# Empirical Evaluation of Public HateSpeech Datasets

Sardar Jaf, and Basel Barakat

***Abstract*—Despite the extensive communication benefits offered by social media platforms, numerous challenges must be addressed to ensure user safety. One of the most significant risks faced by users on these platforms is targeted hatespeech. Social media platforms are widely utilized for generating datasets employed in training and evaluating machine learning algorithms for hatespeech detection. However, existing public datasets exhibit numerous limitations, hindering the effective training of these algorithms and leading to inaccurate hatespeech classification. This study provides a systematic empirical evaluation of several public datasets commonly used in automated hatespeech classification. Through rigorous analysis, we present compelling evidence highlighting the limitations of current hatespeech datasets. Additionally, we conduct a range of statistical analyses to elucidate the strengths and weaknesses inherent in these datasets. This work aims to advance the development of more accurate and reliable machine learning models for hatespeech detection by addressing the dataset limitations identified.**

***Impact Statement*—Hatespeech is a form of abusive language targeted at people based on their personal traits, beliefs, views, etc. The development of tools and resources to efficiently and accurately detect and prevent hatespeech on online platforms is important for the provision of safe online platforms for users. Labeled datasets containing information on hatespeech is one of the most important resources for developing tools for automatic hatespeech processing. However, current public datasets have major limitations. There have been lack of attempts in assessing their suitability for developing Artificial Intelligence (AI) systems to accurately identify hatespeech content. Therefore, this timely project offers extensive empirical evaluation of many public hatespeech datasets. We identify their weaknesses and suitability for designing automated hatespeech classification systems. We offer in depth analyses of several datasets to help future research studies on dataset development avoid introducing weaknesses to their dataset and maximize their suitability for AI systems on hatespeech processing.**

***Index Terms*—Dataset Evaluation, Hatespeech, Hate Classification, Hatespeech Dataset, Hatespeech Dataset Evaluation, Hatespeech Corpus Evaluation**

## I. Introduction

Social media is one of the most widely used online mediums for data sharing and communication by people in modern society, where people can easily share information, news, updates and their opinions on current trends. One of the potential risks associated with easily sharing and publishing information by online users that is accessible worldwide is the integrity of the information. Particularly, the dissemination of hatespeech. Thus, one of the pressing needs for many online platforms and users is to ensure published information on online platforms are free from hatespeech. To address this need, numerous efforts have been made for the provision of datasets to be used in the design of AI systems to efficiently and accurately detect, classify and remove hatespeech content. However, despite the availability of many datasets, which are crucial components for the development of AI powered hatespeech detection/classification systems, the quality of those datasets is questionable. Generally, poor quality datasets would lead to poor AI systems. Therefore, in this paper, we focus on evaluating public hatespeech datasets that are based on social media platforms. Our aim is to empirically evaluate the quality of many public hatespeech datasets in order to assess their suitability for AI hatespeech classifiers.

The evaluation of hatespeech datasets is crucial for the development of more accurate, ethical, and inclusive machine learning models that address the challenges of online hatespeech effectively. Dataset quality is paramount for training effective hatespeech classification models.

Our study on evaluating multiple public hatespeech datasets provides empirically informed assessment of the reliability and suitability of these datasets for building robust hatespeech classifiers. Since available hatespeech datasets are based on data collected from various online platforms such as Reddit, Facebook, X, and YouTube, their features vary significantly from one platform to another. For example, hatespeech on X may differ from that found in YouTube comments or Reddit discussions. Therefore, our evaluation of datasets generated from different platforms helps us identify certain limitations and weaknesses in these datasets that may impact classification models.

Moreover, our study assesses the quality of different datasets, helping to identify high-quality datasets that can be targeted for future classification model development. In addition, our research provides insights into different datasets that can guide future projects towards creating standardized hatespeech datasets and assist policymakers in developing regulations related to online hatespeech and content moderation.

Finally, social dynamics play a critical role in how hatespeech evolves over time. By evaluating many public hatespeech datasets, we can identify patterns of hatespeech that have changed or emerged over time, contributing to a deeper understanding of the phenomenon and informing the development of adaptive, future-proof models.

The contributions of this study are as follows:

- We offer the systematic and empirical evaluation attempt of a large number of datasets.
- We empirically demonstrate that the quality of dataset content has a greater positive impact on AI hatespeech classification than factors such as content volume, context diversity, and data modalities.

This manuscript was submitted for review on 20/10/2023 to IEEE Transactions on Artificial Intelligence.

Sardar Jaf is with School of Computer Science, University of Sunderland, Sunderland, SR1 3SD, United Kingdom (sardar.jaf@sunderland.ac.uk).

Basel Barakat is with School of Computer Science, University of Sunderland, Sunderland, SR1 3SD, United Kingdom (basel.barakat@sunderland.ac.uk).

This paragraph will include the Associate Editor who handled your paper.

- We offer novel approach utilizes hatespeech dataset features to identify correlation between each feature and machine learning classification performance. This approach has the potential to be generalized to datasets in other domain.
- We present a simple yet highly effective baseline deep neural network architecture for hatespeech classification, that outperforms some published binary hatespeech classifiers.

The rest of the paper is organized as follows: in section II, we review published literature on hatespeech datasets and machine learning approaches to hatespeech classification. In section III, we describe our methodology, identifying and selecting public datasets, preprocessing them, normalizing them, binarizing their labels, performing statistical analyses, and developing a baseline hatespeech classifier. We present experimental setup, results, and detailed analysis of the evaluation results of all the selected datasets in section IV. Finally, we conclude our finding and identify potential future work in section VI.

## II. Literature Review

One of the primary sources for collecting big data for text analyses is social media platforms. These platforms have been used by researchers from different disciplines as a data collection source [1]. For academic research projects, social media data has been explored widely for research and practical applications of hatespeech detection, analyses and classification. As a result, many datasets have been compiled from various social media platforms for hatespeech processing. In this section, we will highlight core aspects of some of the published datasets and machine learning applications for the task of hatespeech classification.

**Hatespeech datasets** are largely produced by extracting content (e.g., text, images, memes, videos, emojis etc.) from social media platforms, online forums, blogs and various other online communities.

To develop machine learning approaches to process hatespeech (which involves data analyses, classification, visualization etc.), access to labeled datasets is essential. Since there is no commonly accepted benchmark dataset for processing hatespeech, authors usually collect from online platforms and annotate them using different annotation approaches. This practice resulted in considerable variation in the size of the published datasets, topics, domains, languages, hatespeech categories, platforms, content types, etc. Some datasets are very large (containing over hundred thousand entries [2] [3], [4]) whereas others are small (contain a few thousands entries [5] [6] or few hundreds entries [7]. The main reasons for such data size variation are: (i) as in any text annotation, annotating hatespeech is an extremely time-consuming process, (ii) there are, usually, much fewer hateful than non-hateful (neutral) comments present in sampled data from social media platforms. Therefore, accomplishing this task necessitates the collection of extensive data that can be annotated to identify a substantial number of hatespeech instances. The negative impact of this imbalanced distribution of content types is that it would generally be difficult to build a balanced dataset, where there are equal samples of hateful and neutral content.

TABLE I: Dataset names, platforms, year, and publication sources.

| Datasets | Platform | Year | Ref. |
|---|---|---|---|
| Davidson et al. | Twitter | 2017 | [5] |
| Gibert et al. | Stormfront | 2018 | [14] |
| Gomez et al. | Twitter | 2019 | [3] |
| Kennedy et al. | Twitter, Reddit, YouTube | 2020 | [2] |
| Qian et al. | Gab | 2019 | [13] |
| Salminen et al. | YouTube and Facebook | 2018 | [6] |
| Suryawanshi et al. | Reddit, Facebook, Twitter and Instagram | 2020 | [7] |
| Vidgen et al. A | Dynamically generated | 2021 | [15] |
| Vidgen et al. B | Reddit | 2021 | [12] |
| Waseem and Havoy | Twitter | 2016 | [8] |

Some authors attempted to increase the sample size of hatespeech content whereas keeping the size of data instances to be annotated at a reasonable level, [8][1] proposed an approach to pre-select the text instances to be annotated by querying an online platform (Twitter) for topics that are likely to contain a higher degree of hatespeech (e.g. "Islam terror"). The strength of this approach is it increases the proportion of hatespeech samples in the resulting dataset, and thus resulting in the possibility of achieving a balanced dataset. However, the limitation of this approach is that it focuses the resulting dataset on specific topics and certain subtypes of hatespeech (e.g. hatespeech targeting Muslims)[9].

