although we performed post-processing to avoid disconnected hateful spans, the problem persists. For this post:

(6) Exactly. **This mentality is why Christians are superior. They do not try to be above or better than anybody. Yet atheists & Muslims do. Which is an inferior way of acting**. (HackForums)

BERT+token extracted three spans: *anybody*, *atheists* and *do. which is an inferior way of acting*. The disconnected spans have become incomprehensible and cannot assist human moderators to determine whether the post is hateful.

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

7.2.3 Summary

Although BERT+span and BERT+token achieved prominent performance and correctly extracted hate speech, both systems have limitations.

BERT+span is better for extracting a single span because it outputs a continuous span. In terms of multiple spans, it would either correctly connect all the spans or omit some spans. The first case may deviate from the initial purpose – extract short spans to save human moderators time from reading lengthy posts. Although the second case would not be a problem for human moderators because it correctly outputs a hateful span that helps their classification, it would be an issue for researchers who want to analyse some aspect of hate speech (e.g., target analysis) because there is missing information. For multi-span extraction, BERT+token would be more prominent, but the crucial problem is the discontinuous predictions that often become incomprehensible.

In general, hate speech SE systems are likely to be used in practice. Most companies, such as Twitter, rely on human moderators to review potential harmful posts (Harrison 2019). The proposed hate speech SE models that successfully extracted hateful spans could save moderators time in reading lengthy posts and enable them to focus on accurate classification.

8. Conclusion

This work aimed at understanding hate speech in underground and extremist forums where cybercriminals and extremists communicate with each other and potentially incite crime and abuse against specific social groups. It also explored automated hate speech detection and SE systems.

We introduced a manually labelled hate speech dataset, obtained from HackForums, Stormfront and Incels.co, based on which we analysed the distribution of hate speech and the content in these forums. We found that ideologies and the restrictions of user behaviour affect the amount of hate speech. Hate speech is not prevalent on Hackforums because users focus on monetising their skills and gaining hacking knowledge. In contrast, Stormfront and Incels.co contain more than double the amount of hate speech posts compared to Hackforums. This is expected due to them being driven by hateful ideologies and having little to no content moderation.

This work provided a better understanding of hate speech in these forums. However, the problem of hate speech, in general, remains a challenging task due to the lack of a universal legal definition. This was also reflected in the data annotation. Due to the unavailability of labelled data from the three forums, the data were labelled by three human annotators, and the labelling process was complemented by active learning. The annotation process brought up a number of challenges, including differentiating between offensive language and hate speech. Additionally, the system used in active learning was unable to recognise hate speech against targets that were not present in the training data. This problem also appeared in the multi-platform hate speech classifier due to the imbalanced distribution of hate speech categories in the training data. Future work should ensure a similar distribution of hate speech categories in the training data.

We explored different classifiers trained on a combined dataset from Twitter, HackForums, Stormfront and Incels.co. The performance results have shown that hate speech classifiers would not always benefit from combining data from different platforms. However, we do note that it would benefit the research community if additional training data could be collected and released, focused on hate speech as broadly as possible rather than the subtypes of misogyny and anti-immigrant sentiment contained in HatEval, which was the seed set for our training data.

Finally, this research laid the groundwork for hate speech SE, we fine-tuned BERT and adopted two solutions, namely span prediction and sequence labelling. Both models achieved good results achieving an F1-score of at least 69%. We propose a number of improvements as further research avenues. First, it would be beneficial to create a larger dataset to train the model. Second, because we experimented with basic BERT models, future work could investigate other types of transformers that achieved state of the art in many natural language processing tasks.

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

References

Abadi M., **Agarwal A.**, **Barham P.**, **Chen J.**, **Chen Z.**, **Davis A.**, **Dean J.**, **Devin M.**, **Ghemawat S.**, **Irving G.**, **Isard M.**, **Kudlur M.**, **Levenberg J.**, **Monga R.**, **Moore S.**, **Murray D.G.**, **Steiner B.**, **Tucker P.**, **Vasudevan V.**, **Warden P.**, **Wicke M.**, **Yu Y.** and **Zheng X.** (2016). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation (OSDI'16)*, pp. 265–283.

Alshalan R. and **Al-Khalifa H.** (2020). Hate speech detection in Saudi Twittersphere: A deep learning approach. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, vol. 10. Association for Computational Linguistics, pp. 12–23.

Assimakopoulos S., **Vella Muskat R.**, **van der Plas L.** and **Gatt A.** (2020). Annotating for hate speech: The MaNeCo corpus and some input from critical discourse analysis. In *Proceedings of the 12th Language Resources and Evaluation Conference*, European Language Resources Association, pp. 5088–5097.

