Natural Language Processing (2025), 1–17
doi:10.1017/nlp.2024.62

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Multiclass hate speech detection with an aggregated dataset

Sinéad Walsh[1] and Paul Greaney[1,2]

[1]Department of Computing, Atlantic Technological University, Letterkenny, Co. Donegal, Ireland and [2]Centre for Mathematical Modelling and Intelligent Systems for Health and Environment (MISHE), Atlantic Technological University, Letterkenny, Co. Donegal, Ireland
Corresponding author: Paul Greaney; Email: paul.greaney@atu.ie

(Received 13 November 2023; revised 23 October 2024; accepted 10 December 2024)

Abstract

Detecting and removing hate speech content in a timely manner remains a challenge for social media platforms. Automated techniques such as deep learning models offer solutions which can keep up with the volume and velocity of user content production. Research in this area has mainly focused on either binary classification or on classifying tweets into generalised categories such as *hateful*, *offensive*, or *neither*. Less attention has been given to multiclass classification of online hate speech into the type of hate or group at which it is directed. By aggregating and re-annotating several relevant hate speech datasets, this study presents a dataset and evaluates several models for classifying tweets into the categories ethnicity, gender, religion, sexuality, and non-hate. We evaluate the dataset by training several models: logistic regression, LSTM, BERT, and GPT-2. For the LSTM model, we assess a range of NLP features using a multi-classification LSTM model, and conclude that the highest performing feature combination consists of word *n*-grams, character *n*-grams, and dependency tuples. We show that while more recent larger models can achieve a slightly higher performance, increased model complexity alone is not sufficient to achieve significantly improved models. We also compare this approach with a binary classification approach and evaluate the effect of dataset size on model performance.

Keywords: machine learning; text classification

1. Introduction

Social media has revolutionised the way people communicate and express themselves, enabling people and communities to connect across the world. However, an indisputable drawback of this communication is the use of these platforms to propagate hate and prejudice against individuals due to their protected characteristics.

The definition of hate speech varies across countries, organisations, and studies and is one of the main challenges faced by this area of research (Seemann *et al.* 2023). This has led to issues surrounding dataset annotation, in which annotators have differing opinions on what constitutes hate speech. For example, the American Bar Association states that hate speech is protected by the First Amendment, except if inciting violence (Wermiel 2018). In Ireland, it is an offence to behave or communicate in a threatening, abusive, or insulting way with the intent to instigate hatred against an individual or a group of individuals due to their protected characteristics. These protected characteristics include colour, nationality, national origins, race, religion, sexual orientation, or membership of the travelling community (*Prohibition of Incitement to Hatred Act*, 1989). However, in November 2022 a new bill (*Criminal Justice (Incitement to Violence or Hatred*

and Hate Offences) Bill, 2022) was proposed to update these protected characteristics to include gender, gender identity, and disability, along with creating new hate crime laws. If passed, this definition will then somewhat match the list of characteristics included in the hateful conduct policies of large social media platforms. For example, Twitter's policy defines hateful conduct as a direct attack against other individuals based on their age, caste, disability, disease, ethnicity, gender, gender identity, national origin, race, religion, or sexual orientation (Twitter Inc 2023). Furthermore, other definitions differ on the condition of whether hate speech is hate directed at a group or an individual (MacAvaney *et al.* 2019). In light of these discrepancies, an all-encompassing definition could describe hate speech as insulting, abusive, or threatening language that incites hatred, prejudice, or violence against an individual or group of individuals on account of their protected characteristics. These protected characteristics include age, disability, disease, caste, gender, gender identity, sexual orientation, ethnicity, race, national origin, religion, or membership of an indigenous or ethnic community.

The spread of hate on online platforms is often aided by the anonymity these platforms provide to users, enabling them to spread hate with no 'real-world' consequences for them outside of the virtual environment. However, the emotional consequences for their victims are undeniable, including fear, anger, anxiety, depression, the development of a defensive attitude towards the perpetrator's group, and changes in how they use social media platforms (Williams 2019; Saha, Chandrasekharan, and De Choudhury 2019). The availability of social media means that hate speech can affect people anywhere at any time, even in their own homes, which could be considered a safe space for people in the case of offline hate speech. Furthermore, communicating on social media has become a part of many people's daily lives, and therefore, avoiding these platforms or turning off their phone is often not a viable option, preventing them from escaping online hate. A qualitative study by Ofcom and Traverse (2023) into the impact of online hate on those who have experienced it, reinforces these findings, with participants reporting an observed increase in online hate speech around key events. They also reported feeling embarrassment, particularly when the hate occurred in online spaces visible to friends and family, and disappointment at the lack of repercussions for the perpetrator, despite repeatedly reporting them. Other tactics to avoid online hate involved self-censoring, avoiding certain hateful spaces, and attempting to challenge the online behaviour, which often led to feelings of burnout. Furthermore, some participants reported feeling fear due to online anonymity, which makes it possible that their attacker may be known to them, and as a result, felt unsafe in public spaces and became distrusting of others offline.

The effects of online hate speech are not solely limited to the victims. In terms of hate speech producers, exposure to other online hate speech can reaffirm their hateful views and values, thus encouraging them to further participate in the propagation of online hate. In some cases, exposure to online hate can contribute to the radicalisation of hateful individuals, possibly even motivating them to take real-world violent action in the form of hate crimes (Hassan *et al.* 2022). Suspects involved in many recent hate-motivated terror attacks have been shown to have an extensive record of hateful online commentary across various social media platforms, thus supporting the idea that online hate can be a precursor to offline, often violent, acts of hate (Alnazzawi 2022). Furthermore, although online bystanders are not targeted by the hate and may not agree with the hateful ideologies expressed, studies show that they can also be affected by it. A study by Soral *et al.* (2018) found that witnesses of online hate speech can become desensitised to it, in that they no longer perceive it to be offensive or hateful, which lowers their sympathy towards the victims. This ultimately begins to shape their perception of the victimised group, resulting in the development of prejudices against them and even the fostering of support for radical ideologies and policies. A study by Pluta *et al.* (2023) reiterates this by reporting that exposure to hate speech for only fifteen minutes can reduce people's empathetic brain response towards the pain and suffering of others, regardless of the type of hate. Both of these studies demonstrate that bystander exposure to online hate speech without any initial support for the ideologies expressed contributes to social division,

https://doi.org/10.1017/nlp.2024.62 Published online by Cambridge University Press

thus emphasising the need for hateful content to be quickly removed from platforms to protect members of victimised groups and stop the propagation of hate.

Policing this behaviour to remove this content in a timely manner remains a challenge for social media platforms due to the volume and velocity of data being produced by users (Burnap and Williams 2016). While platform policies, such as Twitter's hateful conduct policy (Twitter Inc 2023), encourage the reporting of such content, this is not an adequate solution as it often does not result in any repercussions for the perpetrator, much to the disappointment of victims, as expressed above. Furthermore, users that are banned can create new accounts to continue to propagate hate speech due to the anonymity that platforms provide. Moderation of hate speech is also hampered by the limitations associated with keyword-based detection approaches, which do not account for the context in which words are used, and therefore, may fail to identify hate speech that is not included in its repository of hateful phrases. Furthermore, users may evade keyword-based detection by obscuring words, such as changing their spelling slightly or replacing letters with similar-looking symbols or numbers (Kovács *et al.* 2021). The solution to these limitations is the use of automated techniques, such as deep learning models, which are less reliant on keywords or phrases and can keep up with the volume and velocity of user content production. Such techniques can effectively remove hateful content across all the protected characteristics in a timely manner, to protect the well-being of all users and prevent social division.