Since there is no commonly accepted benchmark corpus for hatespeech classification, authors usually collect and label their own data [9]. For this reason, most of the available datasets are based on content from one or few data sources. Some of the major sources of datasets are: Yahoo[10] [11] [4], X, formally known as Twitter,[5] [8], Reddit [12], Qian et al.[13], YouTube[2][6], Facebook[6], Stormfront[14] or dynamically generated text[15]. The result of collecting data from different online platforms for creating hatespeech dataset is that the dataset are likely to have different characteristics, and subtypes of hatespeech[9], which is largely because of the nature and purpose of the online platforms. Thus, they may have special characteristics. For instance, a platform especially created for adolescents, one should expect quite different types of hatespeech compared to a platform that is used by a cross-section of the general public because the resulting different demographics will have an impact on the topics discussed and the language used [9].

The above issues related to hatespeech datasets have lead to the creation and availability of several datasets for the task of automated hatespeech classification. Table I contains a list of publicly available hatespeech datasets, which we have evaluated in this study.

**Hatespeech classification** methods for processing hatespeech content, especially classifying social media content as "Hateful" or "Neutral", are largely based on supervised classification method. This method, involves using labeled/annotated data for training machine learning algorithms to classify hatespeech content. Two types of machine learning algorithms are usually used in supervised learning: Shallow learning

[1]The dataset is available at http://github.com/zeerakw/hatespeech

algorithms (such as support vector machine, decision trees, nearest neighbors, etc.) have been widely utilized. (ii) Deep learning algorithms, which mainly cover various types of recurrent neural networks

Other classification methods to hatespeech content employ semi-supervised method, particularly bootstrapping, which can be utilized for different purposes in the context of hatespeech processing. On the one hand, it can be used to obtain additional training data, as it is done in [16]. On the other hand, it can be utilized to build lexical resources that are used as part of the detection process. The authors of [17] apply this method to populate their hate verb lexicon, starting with a small seed verb list, and iteratively expanding it based on WordNet relations, adding all synonyms and hypernyms of those seed verbs.

In recent years, many language models have been extensively explored for text classifications tasks in recent years. Chief among them is BERT, which has contributed to many NLP applications, especially for hatespeech detection, as in the work of [18], [19], [20], [21], [22], [23], [24] where they have utilized BERT as crucial components for building effective hatespeech detection systems. One of the strong feature of BERT is that it is multilingual language model. This aspect of the model prompted many authors to apply it to detecting hatespeech in different languages. [25] fine tuned BERT model for detecting hatespeech in Urdu text, [26] developed their hatespeech system based on BERT for Bengali hatespeech detection, [27], applied BERT for Arabic hatespeech classification task, and [28] used BERT for detecting hatespeech in Spanish text.

## III. Methodology

### A. Data collection

Hatespeech is gaining increasing attention from industry, government organizations and academia. The proliferation of information published on social media platforms provides the means to create datasets for processing, analyzing, detecting, and classifying hatespeech content. Variations in the available datasets (e.g, size, topic, domain, language, hatespeech categories, platform, content type, etc.) could be beneficial. However, it can make it challenging for researchers, and machine learning engineers, to determine which dataset is suitable for training machine learning algorithms for hatespeech classification. For example, a dataset based only on Twitter content may, or may not, be suitable for producing a generalizable machine learning model that perform well on non-twitter data, such as YouTube comments.

The aim of this study is to evaluate multiple publicly available datasets to assess their suitability for training and testing deep learning algorithms for hatespeech classification. We have selected ten datasets from hatespeechdata.com website, which is a widely used platform for hosting hatespeech datasets. Our goal in evaluating multiple datasets is to examine two application aspects of each dataset: (i) to examine the suitability of a dataset in testing the performance of a deep learning based system for hatespeech classification, and (ii) to examine the suitability of a dataset in producing generalizable deep learning model by training a deep learning algorithm on a dataset but testing it on other dataset with different domain, text, genre, and size. The selected dataset names, platforms where the data are collected from, and the publication years of the dataset are presented in Table I. We chose those datasets based on several characteristics: different platforms, dataset size, content type, length of individual text entry, and publication time. Examining dataset with content extracted from various platforms helps us to evaluate the generalizability of deep learning algorithms. Similarly, the different years were chosen to ensure that the evaluation would be generalizable in terms of the evolution of hatespeech patterns over time as some hateful terms might be more popular in specific years. We selected datasets with different sizes because machine learning algorithm performance is usually dependent on dataset size, and thus we can examine the impact of dataset size on machine learning model performance.

The labeled content in each selected dataset varies from one dataset to another. The proposed dataset by [12], has various categories of abuse (e.g., targeted Identity, affiliation, and person), and counter speech. The published dataset by [5] has three classes of content i.e., *Hatespeech*, *Offensive language*, and *Neither*. The datasets from [13] and [15], have only two classes: *Hate/Not Hate*, and *Offensive/Vulgar*. Kennedy et al.'s dataset [2] content is focused on particular categories such as religions, three hate classes for races, and a hate score to indicate the hate level. The dataset from [3] contains text and images. They are labeled as "Hate" or "Not-Hate"."Hate" content is further divided into five classes of different types of hate. Waseem and Havoy's dataset [8], has three classes: "Sexism", "Racism", and "None".

The content of the available datasets has been labeled with different types of hatespeech. Therefore, there are inconsistent labels between the datasets. In order to design and evaluate a supervised deep learning hatespeech classifier trained on the available datasets for binary hatespeech classification, we have converted the various labels in the datasets into either "Hate" or "Not-Hate" labels. Therefore, each dataset has one of two classes ("Hate" or "Not-Hate"). This approach enables us to make the labels in all the selected datasets uniform.

### B. Label binarization

The available datasets have different classes of hatespeech content. since each dataset contains different classes of hatespeech, there is inconsistent hatespeech classes between the datasets. One of the main reasons that published hatespeech datasets have different classes for hatespeech content is because there is no uniform consensus in the research community on the different classes of hatespeech, which is a challenge that requires further effort from the research community to address. Since there is no consensus in the research community on the different types of hatespeech, authors of the published datasets have labeled their data with different classes of hate. Some datasets contain fine-grained categories of hate, where the victims are targeted based on their race; religion; sexuality; ethnicity; gender; etc. [2], [12]. Other datasets contain broad categories such as "hateful", "abusive" or "neutral" [3], [15], [8]. Table II presents a summary of some of the categories

TABLE II: Datasets and examples of content categories

| **Datasets** | Examples content categories |
|---|---|
| Davidson et al. [5] | hatespeech, offensive language, neither |
| Gibert et al. [14] | Hate/not hate, relation, idk/skip |
| Gomez et al. [3] | Hate/not hate |
| Qian et al. [13] | Hate, Offensive/Vulgarity |
| Kennedy et al. [2] | respect, insult, humiliate, status, dehumanize, violence,... (e.g. black, etc), target_religion_... (e.g., atheist, etc), target_origin_... (e.g. immigrant, etc) target_gender_... (e.g. men, etc), target_sexuality_... (e.g., bisexual, etc), target_age_... (e.g., children, etc), target_disability_... (e.g., physical, etc.) |
| Salminen et al. [6] | Hate/Neutral |
| Suryawanshi et al. [7] | Offensive/Non-offensive |
| Vidgen et al. A [15] | Hate/not hate |
| Vidgen et al. B [12] | AffiliationDirectedAbuse, PersonDirectedAbuse, Identity-DirectedAbuse, CounterSpeech |
| Waseem and Havoy [8] | Sexism/racism |