Badjatiya P., **Gupta S.**, **Gupta M.** and **Varma V.** (2017). Deep learning for hate speech detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, pp. 759–760.

Basile V., **Bosco C.**, **Fersini E.**, **Nozza D.**, **Patti V.**, **Rangel Pardo F.M.**, **Rosso P.** and **Sanguinetti M.** (2019). SemEval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pp. 54–63.

Bengio Y., **Ducharme R.**, **Vincent P.** and **Janvin C.** (2003). A neural probabilistic language model. *Journal of Machine Learning Research* **3**, 1137–1155.

Bhalerao R., **Aliapoulios M.**, **Shumailov I.**, **Afroz S.**, **Mccoy D.**, **Levchenko K.** and **Paxson V.** (2018). Mapping the Underground: Towards Automatic Discovery of Cybercrime Supply Chains. **16**. arXiv preprint arXiv:1812.00381.

Binny M., **Saha P.**, **Yimam S.M.**, **Biemann C.**, **Goyal P.** and **Mukherjee A.** (2021). HateXplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 14867–14875.

Bojanowski P., **Grave E.**, **Joulin A.** and **Mikolov T.** (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* **5**, 135–146.

Boser B.E., **Guyon I.M.** and **Vapnik V.N.** (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*. Association for Computing Machinery, pp. 144–152.

Bruwaene D.V., **Huang Q.** and **Inkpen D.** (2020). A multi-platform dataset for detecting cyberbullying in social media. *Language Resources and Evaluation* **54**, 851–874.

Bowman-Grieve L. (2009). Exploring "Stormfront": A virtual community of the radical right. *Studies in Conflict & Terrorism* **32**, 989–1007.

Caines A., **Pastrana S.**, **Hutchings A.** and **Buttery P.** (2018a). Aggressive language in an online hacking forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, vol. 32. Association for Computational Linguistics, pp. 66–74.

Caines A., **Pastrana S.**, **Hutchings A.** and **Buttery P.J.** (2018b). Automatically identifying the function and intent of posts in underground forums. *Crime Science* **7**, 19.

Cer D., **Yang Y.**, **Kong S.**, **Hua N.**, **Limtiaco N.**, St. John R., Constant N., Guajardo-Cespedes M., Yuan S., Tar C., Strope B. and Kurzweil R. (2018). Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, vol. 32. Association for Computational Linguistics, pp. 169–174.

Chhablani G., **Bhartia Y.**, **Sharma A.**, **Pandey H.** and **Suthaharan S.** (2021). NLRG at SemEval-2021 Task 5: Toxic Spans Detection Leveraging BERT-based Token Classification and Span Prediction Techniques. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Association for Computational Linguistics, pp. 233–242.

Chollet F. (2015). Keras. https://keras.io (accessed April 2021).

Cohn D. (2010). Active learning. In *Encyclopedia of Machine Learning*, vol. 32. USA: Springer, pp. 10–14.

Corazza M., **Menini S.**, **Cabrio E.**, **Tonelli S.** and **Villata S.** (2019). Cross-platform evaluation for Italian hate speech detection. In *CLiC-it 2019 – 6th Annual Conference of the Italian Association for Computational Linguistics*, vol. 2481.

Da San Martino G., **Barrón-Cedeño A.**, **Wachsmuth H.**, **Petrov R.** and **Nakov P.** (2020). SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. International Committee for Computational Linguistics, pp. 1377–1414.

Daumé III H. (2009). Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, pp. 256–263.

Davidson T., **Bhattacharya D.** and **Weber I.** (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, pp. 25–35.

Davidson T., **Warmsley D.**, **Macy M.** and **Weber I.** (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, pp. 512–515.

de Gibert O., **Perez N.**, **García-Pablos A.** and **Cuadros M.** (2018). Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online*. Association for Computational Linguistics, pp. 11–20.

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

Devlin J., **Chang M.**, **Lee K.** and **Toutanova K.** (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pp. 4171–4186.

Fleiss J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**, 378–382.

Geng X. and **Smith-Miles K.** (2009). Incremental learning. In *Encyclopedia of Biometrics*. USA: Springer, pp. 731–735.

Gerstenfeld P., **Grant D.** and **Chiang C.** (2003). Hate online: A content analysis of extremist internet sites. *Analyses of Social Issues and Public Policy* **3**, 29–44.

Gokaslan A. and **Cohen V.** (2019). OpenWebText Corpus. http://Skylion007.github.io/OpenWebTextCorpus (accessed May 2021).

Goldberg Y. (2015). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research* **57**, 345–420.

Gurusamy V. and **Kannan S.** (2015). Preprocessing techniques for text mining - An overview. *International Journal of Computer Science & Communication Networks* **5**, 7–16.