Past research in this area has mainly focused on classifying hate speech datasets into categories such as *hate speech*, *offensive*, or *neither*. However, there is limited research into its classification into the category of hate or group it is directed at. This study improves on the performance of existing multi-classification studies, by using different NLP feature combinations with a deep learning long short-term memory (LSTM) model to classify tweets into the categories of ethnicity, gender, religion, sexuality, and non-hate. Thus, in addition to creating an aggregated multiclass dataset, this study aims to answer the following research questions: (a) How does a multiclass dataset affect online hate speech classification performance? (b) How does dataset size influence performance in this scenario?

2. Related work

A wide range of research has been conducted in the area of online hate speech detection. Much of this research has focused on the relatively simple classification of content into categories such as *hate*, *offensive* or *neither*. Some of this research has concentrated on a particular type of hate, such as the study by Vidgen and Yasseri (2020), which evaluated different NLP features, including sentiment scores, POS & NER tags, and GloVe embeddings, across multiple machine learning models to classify tweets into the classes *weak Islamophobia*, *strong Islamophobia*, and *non-Islamophobia*. Some studies have also included multiple categories of hate in their research, such as the study by Qureshi and Sabih (2021), who created a dataset for each of 10 hate categories covering race, religion, ethnicity, gender, and sexual orientation, and similarly assessed multiple NLP feature combinations across a range machine learning models to classify each of these datasets into hate and non-hate. Others included multiple categories within a single simple classification model.

Limited research is available on the classification of hate speech into the category of hate or the targeted protected characteristic, as explored in this study. Waseem and Hovy (2016) conducted a study in this area, in which they collated a dataset of tweets and labelled them as *racism*, *sexism* and *neither*. They then extracted a range of features including user gender, length of tweets, length of user description, character n-grams, word n-grams, and location, and used a grid search with a logistic regression model to assess different combinations of these features. This revealed that using character n-grams up to length 4 and user gender as features achieved the best model performance, with an F1 score of 0.7389. This study did not provide an in-depth analysis of

https://doi.org/10.1017/nlp.2024.62 Published online by Cambridge University Press

results; however, they did mention that they prevented the model from relying on specific hateful keywords by ensuring that they included potentially offensive phrases in the *neither* category. However, they also decided to maintain the category imbalance of approximately 17% racism and 29% sexism in the data in order to mimic the real world, where hate speech is a limited occurrence. A more balanced dataset may have improved their performance by allowing the model to learn further patterns to distinguish between each of the categories.

Waseem (2016a) improved this study by comparing the effect of annotation on classification performance. In this study, he sampled and extended the dataset from the previous study, adding a fourth category for data that contained both racism and sexism. This dataset was then annotated by a set of relevantly knowledgeable people (experts) and by CrowdFlower workers (non-experts) for comparison. Again, a range of NLP feature combinations were assessed, in which it was found that different features performed better with the expert and amateur annotations. However, this study's best multi-category hate classification model performed worse than the previous study, with an F1 score of 0.5343 compared to 0.7389, respectively. This was reportedly due to the model failing to identify the minority classes (*racism* and *both*), and due to false positives in both expert and amateur annotations. The involvement of multiple annotators may make annotator bias less likely, but this alone is not sufficient to prevent bias. The authors suggest that it is possible that there was annotation bias towards labelling hate in the previous study by Waseem and Hovy (2016), as the authors were the only annotators involved, followed by a one-person review. This is supported by the analysis of the overlapping data between the two studies, in which there was a low agreement between this study's annotations and those by Waseem and Hovy (2016). Unfortunately, while this study provides a breakdown of comparison between the amateur and expert annotations for each feature and each feature set, it does not provide a clear performance comparison breakdown between this study and the previous study. Therefore, it is unclear how exactly these two studies were compared with the inclusion of the fourth class. Furthermore, the exact type of model used in this study was not explicitly mentioned.

Another study by Gambäck and Sikdar (2017) aimed to advance this study by using a CNN to classify the same tweets into racism, sexism, both, or none. Word embeddings for the CNN were generated using Word2Vec and through random vectors, along with character *n*-grams, which were assessed both individually and combined via the CNN. This was then compared to the logistic regression model using character *n*-grams by Waseem and Hovy (2016). The Word2Vec embeddings yielded the highest performance with an F1 score of 0.7829, compared to the logistic regression F1 score of 0.7389 achieved by Waseem and Hovy (2016). While the study is compared with Waseem and Hovy (2016), a clear performance comparison breakdown between the two is not given. Therefore, it is unclear how exactly these two studies were compared with the inclusion of the fourth class. A more appropriate comparison may compare this study's CNN model with Word2Vec embeddings, which achieved an F1 score of 0.7829, to the best-performing model by Waseem (2016a), which achieved an F1 score of only 0.5343 using the expert annotations. Similar to Waseem (2016a), this study also reported that their model was unable to identify tweets in the *both* category and struggled with the racism category due to the class imbalance. This highlights the importance of using balanced datasets in multi-category hate speech classification, despite the limited occurrence of hate speech on online platforms.

Other studies have focused on identifying the target of social media posts, such as Zampieri *et al.* (2019a), which introduced the widely used Offensive Language Identification Dataset (OLID) and evaluated support vector machine, BiLSTM and CNN models on it. This task of target identification was similarly considered by Sachdeva *et al.* (2022). Similar to our study, they used the Measuring Hate Speech dataset by Kennedy *et al.* (2020a) to compare the performance of a pre-trained BERT, RoBERTA, and Universal Sentence Encoder in classifying social media comments into 8 identity categories and a further 12 sub-categories. The main categories included age, disability, gender, national origin, political ideology, race, religion, and sexual orientation. The multi-label output of these models was a binary indicator for each target, thus allowing posts

https://doi.org/10.1017/nlp.2024.62 Published online by Cambridge University Press

to have multiple target labels. This study found that the RoBERTA model achieved the best performance with an overall F1 score of 0.647, while the BERT and Universal Sentence Encoder models followed behind with F1 scores of 0.610 and 0.529, respectively. Across the individual categories, the RoBERTA model attained F1 scores between 0.3 and 0.85, with the lower incidence categories producing the lower scores. Unfortunately, while these models were trained using a hate speech dataset, the study solely focused on identifying the target of the posts without determining if the posts were hateful or not, and therefore, cannot be considered a hate speech detection study and is not comparable to our study.