TABLE III: Datasets statistics. Mean/Median, Min/Max, and Variant/Standard Deviation are based on word counts. 'NH' is Not-hate

| Datasets Name | Split (H/NH)<br>Hate % | Mean/Median<br>(H) NH | Min/Max<br>(Hate) NH | VAR/STD<br>(Hate) NH |
|---|---|---|---|---|
| Davidson et al. [5] | (1430 / 1430)<br>50% | (13.9/ 13.0)<br>14.8/15 | (1/ 32)<br>2/32 | (49.2/7.0)<br>45.5/6.7 |
| Gibert et al. [14] | (1437/ 9507)<br>13.1% | (22.0/20.0)<br>17.3/15.0 | (1/349)<br>1/262 | (234.8/15.3)<br>179.5/13.4 |
| Gomez et al. [3] | (112787 / 25263)<br>81.7% | (11.7/11.0)<br>11.5/11 | (2/90)<br>2/81 | (28.4/5.3)<br>28.4/5.3 |
| Kennedy et al. [2] | (46021 / 80624)<br>36.3% | (25.8/19.0)<br>28.1/21 | (1/128)<br>1/128 | (412.4/20.3)<br>507.0/22.5 |
| Qian et al. [13] | (2348 / 25198)<br>8.5% | (27.8/23.0)<br>20.3/15 | (1/191)<br>1/282 | (364.5/ 19.1)<br>274.1/16.6 |
| Salminen et al [6] | (2364 / 858)<br>73.4% | (43.6/34.0)<br>38.2/30.0 | (1/386)<br>1/351 | (1678.1/41.0)<br>1408.5/37.5 |
| Suryawanshi et al. [7] | (303 / 440)<br>40.8% | (45.0/33.0)<br>44.8/32.0 | (4/307)<br>2/268 | (1630.4/40.4)<br>1743.2/41.8 |
| Vidgen et al. A [15] | (22175 / 18969)<br>53.9% | (23.8/15.0)<br>25.1/17 | (1/395)<br>1/408 | (599.6/24.5)<br>621.0/24.9 |
| Vidgen et al. B [12] | (4093 / 19107)<br>17.6% | (39.5/ 19.0)<br>28.7/14.0 | (1/1937)<br>1/1417 | (7690.7/87.7)<br>2908.0/53.9 |
| Waseem and Havoy [8] | (2692 / 7766)<br>25.7% | (16.8/17)<br>14/14 | (1/33)<br>1/38 | (41.3 /6.4)<br>49.3/7.0 |

of hatespeech in each dataset. Some datasets contain very few and general hatespeech content (such as "hate" or "offensive", as in the dataset published by [13] and [5]) whereas other datasets have many fine-grained hatespeech types such as the dataset published by [2], which includes various subtypes of hatespeech based on gender or religion.

Prior to exploring and analyzing the content of the selected datasets for this study, we have binarized the labels (classes). The objective is to ensure all the datasets contain consistent labels ("Hate" or "Not-Hate"). The label binarization process is performed as follows:

- Merge fine-grained labels of hatespeech to broad labels. If the label of content indicates any type of hatespeech, then we convert it to a broad label "Hate". If the label indicates the content is "neutral" or "not hateful", then we convert it to "Not-hate".
- Drop ambiguous labels by discarding any content in a dataset where the content has an ambiguous label, such as "abusive", because such content may not be considered hateful.
- Convert content that is labeled as "neutral", "not-hate", or "not-abusive" to the"Not-hate" label.

Table III presents some statistical information on each dataset after we have binarized the labels as "Hate" or "Not-Hate". The table contains the followings: sample size for each dataset based on the content split between "Hate" and "Not-Hate", total unique word count for "Hate" and "Not-Hate" content, the mean and median, the min/max, and the variance and standard deviation.

TABLE IV: Dataset size: before and after balancing sample size

| | **Initial Binarized Dataset Size** | | | **Balanced Binarized Dataset Size** | | |
|---|---|---|---|---|---|---|
| **Dataset** | Hate | Not-Hate | Total | Hate | Not-Hate | Total |
| Davidsonet al [5] | 1430 | 1430 | 2860 | 1430 | 1430 | 2860 |
| Gibert et al. [14] | 1437 | 9507 | 10944 | 1437 | 1437 | 2874 |
| Gomez et al. [3] | 25263 | 112787 | 138050 | 25263 | 25263 | 50526 |
| Kennedy et al. [2] | 46021 | 80624 | 126645 | 46021 | 46021 | 92042 |
| Qian et al. [13] | 2348 | 25198 | 27546 | 2348 | 2348 | 4696 |
| Salminen et al [6] | 2364 | 858 | 3222 | 858 | 858 | 1716 |
| Suryawanshi et al. [7] | 303 | 440 | 743 | 303 | 303 | 606 |
| Vidgen et al. A [15] | 22175 | 18969 | 41144 | 18969 | 18969 | 37938 |
| Vidgen et al. B [12] | 4093 | 19107 | 23200 | 4093 | 4093 | 8186 |
| Waseem and Havoy [8] | 2692 | 7766 | 10458 | 2692 | 2692 | 5384 |

The label binarization process provides us with a dataset containing consistent labels of "Hate" or "Not-Hate", which we can use to evaluate their suitability for training and testing a deep learning algorithm for binary hatespeech classification. We evaluate a baseline deep learning system on each dataset to assess its performance in two tasks: (i) performing binary classification of hatespeech (i.e., classifying text as either "Hate" or "Not-hate"), and (ii) performing transfer learning classification to test the generalization of the system where we train the system on one dataset and test it on multiple other dataset.

One of the major issues with all the public datasets that negatively impact the performance of machine learning algorithms is the feature imbalance. The sample size for different categories of text is often uneven, with some categories having more content than others. Most of the datasets appear to have more text related to "Not-Hate" than "Hate". To address the data imbalance problem, we balance the sample size of "Hate" and "Not-Hate" content for each dataset before we use them to design and evaluate a baseline deep learning system.

### C. Dataset balancing

The available datasets are imbalanced, as they contain unequal sample size for different text categories. As we discussed in Section III-B, we have binarized the labeled content in each dataset so that they contain only "Hate" and "Not-Hate".

As it can be seen in Table IV, the differences in the sample size for different labeled content is large in all the datasets. Such differences in sample sizes negatively affect the training of machine learning algorithms (including deep learning algorithms), as the algorithms become bias towards the majority sample. To balance the sample size of "Hate" and "Not-Hate" content in each dataset, we apply under-sampling methods, reducing the size of the majority sample to match the size of the minority one. The columns in Table IV show the balanced sample size of "Hate"/"Not-Hate" content for each dataset.

### D. Statistical analysis

To gain a better understanding of the nature and prevalence of hatespeech, this study utilized a quantitative approach to

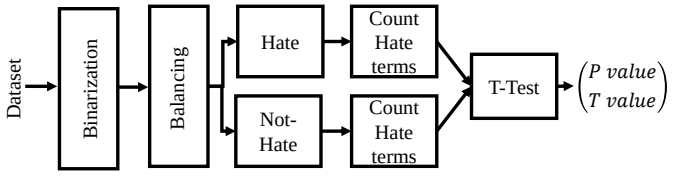



Fig. 1: Block diagram of the T-test procedure used for comparing the usage of hate terms in hateful and non-hateful context in each dataset

analyze the frequency of hate terms in both hateful and non-hateful speech. Our analysis began with collecting, 1523 frequently used hate terms from Hatebase[2], a publicly available database of hatespeech terms. We then conducted an analysis of the usage frequency of these terms in both types of text. To ensure the comparability of the two counts, we utilized the balanced datasets.