Harrison S. (2019). Twitter and Instagram Unveil New Ways to Combat Hate—Again. https://www.wired.com/story/twitter-instagram-unveil-new-ways-combat-hate-again/ (accessed May 2021).

Hatebase Inc. (2020). https://hatebase.org/ (accessed January 2021).

Hinton G.E., **Srivastava N.**, **Krizhevsky A.**, **Sutskever I.** and **Salakhutdinov R.** (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580.

Holpuch A. (2014). Almost 100 hate-crime murders linked to single website. https://www.theguardian.com/world/2014/apr/18/hate-crime-murders-website-stormfront-report (accessed May 2021).

Incels Wiki. (2018). https://incels.wiki/w/Main_Page (accessed May 2021).

Indurthi V., **Syed B.**, **Shrivastava M.**, **Chakravartula N.**, **Gupta M.** and **Varma V.** (2019). 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*. Association for Computational Linguistics, pp. 70–74.

Jaki S., **De Smedt T.**, **Gwóźdź M.**, **Panchal R.**, **Rossa A.** and **De Pauw G.** (2019). Online hatred of women in the Incels.me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict* **7**, 240–268.

Jasser G., **Kelly M.** and **Rothermel A.** (2020). Male supremacism and the Hanau terrorist attack: between online misogyny and far-right violence. *The International Centre for Counter-Terrorism–The Hague* **20**.

Joshi M., **Chen D.**, **Liu Y.**, **Weld D.**, **Zettlemoyer L.** and **Levy O.** (2019). SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics* **8**, 64–77.

Jurkiewicz D., **Borchmann Ł.**, **Kosmala I.** and **Graliński F.** (2020). ApplicaAI at SemEval-2020 Task 11: On RoBERTa-CRF, span CLS and whether self-training helps them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. International Committee for Computational Linguistics, pp. 1415–1424.

Kingma D.P. and **Ba J.** (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*. Conference Track Proceedings, pp. 11–20.

Kleinman S., **Ezzell M.B.** and **Frost A.C.** (2009). Reclaiming critical analysis: The social harms of 'bitch.'. *Sociological Analysis* **3**, 46–68.

Krebs B. (2017). Who Is Marcus Hutchins?. https://krebsonsecurity.com/2017/09/who-is-marcus-hutchins/ (accessed May 2021).

Kwok I. and **Wang Y.** (2013). Locate the hate: Detecting Tweets against blacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, pp. 1621–1622.

Lafferty J.D., **McCallum A.** and **Pereira F.C.N.** (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., pp. 282–289.

Landis J.R. and **Koch G.G.** (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174.

Liu Y., **Ott M.**, **Goyal N.**, **Du J.**, **Joshi M.**, **Chen D.**, **Levy O.**, **Lewis M.**, **Zettlemoyer L.** and **Stoyanov V.** (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

Mandl T., **Modha S.**, **Majumder P.**, **Patel D.**, **Dave M.**, **Mandlia C.** and **Patel A.** (2019). Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*. Association for Computing Machinery, pp. 14–17.

Mathet Y., **Widlöcher A.** and **Métivier J.** (2015). The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics* **41**, 437–479.

Mathew B., **Dutt R.**, **Goyal P.** and **Mukherjee A.** (2019). Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*. Association for Computing Machinery, pp. 173–182.

Mikolov T., **Chen K.**, **Corrado G.** and **Dean J.** (2013). Efficient Estimation of Word Representations in Vector Space. *ICLR Workshop*.

Morio G., **Morishita T.**, **Ozaki H.** and **Miyoshi T.** (2020). Hitachi at SemEval-2020 Task 11: An empirical study of pre-trained transformer family for propaganda detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. International Committee for Computational Linguistics, pp. 1739–1748.

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

Mozafari M., **Farahbakhsh R.** and **Noël C.** (2019). A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. In *International Conference on Complex Networks and Their Applications*, pp. 928–940.

Parmar K., **Limbasiya N.** and **Dhamecha M.** (2018). Feature based composite approach for sarcasm detection using MapReduce. In *International Conference on Computing Methodologies and Communication*, pp. 587–591.

Pastrana S., **Thomas D.R.**, **Hutchings A.** and **Clayton R.** (2018a). CrimeBB: Enabling cybercrime research on underground forums at scale. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, pp. 1845–1854.

Pastrana S., **Hutchings A.**, **Caines A.** and **Buttery P.** (2018b). Characterizing eve: Analysing cybercrime actors in a large underground forum. In *Proceedings of 21st International Symposium*, pp. 207–227.

Pedregosa F., **Varoquaux G.**, **Gramfort A.**, **Michel V.**, **Thirion B.**, **Grisel O.**, **Blondel M.**, **Prettenhofer P.**, **Weiss R.**, **Dubourg V.**, **Vanderplas J.**, **Passos A.**, **Cournapeau D.**, **Brucher M.**, **Perrot M.** and **Duchesnay É.** (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830.