Identification of the target of the social media posts has also been considered, such as in the study by Aditya *et al.* (2022), which also centred on using a dataset included in our study. They aimed to predict whether comments were hateful, offensive or neither, along with predicting the community targeted in the comment using the HaTeXplain dataset (Mathew *et al*. 2021a). For the multi-target classification task, they trained five different models using TF-IDF representations of the posts, each model allowing multiple target-label predictions per post. Their One-Versus-Rest model involved training a binary logistic regression classifier for each target category, thus failing to incorporate the relationships between the target labels in its predictions. Similarly, their Binary Relevance model involved creating a group of independent binary classifiers, which were each trained using target-specific datasets. Unlike these models, the remaining models considered the possible correlations between labels in their predictions. Their Label Powerset model created new labels, one for each possible combination of targets, while their Classifier Chain model, consisting of a model for each target class, utilised the predictions of the previous classifiers in the chain when making the current classifier's predictions. Finally, their multi-label KNN utilises the labels of nearest neighbours in the training set to predict a set of target labels for each social media post. Out of these five models, the One-Versus-Rest model performed the best with a global accuracy of 0.9097, while the others failed to achieve accuracies above 0.4. Unfortunately, it is likely that the One-Versus-Rest model's high accuracy may be attributed to the imbalance in the dataset, with a recall of 0.529 indicating that it failed to detect some of the targeted communities correctly. Furthermore, the study treated the hate speech detection and target identification as two separate tasks and did not report on the model's ability to detect hate speech. Therefore, its performance as a hate speech detection model is unknown, making it incomparable to our study.

In terms of a more advanced and multi-faceted approach to the multi-classification of hate speech, the study by Ousidhoum *et al.* (2019) developed a range of models by combining multiple aspects of online hate speech and multiple languages. They compiled a dataset of English, French and Arabic tweets, which were then labelled for each of the five aspects: the directness of the tweet; whether the tweet was hateful, abusive, offensive, disrespectful, fearful, or normal; the protected characteristic it discriminated against; the targeted group; and how the annotators felt about its content.

For modelling, they used logistic regression with bag-of-words (BOW) features as a baseline, along with a BiLSTM with one hidden layer. Within this they created single-task-single-language, single-task-multi-language, multi-task-single-language, and multi-task-multi-language models. For multi-aspect and multi-language settings, they utilised and evaluated multi-task learning, in which one task may help a related task via weight sharing. Overall, they found that utilising multi-task learning improved model performance when compared to their single-task models, and noted that the BiLSTMs generally outperformed the baseline logistic regression models across the tasks. In terms of the classifying the targeted protected characteristic, their single-task English language BiLSTM model developed for this multi-category classification achieved an F1 score of only 0.42, significantly lower than the previously discussed studies. Furthermore, their multi-task English language model and their baseline BOW-logistic regression model performed no better, with both models producing an F1 score of 0.41. However, it is worth noting that this is not an entirely fair comparison as multi-category hate classification was not the main focus of this study. BiLSTMs

https://doi.org/10.1017/nlp.2024.62 Published online by Cambridge University Press

have also been used for binary hate speech classification, such as Fazil *et al.* (2023) who developed a multi-channel convolution BiLSTM with attention to classify content as hate or non-hate.

A good overview of hate speech detection and classification in the era of LLMs is given by Zampieri *et al.* (2023b). While most of the tasks and studies considered therein again involve binary classification, several address forms of multiclass classification, including type of misogyny (Fersini *et al.* 2022), type of sexism (Kirk *et al.* 2023), and class of target—person, group, other, expletive language (Taulé *et al.* 2021).

3. Dataset construction

An extensive review of publicly available hate speech datasets was carried out. Notable datasets such as OLID, as used in the HatEval & OffensEval studies (Basile *et al.* 2019; Zampieri *et al.* 2019b), and its expanded version SOLID (Rosenthal *et al.* 2020) were not used due to the absence of protected characteristic target labels, and since relabelling datasets such as these were outside the scope of this study. Other well-known datasets which contained multiclass target labels, such as CONAN (Chung *et al.* 2019) and ConvAbuse (Curry, Abercrombie, and Rieser 2021), were also not included as they contained synthetic data and data containing hate towards conversational AI systems, respectively, which may not accurately reflect real-world hate towards people. Another challenge faced during dataset construction included the unavailability of the Twitter API, which prevented the use of datasets which published the tweet IDs without the tweet text, such as the dataset used in the study by ElSherief *et al.* (2018). Thus, due to the limited research in the classification of online hate into the category or targeted group, and the utilisation of the same dataset in multiple studies as discussed in the previous section (Waseem and Hovy 2016; Waseem 2016a; Gambäck and Sikdar 2017), there is a lack of multi-category hate speech datasets across multiple protected characteristics. Four existing datasets containing hate across multiple protected characteristics were therefore relabelled and combined into one corpus of tweets labelled with the hate categories ethnicity, gender, sexuality, religion, and non-hate. The resulting combined dataset could not be published as some of the datasets used have not been made available for this purpose by their respective authors.

3.1 Waseem hate speech dataset

The first dataset chosen was from the study by Waseem (2016a), in which the effect of annotation on classification performance was compared. This study sampled a dataset of tweets from the previous study (Waseem and Hovy 2016), which contained the hate categories *racism*, *sexism*, and *neither*, and extended it via the collection of further tweets and the addition of a fourth category for tweets that contained both racism and sexism. This dataset was then annotated by a set of relevantly knowledgeable people (experts) and by CrowdFlower workers (non-experts) for comparison (Waseem 2016b). We use the expert annotations in our study to ensure that a strict definition was used and maintained as criteria for labelling the tweets. As our aim is to distinguish between hate towards multiple types of protected characteristics, tweets labelled with both racism and sexism were removed. Upon inspection of the tweets, it became apparent that the hateful context of the 93 racism-labelled tweets relied heavily on their hateful hashtags, which would be later removed during data cleaning, thus removing the hateful content within them. Therefore, these 93 tweets were removed to avoid negatively impacting model performance. The remaining tweets were then relabelled in accordance with our selected categories for hate speech, so that tweets containing sexism were relabelled as *gender* and tweets in the *neither* category were relabelled as *non-hate*. This resulted in a dataset of 5,580 non-hate- and 891 gender-labelled tweets. To reduce the class imbalance in this dataset, the non-hate tweets were shuffled before a sample was taken, of size equal to the number of tweets in the gender category, thus balancing the number of non-hate tweets to 891.

https://doi.org/10.1017/nlp.2024.62 Published online by Cambridge University Press

Table 1. Breakdown of dataset by source and class

Dataset	Ethnicity	Gender	Religion	Sexuality	Non-hate	Total
Waseem	0	891	0	0	891	1,783
Berkeley	1,493	1,441	715	927	1,089	5,443
HateXplain	339	14	89	101	136	679
Large-scale hate speech	405	0	328	0	366	1,099
	2,015	2,347	1,132	1,028	2,482	9,004

3.2 Berkeley measuring hate speech dataset

The second dataset was used in the study by Kennedy *et al*. (2020a), which involved modelling online hate speech as a continuous spectrum. The study proposed that classifying online hate speech into discrete categories failed to capture the underlying spectrum and that assigning hate speech as a continuous score would allow a moderation policy to be imposed based on the severity via hate score thresholds. This study collected hate speech from Twitter, YouTube, and Reddit across the protected characteristics race, religion, ethnicity, nationality, gender, gender identity, sexual orientation, and disability. These tweets and comments were then labelled by an annotator, who identified the targeted protected characteristic and gave them ordinal ratings across 10 different components based on a given hate scale. A deep learning model was then trained, which incorporated annotator bias, to predict these ordinal labels for the tweets and comments. The predicted ordinal labels were then converted to a continuous score using the Rasch Item Response Theory. This continuous score was given with the dataset, in which they reported that a score of greater than or equal to 0.5 is considered hate speech (Kennedy *et al*. 2020b).