We conducted a T-test for the means of the counted hate terms frequency using python scientific computing package [29]. The T-test generated two matrices for the comparison, the $t$-value and the $p$-value, providing a quantitative measure of the significance of differences in the usage of hate terms between hateful and non-hateful text. This allowed us to compare the usage of the hate terms in both types of text within each dataset, and to draw conclusions about the prevalence and nature of hatespeech in the datasets. Fig 1 presents a block diagram of the used T-test procedure. After the a dataset is binarized, we balance the sample size between "Hate" and "Not-Hate" content. Then we count the "Hate" terms that appear in both contexts, "Hate" and "Not-Hate". Finally, we calculate T-Test to obtain T and P values.

### *E. Hatespeech classification*

**Text preprocessing: cleaning and normalization.** Since the content of the published datasets is collected from different online platforms (e.g., Twitter, Facebook, YouTube comments, etc.), they have different features such as structure, topic, user writing style, etc. We stored each dataset in comma separated value files (CSV) with two columns (text and label). The text column contains the text content and the label column contains one of two values, "hate" or "not-hate". We have performed the following transformations on the text before training and evaluating our model:

- Lower casing. We convert all the text to lower case English characters.
- Removing non-English text. We remove content that is not part of the English alphabet.
- Normalizing emojis by replacing them with token "<EMOJ>".
- Normalizing tag. We transform all hashtags to the token "<HASHTAG>" and all usernames to the token "@USER".
- Removing duplication. We remove sequentially duplicated items such as words, spaces, characters (except when they are part of the word, e.g., "different" we keep it unchanged to keep the word spelling intact), etc.
- Removing punctuation. We remove all the English and non-English punctuation.
- Removing stop words. We remove all stop words such as 'a', 'the', 'of' etc.
- Normalizing URL. We transform all hyperlinks and website addresses to the token "<URL>".
- Normalizing HTML elements. We convert all named and numeric character references from HTMLs such as ">" and "&#amps;" in the text to their corresponding Unicode characters "<" and "&", respectively.
- Removing new line in text. We remove all new line in each text to create a single line text.

**Deep learning model implementation using BERT.** We have implemented a baseline deep learning text classification system. We have trained and tested the system on ten publicly available datasets.

Our model is based on Bidirectional Encoder Representations from Transformers model (BERT) [30], a widely utilized deep learning algorithm for text classification.

The primary reasons for using BERT in our study offers many benefits. One of the main advantages is its ability to provide contextual understanding around words. This feature is crucial for our model because the context surrounding certain words helps determine whether they are hateful or benign. For example, the word "shoes" on its own is neutral but using it in a sentence to describe or compare a person to them would be offensive, and usually used in hateful manner, in certain cultures and societies. Also, BERT's ability to handle such subtleties and nuances enhances machine learning models by distinguishing hatespeech from sarcasm. Additionally, BERT provides dense and meaningful vector representations for words and sentences, which significantly improves the performance of machine learning classifiers in identifying hatespeech patterns in data.

Moreover, BERT is a flexible language model that can be easily fine-tuned on different datasets. This feature was particularly useful in our study, as we fine-tuned the model on ten different datasets. BERT has been trained on 110 million parameters. The main advantage of using a pre-trained model (such as BERT) is the significant reduction in training time. Furthermore, since natural language is inherently ambiguous– and many words in English and other languages carry varying levels of ambiguity–BERT's integration of attention mechanism methods enable the model to focus on the most relevant parts of a sentence. This capability makes it robust against the ambiguous expressions often found in hatespeech content.

By leveraging ten datasets with diverse features, sizes, platform-dependent data, hatespeech types, and more, BERT's generalization capabilities allows us to adapt and fine-tune the model effectively across datasets based on different social media platforms (e.g., Facebook, YouTube comments, X, Reddit, etc.).

Finally, BERT model is a multilingual language model trained on a very large text data. BERT model is a multi-layer bidirectional Transformer, which is a deep learning model used in several Natural Language Processing tasks [30]. It has

[2]Available at: https://hatebase.org/

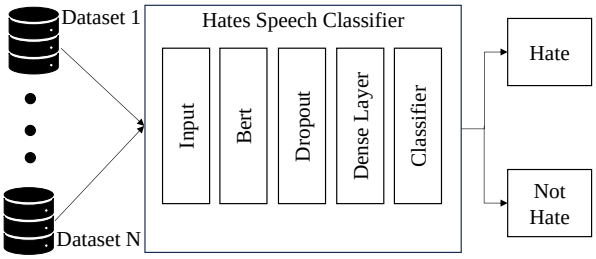



Fig. 2: Baseline architecture

consistently demonstrated strong performance across various NLP tasks, particularly in text classification. This makes it an excellent choice for our classification task.

**Data transformation.** After preprocessing and normalizing the text in each dataset III-E) we have tokenised the input data (textual information) using BERT, which is an essential step for training any deep learning algorithms. Next, the vector representation of the data is processed by a dropout layer with a dropout rate of 0.3. This step involves randomly excluding 30% of the training data during the training phase of the model in order to prevent the algorithm from memorizing the data pattern from the dataset, as is usually referred to as overfitting.

The output from the dropout layer is used as input to a single deep learning dense layer of neural network. We optimized the model learning capacity using Adam optimizer with learning rate of $2e-5$. During the training phase, we computed the training error rate using Binary Cross Entropy Loss function. Fig 2 shows the components of our model architecture.

In order to determine sentence boundary and length, we use BERT to assign special tokens to sentences, the "CLS", "SEP" and "PAD" token. The "CLS" token is to indicate for the model to identify the start of a sentence. The "SEP" token to indicate the end of a sentence for the model. Therefore, using "CLS" and "SEP" informs the model of the start and end of a sentence. We use the "PAD" token is used to assign extra empty token(s) to sentences in order for all input sentences to have equal token length. For example, if we restrict the model to learn from sentences of 50 tokens long, and a sentence contains 30 tokens, we would append it 20 empty tokens ("PAD") to make it a 50 tokens long sentence. We set the maximum length of sentences to 128 tokens. We add padding token "PAD" to sentences with less than 128 tokens in order to reach 128 tokens. Any sentence over 128 tokens long will be truncated. Moreover, BERT employs attention mask where tokens that represent words are masked with the value of 1 and tokens that represent nothing (e.g., "PAD") are masked with the value of 0. This masking mechanism allows BERT to determine what token(s) to retain or discard during the learning phase of the model. Each sentence, after being separated into individual words, will be encoded into vectors. We use the bert-base-cased pre-trained version for our model.

**Deep learning model Evaluation.** We follow a standard approach for evaluating a machine learning algorithm for a supervised binary classification task. The evaluation metrics used in this study are based on three standard measures for classification task: Recall, Precision and Weighted F1-score.

## IV. Dataset evaluation and results

### A. Dataset evaluation

We empirically evaluated ten different datasets using a baseline classifier. For each dataset evaluation, we divided the dataset into two parts: a training set and a test set, ensuring that the content of the two sets was distinct. We allocated 80% of each dataset to the training set for training the baseline classifier and 20% to the test set for evaluating the classifier's performance. within the training set, we further partitioned the data by allocating 90% for training the deep learning architecture and reserving 10% for validating the training accuracy during the training stage.

We conducted two types of evaluations on each dataset:

- **Mono-dataset evaluation**. In the model evaluation process, we trained and tested our baseline classifier, individually on each dataset, to classify their content as either "Hate" or "Not-Hate". The aim of this experiment is to assess the suitability of each dataset for binary classification of hatespeech content.
- **Generalized learning evaluation**. In this evaluation, we used the same baseline classifier that we used for the mono-dataset evaluation, but we trained and tested it differently. In this evaluation, we have conducted multiple experiments on the baseline classifier. In each experiment, we trained the model on one dataset and tested it on nine other datasets. For this experiment, our approach is: given dataset $d$ is a member of a set of dataset $D^{i,\dots,n}$, we trained our model on $d^i$ and tested it on all the dataset in $D$ except $d^i$. We trained the baseline classifier on the train set of dataset $d^i$ and tested it on the test set of the dataset $D^{j\cdots n}$, where $d^j$ is the dataset in $D$ that is not the same as the test set of the dataset $d^i$, as demonstrated in Algorithm 1. This approach is similar to the transfer learning approach, however, the classifiers here are not fine tuned on the test sets.
  This evaluation method allows us to assess the dataset's suitability for training a deep learning model that can be effectively generalized to other datasets, which may exhibit different features compared to the one used for training. These variations may include differences in data content published on diverse online platforms, encompassing variations in users' writing styles, content topics, dataset sizes, and other relevant characteristics. We utilize this evaluation to determine each dataset's suitability for producing a deep learning model capable of generalizing to classifying "Hate" and "Not-Hate" content published on different online platforms.