Pennington J., **Socher R.** and **Manning C.D.** (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Pradha S., **Halgamuge M.N.** and **Tran Quoc Vinh N.** (2019). Effective text data preprocessing technique for sentiment analysis in social media data. In *11th International Conference on Knowledge and Systems Engineering (KSE)*, pp. 1–8.

Rajpurkar P., **Zhang J.**, **Lopyrev K.** and **Liang P.** (2016). SQuAD: 100000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 2383–2392.

Reja M. (2021). Trump's 'Chinese Virus' tweet helped lead to rise in racist anti-Asian Twitter content: Study. https://abcnews.go.com/Health/trumps-chinese-virus-tweet-helped-lead-rise-racist/story?id=76530148 (accessed May 2021).

Salminen J., **Hopf M.**, **Chowdhury S.A.**, **Jung S.**, **Almerekhi H.** and **Jansen B.J.** (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences* **10**(1), 1–34.

Schafer J. 2002. Spinning the Web of hate: Web-based hate propagation by extremist organizations. *Journal of Criminal Justice and Popular Culture* **9**, 69–88.

Smith K.L. (2018). Twitter Is Deleting Accounts And These Are The Words That Might Get You Suspended. https://www.popbuzz.com/internet/social-media/twitter-account-suspension-trigger-words/ (accessed May 2021).

Sparck Jones K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* **28**, 11–21.

Stricker G. (2014). The 2014 #YearOnTwitter. https://blog.twitter.com/official/en_us/a/2014/the-2014-yearontwitter.html (accessed May 2021).

Tranchese A. and **Sugiura L.** (2021). "I Don't Hate All Women, Just Those Stuck-Up Bitches": How incels and mainstream pornography speak the same extreme language of misogyny. *Violence Against Women* **27**, 2709–2734.

UN. (2020). https://www.un.org/en/genocideprevention/documents/UN

Vaswani A., **Shazeer N.**, **Parmar N.**, **Uszkoreit J.**, **Jones L.**, **Gomez A.N.**, Kaiser ŁL. and Polosukhin I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., pp. 6000–6010.

Vu X., **Vu T.**, **Tran M.**, **Le-Cong T.** and **Nguyen H.T.M.** (2020). HSD Shared Task in VLSP Campaign 2019:Hate Speech Detection for Social Good. arXiv preprint arXiv:2007.06493.

Vu A.V., **Wilson L.**, **Chua Y.T.**, **Shumailov I.** and **Anderson R.** (2021). ExtremeBB: Enabling Large-Scale Research into Extremism, the Manosphere and Their Correlation by Online Forum Data. arXiv preprint arXiv:2111.04479.

Warner W. and **Hirschberg J.** (2012). Detecting hate speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, pp. 19–26.

Waseem Z., **Davidson T.**, **Warmsley D.** and **Weber I.** (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, pp. 78–84.

Waseem Z. and **Hovy D.** (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, pp. 88–93.

Williams M.L., **Burnap P.**, **Javed A.**, **Liu H.** and **Ozalp S.** (2020). Hate in the machine: Anti-black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *British Journal of Criminology* **60**, 93–117.

Wolf T., **Debut L.**, **Sanh V.**, **Chaumond J.**, **Delangue C.**, **Moi A.**, **Cistac P.**, **Rault T.**, **Louf R.**, **Funtowicz M.**, **Davison J.**, **Shleifer S.**, **von Platen P.**, **Ma C.**, **Jernite Y.**, **Plu J.**, **Xu C.**, **Le Scao T.**, **Gugger S.**, **Drame M.**, **Lhoest Q.** and **Rush A.** (2020). Transformers: State-of-the-Art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, pp. 38–45.

Wulczyn E., **Thain N.** and **Dixon L.** (2016). Ex Machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web (WWW'17)*. International World Wide Web Conferences Steering Committee, pp. 1391–1399.

https://doi.org/10.1017/S1351324922000262 Published online by Cambridge University Press

Zhang Z., **Robinson D.** and **Tepper J.A.** (2018). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In *Proceedings of The Semantic Web*, pp. 745–760.

Zhu Y., **Kiros R.**, **Zemel R.**, **Salakhutdinov R.**, **Urtasun R.**, **Torralba A.** and **Fidler S.** (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, pp. 19–27.

Cite this article: Zhou L, Caines A, Pete I and Hutchings A (2023). Automated hate speech detection and span extraction in underground hacking and extremist forums. *Natural Language Engineering* **29**, 1247–1274. https://doi.org/10.1017/S1351324922000262