Our study uses the tweets from this dataset, removing the Reddit and YouTube comments. The dataset has a column of discrete hate speech labels with values 0,1,2, and a column of continuous hate speech scores. In the absence of an explanation for the discrete categories, hateful tweets were identified as those with a hate speech score greater than or equal to 0.5, and those with a score less than this were labelled as *non-hate*. The hateful tweets were then labelled using the annotator-identified target characteristics. Tweets containing multiple identified characteristics were removed to avoid negatively impacting model performance, with two exceptions. Race and origin targeting tweets were combined and labelled as *ethnicity*, since these categories were not explicitly distinguished in the other datasets. Tweets targeting sexuality were labelled as *sexuality* if the identified target was sexuality only or if sexuality and gender were identified and the tweet contained a hateful sexuality-based insult. This was carried out to increase the number of tweets in this category as it was discovered that many of the tweets identified as targeting sexuality, were also identified as targeting gender due to the hate being aimed at a specific combination of gender and sexual orientation. The breakdown of the resulting dataset is given in Table 1, where again the number of non-hateful tweets has been balanced with the average number of tweets in the other categories.

3.3 HateXplain dataset

The HateXplain dataset was used in the study by Mathew *et al*. (2021a) and was made publicly available as the first benchmark dataset covering multiple aspects of online hate speech. This study involved demonstrating that incorporating human rationale into deep learning models can improve the explainability of binary hate classification across characteristics, including race, religion, gender, and sexual orientation, by enabling the model to make more human-like decisions.

https://doi.org/10.1017/nlp.2024.62 Published online by Cambridge University Press

To do this, they collated a dataset of tweets and comments from the platform Gab, before asking Amazon Mechanical Turk workers to not only label the tweets as hate or non-hate but also select the parts of the tweet that led them to their decision and identify the community that the hate targeted (Mathew *et al*. 2021b).

Again, the tweets from this dataset were retained, before hateful, offensive, and non-offensive tweets were identified via the mode of the supplied annotator votes. Similarly, the target of each tweet was determined using the mode of the target votes. Tweets which were determined as targeting protected characteristics which were not included in this study were removed from the dataset. Offensive and non-offensive tweets were then relabelled as *non-hate* before the remaining hateful tweets were relabelled based on their voted target. Tweets whose voted target was Africans, Caucasians, Refugees, Asians, Arabs, Hispanic, or Indigenous people were relabelled as *ethnicity*. Those targeting Islamic, Jewish, Hindu, Christian, or Buddhist individuals were relabelled as *religion*, tweets targeting people based on sexuality were relabelled as *sexuality*, and those targeting men or women were relabelled as *gender*. The breakdown of the resulting dataset is given in Table 1, where again the number of non-hateful tweets has been balanced with the average number of tweets in the other categories. Unlike the other datasets, these tweets were provided in tokenised form, and so the tweets were detokenised using the NLTK library to simplify the process of data cleaning and preprocessing after dataset integration.

3.4 Large-scale hate speech dataset

The fourth and final dataset we integrated was used in the study by Toraman *et al*. (2022b), which collated a dataset of English and Turkish tweets across the categories religion, gender, race, politics, and sport. They then classified these tweets into hate, offensive or neither and evaluated the decay and recovery rates in the cross-domain transfer of hate in which the model is trained using one category and tested on the others. This analysis found that recovery rates are generally high across domains with an average recovery of 96% across the English tweets. Here we use version 2 of their dataset, which only includes tweets with an annotator agreement greater than 80%, resulting in a more reliably annotated dataset (Toraman *et al*. 2022a).

The Turkish tweets were removed from the dataset, retaining English tweets only. Offensive and normal tweets were relabelled as *non-hate*, while the remaining hateful tweets were relabelled using their topic label. Hateful sports and politics tweets were removed from the dataset, as these are not protected characteristics, and therefore do not constitute hate speech. Gender-targeted tweets were also removed from the dataset, as these tweets combined both gender-based hate and sexuality-based hate into one category, with no effective way to separate them. Like the previous datasets, hate speech race-labelled tweets were relabelled as *ethnicity*, while tweets with the religion topic label were relabelled as *religion*. The breakdown of the resulting dataset is given in Table 1, where again the number of non-hateful tweets has been balanced with the average number of tweets in the other categories.

3.5 Dataset integration

The four datasets were then integrated into one complete dataset, containing the tweet IDs, the tweets, and the hate labels. The majority of the tweets came from the Berkeley Measuring Hate Speech dataset, making up 60.5% of the tweets, while the HateXplain dataset contributed the least at 7.5%. This final dataset contained 2,347 gender, 2,015 ethnicity, 1,132 religion, 1,028 sexuality, and 2,482 non-hate tweets before cleaning and preprocessing. Thus, the dataset was slightly class-imbalanced, which could cause a possible decrease in classification performance in the religion and sexuality categories.

The number of non-hate tweets in each dataset was balanced to enhance the model's ability to learn hate and non-hate specific patterns, similar to the studies by Qureshi and Sabih (2021);

https://doi.org/10.1017/nlp.2024.62 Published online by Cambridge University Press

Mathew *et al.* (2021a), and Kennedy *et al.* (2020a). This avoids any negative impact on classification performance caused by class imbalance, particularly when the number of examples in the hateful classes is low, as evidenced in the study by Toraman *et al.* (2022b).

The non-hate class in each dataset contained a large number of tweets as this class combined the offensive and non-hate categories in three of the datasets, while the fourth was imbalanced to begin with. Balancing was implemented before integration to ensure that the non-hate category contained a proportional amount of tweets from every dataset. This was carried out to enable the model to distinguish between hateful and non-hateful tweets containing content about similar topics, as data is often collected based on specific events, keywords, and phrases. This encourages the model to learn the hateful patterns and discourages it from learning topic-related keywords, which may result in false positives.

Once the individual datasets were integrated into one complete dataset, the tweets were cleaned and preprocessed to remove user mentions, URLs, hashtags, emojis, twitter reserved phrases, numbers and symbols. This was done using the Tweet Preprocessor library. Contracted phrases, such as 'don't', 'she'd', or 'you're', were converted to their expanded form, thus ensuring consistency in phrasing across the tweets enabling the model to recognise them as part of a pattern. As the final cleaning steps, duplicate, null, and blank tweets were removed from the dataset. When removing duplicate tweets, only those labelled as *ethnicity*, *gender*, or *non-hate* were considered to maintain the class balance in the dataset. Although the remaining religion and sexuality classes contained duplicates, these were retained to produce the same effect as the random oversampling technique, which would have been required to rebalance the dataset if all duplicates were removed. Duplicates which appeared in the validation and test sets were removed.