**Algorithm 1** Generalized learning evaluation

**Require:** Dataset set $D = \{d^1, d^2, \dots, d^n\}$
1: **for** $i = 1$ to $n$ **do**
2: Train the model on $d^i$
3: **for** $j = 1$ to $n$ **do**
4: Test the model on $d^j \in D$ and $j \neq i$
5: **end for**
6: **end for**

TABLE V: System performance - mono-dataset classifier performance rank based on weighted F1 score

| **Model** | Weighted F1-score | |
|---|---|---|
| Davidson et al. [5] | 0.930 | 1 |
| Suryawanshi et al. [7] | 0.902 | 2 |
| Salminen et al. [6] | 0.884 | 3 |
| Waseem and Havoy [8] | 0.879 | 4 |
| Kennedy et al. [2] | 0.840 | 5 |
| Qian et al. [13] | 0.816 | 6 |
| Vidgen et al A. [15] | 0.789 | 7 |
| Gibert et al. [14] | 0.777 | 8 |
| Vidgen et al. B [12] | 0.740 | 9 |
| Gomez et al. [3] | 0.697 | 10 |

TABLE VI: System performance - mono-dataset classifier

| **Model** | Weighted F1 | Recall | Precision |
|---|---|---|---|
| Gomez et al. [3] | 0.697 | 0.694 | 0.712 |
| Vidgen et al. B [12] | 0.740 | 0.739 | 0.745 |
| Gibert et al. [14] | 0.777 | 0.774 | 0.803 |
| Vidgen et al. [15] | 0.789 | 0.789 | 0.790 |
| Qian et al. [13] | 0.817 | 0.817 | 0.817 |
| Kennedy et al. [2] | 0.841 | 0.840 | 0.841 |
| Waseem and Havoy [8] | 0.879 | 0.879 | 0.882 |
| Salminen et al. [6] | 0.884 | 0.884 | 0.885 |
| Suryawanshi et al. [7] | 0.902 | 0.902 | 0.910 |
| Davidson et al [5] | **0.930** | **0.930** | **0.930** |

### *B. Results*

We have conducted multiple empirical evaluations of our baseline classifier, which we described in section III-E, on ten publicly available datasets. In this section, we report the empirical evaluation outcomes of the effectiveness of each dataset for training and testing a baseline deep learning classifier to examine the suitability of different public dataset for hatespeech classification.

For each dataset evaluation, we conducted two experiments: (i) we evaluated the suitability of each dataset for training and testing a baseline classifier for the binary classification of hatespeech content. The training and testing samples are from the same dataset. Thus, we refer to this experiment as "mono-dataset experiment", and we refer to the baseline classifier in this experiment as "mono-dataset classifier". (ii) The second experiment involved testing the suitability of each dataset to produce a generalized baseline classifier, which has the same architecture as the mono-dataset classifier but trained and tested in a generalizable approach. We trained the baseline classifier on a dataset and tested it on nine other datasets, excluding the dataset that we used for training the classifier. We refer to this experiment as "generalized learning experiment", and we refer to the baseline classifier as "generalized classifier".

For each experiment–due to space limitation– we report the system performance using weighted F1-score, which is based on the harmonic mean of recall and precision.

Table V, shows the ranking of different datasets in ascending order based on the classifier's performance, measured by weighted F1-score.

### *C. Mono-Dataset Experiment.*

Table VI presents the performance of the mono-dataset classifier. Out of the ten selected datasets, the model performs best when trained on the dataset published by Davidson et al. [5] achieving a weighted F1-score of 0.930. The second-best performance is based on the Suryawanshi et al. [7] dataset with a weighted F1-score of 0.902, which is slightly behind Davidson et al. [5]'s dataset . The classifier performed worst when trained and tested on the dataset published by [3], producing a weighted F1-score of 0.697.

The classifier achieved moderate performance (between 0.81 and 0.88 of weighted F1-score) when evaluated on the following datasets: Qian et al. [13] (0.817), Kennedy et al. [2] (0.841), Waseem and Havoy [8] (0.879), and Salminen et al. [6] (0.884). Furthermore, the classifier produced a weighted F1-score between 0.71 and 0.78 when trained and tested on the following datasets: Vidgen et al. [12] (0.740), Gibert et al. [14] (0.777), and Vidgen et al. A [15] (0.789).

The dataset from Gomez et al. [3] is the least effective for training and testing a baseline neural network classifier, despite being a large dataset, which, theoretically, should be beneficial for deep neural network algorithm training. Davidson et al. [5]'s dataset is ranked first for training a baseline classifier, with Suryawanshi et al. [7]'s dataset (ranked second) narrowly behind. Several other datasets performed moderately: Salminen et al. [6] (0.884), Waseem and Havoy [8] (0.879), Kennedy et al. [2] (0.840) and Qian et al. (0.816). The dataset from Vidgen et al. A [15], Gibert et al. [14], and Vidgen et al. [12] produced weight F1-score of 0.789, 0.777 and 0.740, respectively.

In comparison with the reported hatespeech classifiers proposed by some of the authors of the published datasets, Table VII the recall, precision, and F1-score of our proposed baseline system compared against other published binary hatespeech classifiers. The table demonstrates that in some cases, our baseline system trained on the binary labels of the selected datasets outperforms the binary systems proposed by the dataset authors.

TABLE VII: System performance comparison - mono-dataset classifier performance against published works.

| **Datasets** | Our baseline system | | | Published systems by dataset authors | | |
|---|---|---|---|---|---|---|
| | Recall | Precision | Weighted F1 | Recall | Precision | Weighted F1 |
| Davidson et al. [5] | 0.930 | 0.930 | 0.930 | 0.90 | 0.91 | 0.90 |
| Salminen et al. [6] | 0.884 | 0.885 | 0.884 | _ | _ | 0.96 |
| Waseem and Havoy [8] | 0.879 | 0.882 | 0.879 | 0.729 | 0.774 | 0.739 |
| Qian et al. [13] | 0.817 | 0.817 | 0.816 | _ | _ | 0.896 |

### *D. generalized Learning Experiment.*

In this experiment, we evaluated our baseline classifier by training it on one dataset and testing it on nine other datasets, excluding the dataset used for training. We repeated the experiment for all the ten datasets. The results from this experiment provide a clear indication of the suitability of each dataset for producing a classifier that can be generalized to unseen data. We refer to the baseline classifier in this experiment as "generalized classifier".[3]

Table VIII shows the performance of the classifier during this experiment. The first column contains the dataset used for training the classifier. The other columns (column 1 to

[3]Note: this type of evaluation could also be referred to as "transfer learning".

TABLE VIII: System performance: generalized classifier performance based on weighted F1-score

| **Model\Dataset** | Davidson et al. | Waseem and Havoy | Vidgen et al. B | Salmenin et al. | Gomez et al. | Kennedy et al. | Vidgen et al. A | Suryawanshi et al. | Qian et al. | Gibert et al. | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Gomez et al. [3] | 0.292 | 0.373 | 0.363 | 0.341 | * | 0.395 | 0.425 | 0.486 | 0.253 | 0.341 | 0.363 |
| Suryawanshi et al. [7] | 0.506 | 0.625 | 0.561 | 0.602 | 0.481 | 0.580 | 0.446 | * | 0.556 | 0.425 | 0.531 |
| Waseem and Havoy [8] | 0.618 | * | 0.509 | 0.463 | 0.459 | 0.519 | 0.505 | **0.722** | 0.530 | 0.467 | 0.532 |
| Kennedy et al. [2] | 0.803 | 0.427 | 0.574 | 0.726 | 0.392 | * | 0.523 | 0.468 | 0.662 | 0.655 | 0.581 |
| Davidson et al. [5] | * | 0.435 | 0.586 | 0.753 | 0.410 | **0.729** | 0.519 | 0.452 | 0.714 | 0.647 | 0.583 |
| Gibert et al. [14] | 0.623 | 0.415 | 0.557 | 0.750 | **0.578** | 0.653 | 0.578 | 0.441 | 0.690 | * | 0.587 |
| Salmenin et al. [6] | 0.751 | 0.580 | 0.610 | * | 0.467 | 0.723 | 0.477 | 0.445 | 0.723 | 0.618 | 0.599 |
| Vidgen et al A. [15] | 0.762 | 0.516 | 0.590 | 0.596 | 0.443 | 0.692 | * | 0.419 | 0.704 | 0.718 | 0.605 |
| Vidgen et al. B [12] | 0.782 | 0.653 | * | 0.808 | 0.389 | 0.541 | **0.583** | 0.571 | **0.756** | **0.752** | 0.649 |
| Qian et al. [13] | **0.821** | **0.661** | **0.678** | **0.820** | 0.474 | 0.595 | 0.579 | 0.548 | * | 0.724 | **0.656** |
| Mean | **0.662** | 0.521 | 0.559 | 0.651 | 0.455 | 0.603 | 0.515 | 0.506 | 0.621 | 0.594 | |

TABLE IX: System performance: generalized classifier ranking based on overall mean weighted F1-score, list in ascending order by Mean score

| **Model** | Weighted F1-score | Rank |
|---|---|---|
| Qian et al. [13] | 0.656 | 1 |
| Vidgen et al. B [12] | 0.649 | 2 |
| Vidgen et al. A [15] | 0.605 | 3 |
| Salminen et al. [6] | 0.599 | 4 |
| Gibert et al. [14] | 0.587 | 5 |
| Davidson et al. [5] | 0.583 | 6 |
| Kennedy et al. [2] | 0.581 | 7 |
| Suryawanshi et al. [7] | 0.531 | 8 |
| Waseem and Havoy [8] | 0.532 | 9 |
| Gomez et al. [3] | 0.363 | 10 |

10) contain the datasets used for testing the performance of the classifier. An asterisk '*' in each column indicates the classifier is not tested on the dataset specified in that column. For example, the '*' in the first row of the second column indicates the classifier is trained on Davidson et al.'s [5] dataset but not tested on that dataset. This is due to the nature of transfer learning method. Thus, there is one '*' in each row. The numbers in the rows represent the performance of each classifier when tested on all the datasets except the one that is used for training.

The database published by [13] is one of the most suitable one for producing a generalized deep learning baseline classifier that performs well on multiple hatespeech datasets. When the baseline classifier was trained on Qian et al. [13]'s dataset, it performed well on four out of nine datasets (namely, Davidson et al. [5], Waseem and Havoy [8], Vidgen et al. A [15], and Salmenin et al. [6]). The classifier's performance on this dataset is shown in bold in the 10th row of Table VIII.

Second to Qian et al. [13]' dataset is Vidgen et al. B [12]'s dataset, which performed well on three out of nine datasets (namely, Vidgen et al. [15], Qian et al. [13], and Gibert et al. [14]). The classifier performed well on only one out of nine datasets when trained on the following dataset: Davidson et al. [5], Waseem and Havoy [8], and Gibert et al. [14]. The classifier trained on Salmenin et al. [6], Gomez et al. [3], Kennedy et al. [2], Vidgen et al. [12], and Suryawanshi et al. [7] did not perform better than those trained on other datasets. However, it should be noted that their mean score across the nine datasets affected their ranking performance.

Table IX shows the ranking of the generalized classifier performance for each dataset based on the overall mean performance across the nine datasets used for testing the classifier. The average mean score of the classifier indicates the generalization level of the classifier when trained on each dataset for hatespeech binary classification, i.e., the suitability of a dataset for training a baseline deep learning classifier in a generalized learning setting.

The classifier ranked first when trained on the Qian et al. [13] dataset and tested on nine other datasets, achieving 0.656. The classifier trained on the Vidgen et al. B [12] dataset ranked second with a mean of 0.649The classifier performed worst when trained on Gomez et al. [3] dataset and tested on the other nine datasets, achieving the lowest mean weighted F1-score of 0.363. This poor performance indicates that the Gomez et al. [3] dataset is the least suitable for producing a baseline deep learning classifier in a generalized learning setting.

The other datasets (Vidgen et al. A. [15], Salminen et al. [6], Gibert et al. [14], Davidson et al. [5], Kennedy et al. [2], Suryawanshi et al. [7], and Waseem and Havoy [8]) achieved mean weighted F1-score between 0.532 and 0.599, as shown in Table IX.

## V. Discussion

The classifier performed well when trained and tested on a single dataset at a time, which we referred to in Section IV as mono-dataset experiment. It produced a weighted F1-score between 0.81 and 0.93 for six out of ten datasets that we used for training and testing the classifier. In order to evaluate each datasets, we applied a generalized learning approach in a second experiment, referred to as "generalized learning experiment". In this experiment, we trained the classifier on one dataset and tested it on nine other datasets. This approach allows us to rigorously examine the suitability of a dataset for training and testing a classifier in a generalized learning setting, which provides a more rigorous evaluation than mono-dataset evaluation.

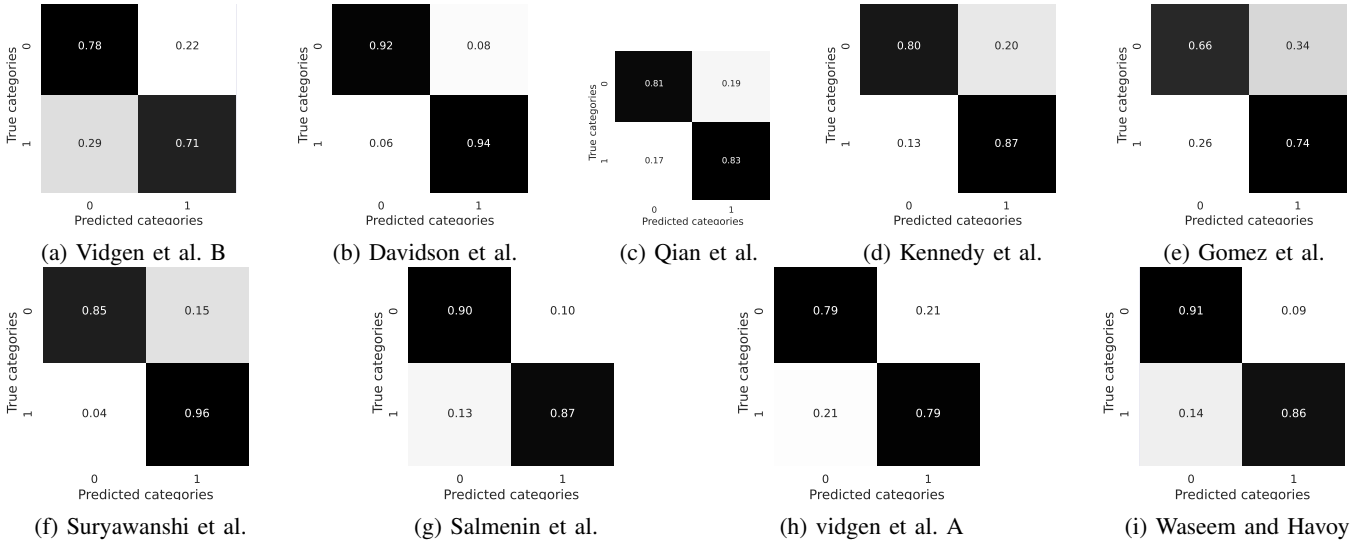
We found that the classifier's performance varied depending on the dataset used for training. In this section, we will highlight some of the major features of the datasets that have potentially influenced the classifier's performance. Additionally. we will present several confusion matrices to illustrate some of the classification errors the classifier made when applied to different dataset.