The tweet labels were then encoded for use with the model, with 0 indicating non-hate, 1 indicating ethnicity, 2 indicating gender, 3 indicating religion, and 4 indicating sexuality. Named entity recognition (NER) was then applied to the full dataset of tweets before the dataset was split into 80% training, 10% validation, and 10% test data using stratified sampling to maintain an equal proportion of classes in each split. The resulting training data contained 7,511 tweets, while the validation and test data both contained 939 tweets.

4. Feature extraction

To evaluate our dataset, we extract a range of features from the dataset using NLP techniques as follows.

1. Named Entity Recognition (NER): Stanford's English 3-class caseless model with distributed similarity-based features was used to tag entities within the tweets (Finkel, Grenager, and Manning 2005; Bosu 2018), which included the entity types location, person, and organisation. The NER tags isolated from each tweet were then encoded for use as an input to the model using SciKit-Learn's CountVectorizer.
2. Character & Word *N*-Grams: Character *n*-grams of lengths 1–4 and word *n*-grams of lengths 1–3 were extracted using SciKit-Learn's TF-IDF Vectoriser with a maximum document frequency of 0.6. This means that *n*-grams with a document frequency greater than this were excluded from the TF-IDF matrix, since they are likely not contributing to any hate pattern. Due to the way in which TF-IDF matrices are computed, if a *n*-gram does not appear in a tweet, the corresponding TF-IDF weight is zero. Thus, these matrices are generally sparse. Therefore, latent semantic analysis (LSA) was applied to the TF-IDF matrices to reduce dimensionality, retaining 95% of the cumulative explained variance in the character *n*-grams and 73.8% in the word *n*-grams.
3. Part of Speech (POS) Tagging: This was carried out using the Stanza library, following which the tag sequences were converted to TF-IDF matrices using SciKit-Learn's TF-IDF

https://doi.org/10.1017/nlp.2024.62 Published online by Cambridge University Press

Vectorizer. Since there were only 16 possible tags for each tweet, dimensionality reduction was not required.

4. Dependency Parsing: This was carried out using the Stanza library, before the resulting head word-relationship tuples were converted into unigrams to represent them as TF-IDF matrices. LSA was then applied to reduce matrix dimensionality, retaining 95% of the cumulative explained variance in the data.
5. Sentiment Scoring: Valence Aware Dictionary and sEntiment Reasoner (VADER), a lexicon rule-based approach specifically designed for social media text, was used for sentiment scoring. VADER's Sentiment Intensity Analyser was used to return a compound intensity score between −1 and 1 for each of the tweets in which scores close to −1 indicate a negative sentiment and scores close to 1 indicate a positive sentiment. These scores were then standardised to between 0 and 1 using SciKit-Learn's Standard Scaler to match the scale of the other features, which did not contain negative values.

5. Results & discussion

5.1 LSTM model

We first train and evaluate a long short-term memory network (LSTM) for the multi-category hate classification of tweets. The first layer of an LSTM is usually an embedding layer, which converts the input sequences to learnable continuous representations which capture the token significance and semantic relationships used to make predictions. Here, however, these continuous representations were captured during NLP feature extraction. Different combinations of these features were evaluated during feature selection, to determine which of them provide the best-performing representations for model classification. Therefore, the embedding layer was not required, and the first layer of this LSTM is the LSTM layer itself.

Feature selection was carried out using an LSTM model consisting of one LSTM layer with a width of 16 nodes. Each feature combination was trained and evaluated for 15 epochs, recording the training and validation loss and accuracy at every epoch.

Word *n*-grams were used as a baseline feature, to which one feature was added to the combination at a time in the order of character *n*-grams, dependency tuples, POS tags, NER tags, and sentiment scores. Each additional feature's impact on model performance was evaluated using the maximum validation accuracy within the 15 epochs. If the additional feature improved the model's maximum validation accuracy, the feature was retained in the combination, to which the next feature was added. If the additional feature reduced the model's maximum validation accuracy, it was removed from the combination before the next feature was added. Thus, the best combination of features was determined to be those which produced the highest maximum validation accuracy within 15 epochs of training. The maximum validation accuracy was chosen, as opposed to the final validation accuracy at epoch 15, to mitigate the impact of overfitting. The results of this process are given in Table 2 and Figure 1, and show that word *n*-grams, character *n*-grams, and dependency tuples contribute to increasing model accuracy, while POS tags, NER tags, and sentiment scores do not lead to any model improvement.

Following feature selection, the model was tuned using the chosen input feature combination by conducting a grid search over the range of parameters shown in Table 3. Using the results of the relatively simple feature selection model as a starting point, different sets of hyperparameters, varying the dropout probability, number of layers, number of nodes, and number of epochs were evaluated across a total of 360 different configurations. Within this, overfitting was addressed via dropout and early stopping techniques, while the model complexity was tuned with the aim of finding the set of hyperparameters which maximised the model's validation accuracy.

https://doi.org/10.1017/nlp.2024.62 Published online by Cambridge University Press

Table 2. Feature selection results

Feature set	NLP features	Max validation accuracy
1	Word *N*-grams	0.6994
2	Word & character *N*-grams	0.7091
3	Word & character *N*-grams, dependency tuples	0.7112
4	Word & character *N*-grams, dependency tuples, POS tags	0.7069
5	Word & character *N*-grams, dependency tuples, NER tags	0.6994
6	Word & character *N*-grams, dependency tuples, sentiment scores	0.7037

Figure 1. Feature selection—feature set 3 training & validation graphs.

The final trained model was then evaluated using unseen test data, as shown in the confusion matrices in Figure 2 and evaluated via the accuracy, macro-averaged precision, recall, and F1 score metrics shown in Table 4.

These scores are broken down at a class level in the confusion matrices in Figure 2, which shows that the model achieved the highest classification performance in the sexuality category, with only 12% incorrect predictions in this class, despite it containing the lowest number of tweets in the dataset. Interestingly, although the non-hate category contained the highest number of tweets, it was the category with which the model struggled most, with 41% incorrect predictions. Most of the non-hate misclassifications occurred between the non-hate and gender classes, closely followed by the ethnicity class. These misclassifications may stem from the model struggling to distinguish between offensive tweets in the non-hate class and hateful tweets in the other classes.

5.2 Other models

Several other models were also evaluated to benchmark their performance on our dataset. Using Scikit-learn, a standard logistic regression model was evaluated using cross-validation and then trained and evaluated on the same splits as the LSTM model. Two further models were also considered: a BERT (base uncased) model and a GPT-2 model, both of which were fine-tuned from their pre-trained weights. The results of these experiments are given in Table 5.

https://doi.org/10.1017/nlp.2024.62 Published online by Cambridge University Press

Table 3. Hyperparameter tuning value ranges with optimal values determined by grid search

Hyperparameter	Value range	Optimal value
LSTM layers	1, 2, 3	1
Layer width	16, 32, 64, 128, 256	128
Dropout probability	0.1, 0.2, . . . , 0.8	0.8
Learning rate	$10^{-4}, 10^{-3}, 10^{-2}$	10^{-2}

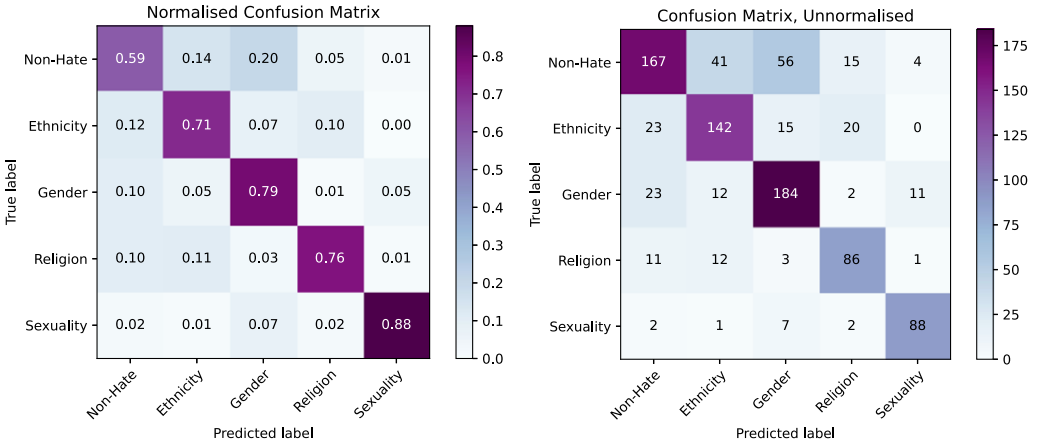

Figure 2. LSTM model—test set normalised & unnormalised confusion matrices.

5.3 Comparison with binary classification

To determine the efficacy of the multiclass approach, we trained the GPT-2 model to perform binary classification in two scenarios: first, directly on the aggregated dataset, with the hate categories mapped to a single category; and secondly, balancing this single category with the non-hate tweets in the dataset. The first experiment yielded an F1 score of 76.59% while the second gives an F1 score of 76.27%. Confusion matrices for these experiments are shown in Figure 3. The models exhibit a better overall performance on detecting hate, likely due to a reduced level of confusion between classes in the multiclass approach.

5.4 Effect of dataset size

To examine the role of dataset size in model performance, we fine-tuned the GPT-2 model on successively larger subsets of the dataset. The results of this are given in Table 6, with confusion matrices for each subset model shown in Figure 4. From this, it is evident that the size of the dataset plays a significant role in model performance. Smaller datasets lead to many tweets in hate categories being misclassified as non-hate. This is particularly evident for the 20% subset but can be seen across all subsets.

5.5 Discussion, contributions & limitations

While many studies have considered the task of classifying hate, little attention has been given to multiclass classification across the full range of different categories considered in this study.

https://doi.org/10.1017/nlp.2024.62 Published online by Cambridge University Press

Table 4. LSTM model—test set evaluation metrics

Metric	Score
Accuracy	0.7188
Precision	0.7297
Recall	0.7447
F1 score	0.7423

Table 5. Comparison of model test scores on the aggregated dataset

Model	F1 score
Logistic regression	0.7391
LSTM	0.7423
Fine-tuned BERT (base uncased)	0.7593
Fine-tuned GPT-2 (124M parameters)	0.7220

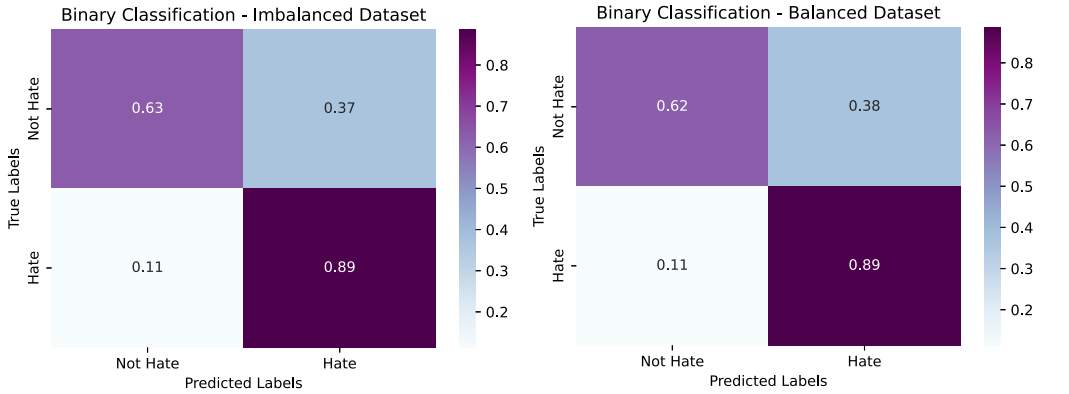

Figure 3. Confusion matrices for binary classification models.

Although the dataset presented here was compiled with varying definitions of hate speech, the models trained on it have achieved a classification performance that is comparable with similar studies, which annotated their tweets more consistently according to their own individual definitions. Our study also has the advantage of detecting a larger proportion of hateful tweets, across the four protected characteristics ethnicity, gender, religion, and sexuality.

It is evident from our results that larger models, such as fine-tuned BERT or GPT models, achieve similar or slightly higher F1 scores on this dataset, indicating that larger models may lead to improved model performance. However, there is significant room for improvement here which our results indicate cannot be achieved by considering model complexity alone. While other studies have considered a narrower range of categories, which may reduce model confusion, it should be noted that this may prevent detection of hate speech categories for which comparatively less data is available. Taken together with the results of the models trained on data subsets, this suggests that the availability of further data could lead to improved performance.

https://doi.org/10.1017/nlp.2024.62 Published online by Cambridge University Press