### A. *p-Test*

We have calculated two statistical measures ($P-test$ and $T-test$) using the content of each dataset based on the labels "Hate" or "Not-Hate". Details of our approach to computing these statistics are presented in section III-D. Since machine learning algorithm performance is based on its learning from

This article has been accepted for publication in IEEE Transactions on Artificial Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAI.2025.3564605






(a) Vidgen et al. B (b) Davidson et al. (c) Qian et al. (d) Kennedy et al. (e) Gomez et al.

(f) Suryawanshi et al. (g) Salmenin et al. (h) vidgen et al. A (i) Waseem and Havoy

Fig. 3: Confusion matrix: mono-dataset classification error analyses

TABLE X: Hate/Not-Hate Hate terms t-test

| **Dataset** | T-value | P-value | Rank P-value | Rank by T-value |
|---|---|---|---|---|
| Qian et al [13] | -3.9461 | 0.0001 | 1 | 1 |
| Gibert et al. [14] | -3.5345 | 0.0004 | 2 | 2 |
| Kennedy et al [2] | -3.4692 | 0.0005 | 3 | 3 |
| Vidgen et al. B [12] | -3.4119 | 0.0007 | 4 | 4 |
| Vidgen et al. A [15] | -2.3177 | 0.0205 | 5 | 5 |
| Waseem and Havoy [8] | -1.7845 | 0.0744 | 6 | 6 |
| Salmenin et al. [6] | -1.7490 | 0.0804 | 7 | 7 |
| Suryawanshi et al. [7] | -1.6805 | 0.0930 | 8 | 8 |
| Davidson et al. [5] | -1.6402 | 0.1011 | 9 | 9 |
| Gomez et al. [3] | 0.0067 | 0.9947 | 10 | 10 |

identifiable patterns in a given dataset, the $p$ value offers good indication on the available patterns in the ten datasets we have chosen for this study. A $p$ value of 0 indicates that the patterns in the dataset occurred by chance, which may reflect poor dataset annotation. A $p$ value of 1 indicates there is no difference in the patterns in the dataset. Our data analyses results are shown in Table X.

The result of our analyses highlighted that the dataset from [3] has consistently performed poorly in both of the experiments: mono-dataset experiment and generalized learning experiment. From table X it can be noted that $p$ the value for this dataset is very close to 1 (0.9947), which means there is no recognizable pattern between "Hate" and "Not-Hate" content in this dataset. The lack of distinguishable patterns in the dataset highlights the main reason for the baseline classifier failing to learn sufficiently from this dataset, hence performing poorly, producing a weighted F1-score of 0.363 in the generalized learning experiment.

In contrast, as can be seen from table IX, the classifier produced a weighted F1-score of more than 0.5 for all those datasets with $p$ value 0.0001 and ¡0.1011, indicating that the model learned sufficient patterns to produce a weighted F1-score of over 0.531.

### *B. Confusion matrix*

In supervised text classification, machine learning algorithms learn from a set of labeled data. Any given labeled dataset contain annotation errors due to many reasons (e.g., annotation procedure, annotator competency, data quality checking, ambiguities in natural language, etc.). Thus, machine learning algorithms are expected to make mistakes since they learn from annotated data. We use confusion matrices to highlight the classification errors the proposed model made in each experiment (mono-dataset and generalized learning experiments). Due to space limitations, we provide comprehensive details on the classification errors of the best and worst performing classifiers compared to other models.

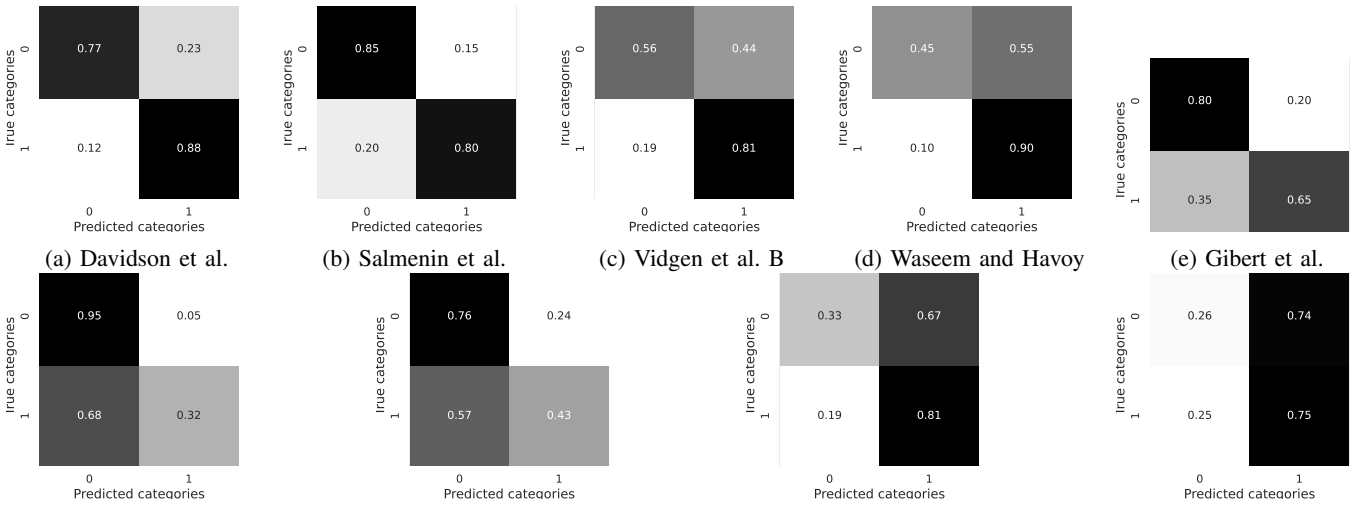
*1) mono-dataset classifier error analyses:* Training the classifier on the dataset published by Davidson et al. [5] produced the lowest classification error rate. The classifier correctly classified 92% "Not-Hate' content and 94% "Hate" content. However, the classifier miss-classifies 8% of "Not-Hate" content as "Hate" and 6 of it "Hate" content as "Not-Hate". The classifier's misclassification total error rate between the classes "Hate" and "Not-Hate" is 14%, as shown in Fig 3b, which is lower compared to when the classifier is trained on other dataset.

Comparing the confusion matrix in Fig 3, it appears that the largest misclassification error rate is produced by the classifier when trained on the Gomez et al. [3] dataset, as shown in Fig 3e. The classifier makes a 26% misclassification error rate for "Hate" content and a 34% error rate for "Not-Hate' content. The classifier seems to perform consistently when trained on Vidgen et al. A's [15] dataset, correctly classifying 79% of both"Hate" and "Not-Hate" content, with 21% error rate for both types of contents. This 21% misclassification error rate, when training the classifier on Vidgen et al. A's [15] dataset, is the second-largest misclassification error rate generated by the classifier after the Gomez et al. [3] trained classifier. The confusion matrix in Fig 3h presents the classification error rate of the classifier when trained on Vidgen et al. A. [15]

The confusion matrices for Vidgen et al. B (Fig 3a), Salmenine et al. (Fig 3g) and Waseem and Havoy (Fig 3i) show that the classifier performs better at classifying "Not-Hate" content than "Hate" content, with a lower misclassification rate

This article has been accepted for publication in IEEE Transactions on Artificial Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAI.2025.3564605






(a) Davidson et al. (b) Salmenin et al. (c) Vidgen et al. B (d) Waseem and Havoy (e) Gibert et al.

(f) Kennedy et al. (g) vidgen et al. (h) Suryawanshi et al. (i) Gomez et al.