Table 6. Comparison of model test scores on smaller subsets of the dataset

Data subset	F1 score
20%	0.3419
40%	0.5795
60%	0.6520
80%	0.6888

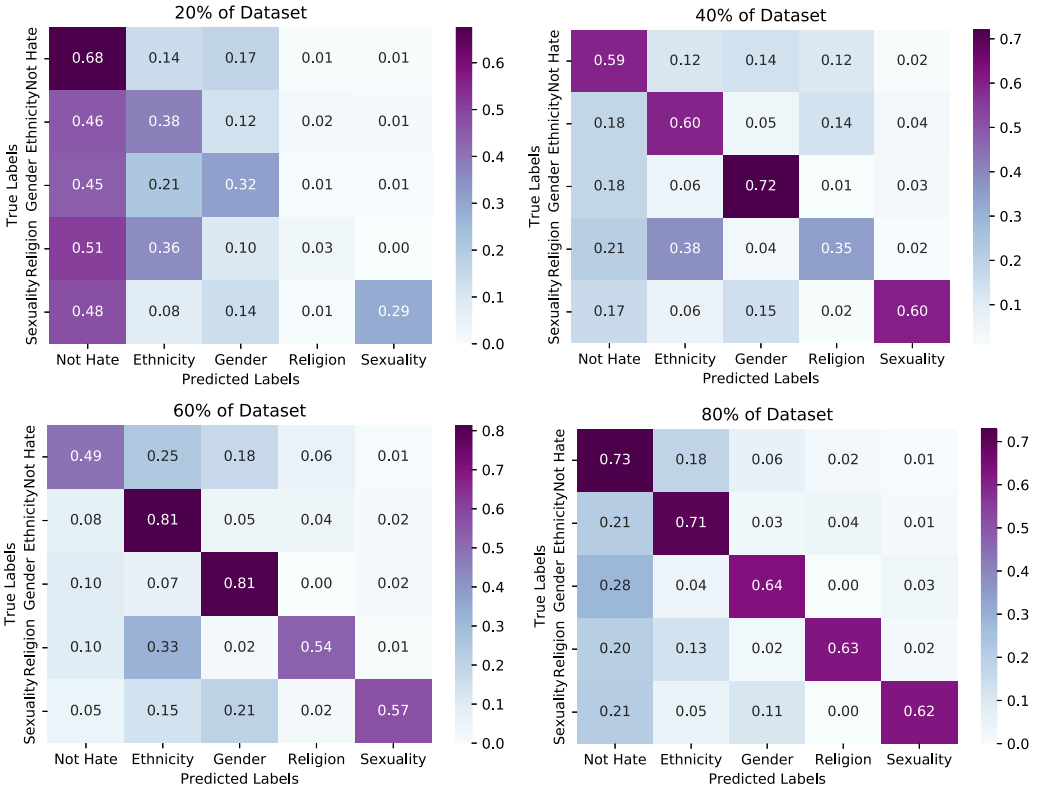

Figure 4. Confusion matrices for subset models.

One of the principal advantages of our approach is that it is capable of classifying hate towards four protected characteristics, with comparable performance, compared to either binary hate vs non-hate classification or the two specific characteristics classified in previous studies. While balancing the dataset makes it less similar to the class proportions of real data, it has been widely observed that models cannot learn well from data where the class imbalance is very pronounced, so it is essential to address the imbalance in some way as we have done here. This is a relatively simple method of addressing this problem and would merit further investigation.

The achievement of a good classification performance on four protected characteristics may also be seen as a limitation, as there are many more protected characteristics that could be included. As mentioned at the outset, the varying definitions of hate speech across countries,

https://doi.org/10.1017/nlp.2024.62 Published online by Cambridge University Press

organisations, and studies are a major challenge in hate speech research. An all-encompassing definition of hate speech could include the protected characteristics age, disability, disease, caste, gender, gender identity, sexual orientation, ethnicity, race, national origin, religion, and membership of an indigenous or ethnic community. While this study combined race, origin, and ethnicity into one category, there are still many more protected characteristics not included in this model, which are essential to protect all individuals and groups from online hate speech. However, due to the limited number of multi-category hate speech datasets available, which were developed with differing definitions of hate speech, this would require the collection and annotation of a new, much larger corpus of tweets, which is outside the scope of this study.

This lack of a universally accepted definition of hate speech has very likely impacted the classification performance of our model. Variations in definitions and the strictness of criteria followed by annotators when labelling the tweets can make a significant difference in model performance, with the difference between offensive and hateful content likely leading to particular confusion.

Lack of definitions used in determining hate speech and in quantifying levels of hate is also an issue for some of the datasets used in constructing our dataset, as is the inclusion of categories which do not strictly fall under the definition of hate speech. Although these were removed when compiling the dataset, the strictness of criteria followed during the annotation process could be questioned. For these reasons, it is likely that the final dataset contained inconsistently annotated tweets, in which some offensive tweets were labelled as *hate* instead of *non-hate*. It is possible therefore that the models are provided with conflicting examples of hateful tweets across the four compiled datasets, which may affect model performance. This is a potential source of misclassifications between the non-hate class and the gender and ethnicity classes. This limitation could be improved in all hate speech studies by the creation of a universal hate speech definition for use by the research community when creating datasets. Along with a universally accepted definition, a universally accepted procedure for annotating datasets would not only improve the quality of hate speech research but would also make research in this area more comparable.

6. Conclusion

Monitoring of social media platforms to remove hateful content in a timely manner remains a challenge for social media platforms due to the volume and velocity of data produced by users. This study addressed this problem by compiling a dataset and classifying tweets into the categories ethnicity, gender, religion, sexuality, and non-hate. Although the dataset was compiled with varying definitions of hate speech, the models considered have achieved a classification performance that is comparable to similar studies, which annotated their tweets more consistently according to their own individual definitions.

Future work could involve addressing the limitations discussed previously, which mainly centre around the definition of hate speech, dataset annotation issues, and class imbalance. Although the successful classification of the four protected characteristics, ethnicity, gender, religion, and sexuality, represents an improvement compared to previous studies, there are many more protected characteristics yet to be classified, such as age, disability, disease, caste, gender identity, and membership of an indigenous or ethnic community. A more comprehensive dataset of tweets containing hate and non-hate towards the four classified characteristics, and as many of the remaining unclassified characteristics as possible, would likely be a useful extension of this work.

As regards the dataset, future improvements would also aim to address the inconsistency in hate speech definition in the annotation of these tweets. A universally accepted definition of hate speech across the research community, and corresponding annotation criteria, would be very helpful here. As recommended by Ross *et al.* (2017), the labelling criteria should not be treated as a binary decision, and instead, should be made up of a set of questions to identify hate speech and its targeted protected characteristic in the form of a key, with a particular focus on distinguishing

https://doi.org/10.1017/nlp.2024.62 Published online by Cambridge University Press

between offensive tweets and hateful tweets to avoid providing the model with conflicting training data.

Competing interests. The author(s) declare none.

References

Aditya A., Vinod R., Kumar A., Bhowmik I. and Swaminathan J. (2022). Classifying Speech into Offensive and Hate categories along with Targeted Communities using Machine Learning. In 2022 International Conference on Inventive Computation Technologies (ICICT), Nepal, IEEE, pp. 291–295.

Alnazzawi N. (2022). Using Twitter to detect hate crimes and their motivations: The hatemotiv corpus. *Data* **7**(6), 69.

Basile V., Bosco C., Fersini E., Nozza D., Patti V., Pardo F. M. R., 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, 6–7 June 2019, Minneapolis, MN, pp. 54–63.

Bosu A. (2018). Stanford NER English 3-class caseless model. available: https://github.com/amiangshu/SentiSE [accessed 18 Aug 2023].

Burnap P. and Williams M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science* **5**, 1–15.

Chung Y.-L., Kuzmenko E., Tekiroglu S. S. and Guerini M. (2019). CONAN–COunter NArratives through Nichesourcing: a multilingual dataset of responses to fight online hate speech. arXiv preprint arXiv:1910.03270.

Criminal Justice (Incitement to Violence or Hatred and Hate Offences) Bill (2022). No 105/2022, Government of Ireland, available: https://www.oireachtas.ie/en/bills/bill/2022/105/?tab=bill-text [accessed 06 Jun 2023].

Curry A. C., Abercrombie G. and Rieser V. (2021). ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI, arXiv preprint arXiv:2109.09483.

ElSherief M., Kulkarni V., Nguyen D., Wang W. Y. and Belding E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. In Proceedings of the international AAAI conference on web and social media, ICWSM 2018, 25–28 June 2018, Pao Alto, CA, pp. 42–51.

Fazil M., Khan S., Albahlal B. M., Alotaibi R. M., Siddiqui T. and Shah M. A. (2023). Attentional multi-channel convolution with bidirectional LSTM cell toward hate speech prediction. *IEEE Access* **11**, 16801–16811.

Fersini E., Gasparini F., Rizzi G., Saibene A., Chulvi B., Rosso P., Lees A. and Sorensen J. (2022). Semeval-2022 task 5: Multimedia automatic misogyny identification. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, pp. 533–549.

Finkel J. R., Grenager T. and Manning C. D. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pp. 363–370.

Gambäck B. and Sikdar U. K. (2017). Using convolutional neural networks to classify hate-speech. In Proceedings of the First Workshop on Abusive Language Online, Vancouver, BC, Canada: Association for Computational Linguistics, pp. 85–90.

Hassan G., Rabah J., Madriaza P., Brouillette-Alarie S., Borokhovski E., Pickup D., Varela W., Girard M., Durocher-Corfa L. and Danis E. (2022). Protocol: Hate online and in traditional media: A systematic review of the evidence for associations or impacts on individuals, audiences, and communities. *Campbell Systematic Reviews* **18**(2), e1245.

Kennedy C. J., Bacon G., Sahn A. and von Vacano C. (2020a). Constructing interval variables via faceted Rasch measurement and multitask deep learning: A hate speech application. arXiv preprint arXiv:2009.10277.

Kennedy C. J., Bacon G., Sahn A. and von Vacano C. (2020b). Measuring hate speech [dataset]. available: https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech [accessed 19 Jul 2023].

Kirk H. R., Yin W., Vidgen B. and Röttger P. (2023). Semeval-2023 task 10: Explainable detection of online sexism. arXiv preprint arXiv:2303.04222.

Kovács G., Alonso P. and Saini R. (2021). Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources. *SN Computer Science* **2**, 1–15.

MacAvaney S., Yao H.-R., Yang E., Russell K., Goharian N. and Frieder O. (2019). Hate speech detection: Challenges and solutions. *PloS One* **14**(8), e0221152.

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

Mathew B., Saha P., Yimam S. M., Biemann C., Goyal P. and Mukherjee A. (2021b). Hatexplain: A benchmark dataset for explainable hate speech detection [dataset]. available: https://huggingface.co/datasets/hatexplain [accessed 19 Jul 2023].

Ofcom and Traverse (2023). Qualitative research into the impact of online hate. Technical report, Ofcom, London, United Kingdom. available: https://www.ofcom.org.uk/__data/assets/pdf_file/0020/252740/qual-research-impactof-online-hate.pdf [accessed: 15 Jun 23].

https://doi.org/10.1017/nlp.2024.62 Published online by Cambridge University Press

Ousidhoum N., Lin Z., Zhang H., Song Y. and Yeung D.-Y. (2019). Multilingual and multi-aspect hate speech analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, pp. 4675–4684.

Pluta A., Mazurek J., Wojciechowski J., Wolak T., Soral W. and Bilewicz M. (2023). Exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others' pain. *Scientific Reports* **13**(1), 4127.

Qureshi K. A. and Sabih M. (2021). Un-compromised credibility: Social media based multi-class hate speech classification for text. *IEEE Access* **9**, 109465–109477.

Rosenthal S., Atanasova P., Karadzhov G., Zampieri M. and Nakov P. (2020). SOLID: A large-scale semi-supervised dataset for offensive language identification. arXiv preprint arXiv:2004.14454.

Ross B., Rist M., Carbonell G., Cabrera B., Kurowsky N. and Wojatzki M. (2017). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. arXiv preprint arXiv:1701.08118.

Sachdeva P., Barreto R., Von Vacano C. and Kennedy C. (2022). Targeted Identity Group Prediction in Hate Speech Corpora. In Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pp. 231–244.

Saha K., Chandrasekharan E. and De Choudhury M. (2019). Prevalence and psychological effects of hateful speech in online college communities. In Proceedings of the 10th ACM Conference on Web Science, pp. 255–264.

Seemann N., Lee Y. S., Höllig J. and Geierhos M. (2023). The problem of varying annotations to identify abusive language in social media content. *Natural Language Engineering* **29**(6), 1561–1585.

Soral W., Bilewicz M. and Winiewski M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior* **44**(2), 136–146.

Taulé M., Ariza A., Nofre M., Amigó E. and Rosso P. (2021). Overview of detoxis at iberlef 2021: Detection of toxicity in comments in spanish. *Procesamiento del lenguaje natural* **67**, 209–221.

Toraman C., Şahinuç F. and Yilmaz E. H. (2022a). Large-scale hate speech dataset [dataset]. v2. available: https://github.com/avaapm/hatespeech [accessed 20 Jul 2023].

Toraman C., Şahinuç F. and Yilmaz E. H. (2022b). Large-scale hate speech detection with cross-domain transfer. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, pp. 2215–2225.

Twitter Inc (2023). Twitter's policy on hateful conduct — Twitter help. available: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy [accessed 06 Jun 2023].

Vidgen B. and Yasseri T. (2020). Detecting weak and strong Islamophobic hate speech on social media. *Journal of Information Technology & Politics* **17**(1), 66–78.

Waseem Z. (2016a). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate speech detection on Twitter. In Proceedings of the First Workshop on NLP and Computational Social Science, Austin, Texas: Association for Computational Linguistics, pp. 138–142.

Waseem Z. (2016b). Are you a racist or am I seeing things? Annotator Influence on Hate speech detection on Twitter [dataset]. available: https://github.com/zeeraktalat/hatespeech [accessed 19 Jul 2023].

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, pp. 88–93.

Wermiel S. J. (2018). The ongoing challenge to define free speech. *Human Rights Magazine* **43**(4), 82. available: https://www.americanbar.org/groups/crsj/publications/human_rights_magazine_home/the-ongoing-challenge-todefine-free-speech/the-ongoing-challenge-to-define-free-speech/ [accessed 08 Jun 2023].

Williams M. (2019). Hatred behind the screens: A report on the rise of online hate speech. Technical report, HateLab, Cardiff University and Mishcon de Reya, Cardiff, United Kingdom. available: https://orca.cardiff.ac.uk/id/eprint/127085/ [accessed: 09 Jun 2023].

Zampieri M., Malmasi S., Nakov P., Rosenthal S., Farra N. and Kumar R. (2019a). Predicting the type and target of offensive posts in social media. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Vol. **1**, pp. 1415–1420.

Zampieri M., Malmasi S., Nakov P., Rosenthal S., Farra N. and Kumar R. (2019b). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). available: https://github.com/zeeraktalat/hatespeech [accessed 19 Jul 2023].

Zampieri M., Rosenthal S., Nakov P., Dmonte A. and Ranasinghe T. (2023b). Offenseval 2023: offensive language identification in the age of large language models. *Natural Language Engineering* **29**(6), 1416–1435.

Cite this article: Walsh S and Greaney P. Multiclass hate speech detection with an aggregated dataset. *Natural Language Processing* https://doi.org/10.1017/nlp.2024.62