Fig. 4: Confusion matrix: generalised model train on Qian et al. dataset and tested on multiple dataset

for "Not-Hate" content. In contrast, the confusion matrices for Davidson et al. (Fig 3b), Qian et al. (Fig 3c), Kennedy et al. (Fig 3d), Gomez et al. [3] (Fig 3e), and Suryawanshi et al. (Fig 3f) show the model miss-classifies "Hate" content less than "Not-Hate" content.

We found that five out of the ten trained classifiers on different dataset, misclassify "Not-Hate" content more than "Hate" content. Three trained classifiers misclassify "Hate" content more than "Not-Hate" content. The exception is the classifier trained on Vidgen et al. A's dataset [15], which has equal error rates for both "Hate" and "Not-Hate".

The lowest misclassification rate (4%) for "Hate" content comes from the classifier trained on the Suryawanshi et al. dataset, indicating that this model achieves the highest correct classification rate (96%) for "Hate" content. The smallest misclassification error rate for "Not-Hate" comes from the classifier trained on Davidson et al. [5], which means that this classifier achieves the highest correct classification rate (92%) for "Not-Hate" content.

Although the model trained on the Suryawanshi et al. dataset produces the highest correct classification rate of "Hate" content, it falls behind the classifier based on Davidson et al. [5] dataset due to its misclassification error rate for "Not-Hate" content. In the Suryawanshi et al. based classifier is nearly twice as high as that of the Davidson et al.'s dataset-based classifier (15% vs 8%).

*2) generalized model's error analyses:* In this section, we examine the confusion matrix graphs to analyze the errors made by each classifier when evaluated on transfer learning performance.

In the generalized learning experiment, we trained our classifier on one dataset and tested it on the remaining nine datasets. This experiment produced ten classifiers, each of which tested on nine datasets, excluding the one was used for training. For each experiment, we obtained one confusion matrix, resulting in nine confusion matrices per classifier. Therefore, this substantial experiment produced a total of ninety confusion matrices. Due to space limitations, we will focus our discussion only on a subset of confusion matrices. Specifically, we will examine the classification errors of the proposed model that have the highest or the lowest mean score of weighted F1-score when tested on nine different datasets. As shown in Table IX, the highest weighted F1-score (0.656) produced by the model trained on the Qian et al. [13]'s dataset, while the lowest weighted F1-score (0.636) was produced by the model trained on Gomez et. al. [3]'s dataset.

The confusion matrices in Fig 4 and 5. are presented in ascending order based on the weighted F1-score, which is shown in Table VIII.

We grouped our analyses of the confusion matrices based on the following criteria: i) the proposed model produced the highest mean weighted F1-score compared to the other nine models, ii) the proposed model produced the lowest mean weighted F1-score.

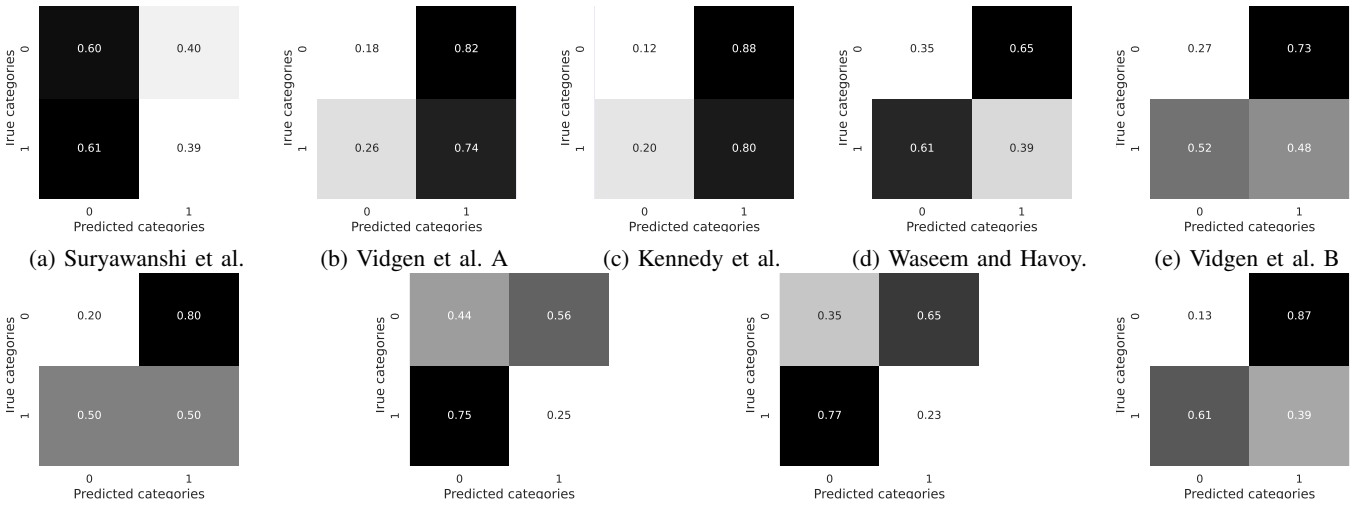
**generalized model trained on Qian et al. dataset** The model trained on the Qian et al. [13] dataset performed the best on four out of nine datasets: Davidson et al. [5], Salmenine [6], Vidgen et al. B [12], and Waseem and Havoy [8] ( 4a 4d). In these cases, the model correctly classified "Hate" content more accurately than "Not-Hate" content, with the exception of the Salminen et al. [6] dataset, where that model classified "Not-Hate" more accurately than "Hate" by a margin of 5%.

Although the model didn't produce a high weighted F1-score when tested on the Kennedy et al. [2] dataset, it appears to perform well in correctly classifying "Not-Hate" content. As shown in Fig 4f, the model has misclassification error rate of just 5% for "Not-Hate" content. However, the model has large misclassification error rate 68% for "Hate" content, which is the main reason the model didn't produce high weighted F1-score on this dataset compared to the performance on other dataset.

In contrast, the model's large misclassification error rate of 74% for "Not-Hate" content also negatively impacts the model's weighted F1-score. The Suryawanshi et al. [7]. dataset seem to come second after Gomez et al. [3] dataset in misclassifying "Not-Hate" content, with an error rate of 67%.

This article has been accepted for publication in IEEE Transactions on Artificial Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAI.2025.3564605






(a) Suryawanshi et al. (b) Vidgen et al. A (c) Kennedy et al. (d) Waseem and Havoy. (e) Vidgen et al. B

(f) Salmenin et al. (g) Gibert et al. (h) Davidson et al. (i) Qian et al.

Fig. 5: Confusion matrix: generalized classifier trained on Gomez et al. dataset and tested on multiple dataset

**generalized model trained on Gomez et al. dataset** Fig 5 shows the confusing matrix for the model trained on Gomez et al. [3] dataset.

The model trained on the Gomez et al. [3]. dataset demonstrated the worst performance compared to all the other models . We refer to this model as "Gomez et al. model". The confusion matrix showing the errors made by this model when tested on nine datasets is shown in Fig 5.

The model higher misclassification error rate on "not-Hate" content than "Hate" content, with 61% of "Not-Hate" being classified as "Hate", as shown in the Confusion matrix in Fig 5a. In contrast, the model makes significant misclassification errors for "Hate" content, as presented in the confusion matrices in Fig 5b and 5c, where the misclassification error rate of "Hate" content exceed 80%.

For several other datasets, the model seems to struggle to correctly classify either types of content ("Hate" and "Not-Hate"). For the Waseem and Havoy [8] dataset, the correct classification rate does not exceed 35%, as shown in Fig 5d. For the Gibert et al. [14], Davidson et al. [5] and Qian et al. [13] datasets, the model has a misclassification error rate between 56% to 87% for either "Hate" or "Not-Hate" content. The exception is the Salmenine et al. [6] dataset where the model make 50/50 misclassification error of "Not-Hate" content, but a large error rate of 80% for "Hate" content.

*3) Model performance analysis:* To develop an efficient hatespeech classifier, it is crucial to train the model on a high-quality dataset that provides sufficient information for accurate classification and real-world implementation. To gain a deeper understanding of the role of features in the effectiveness of the classifier, we conducted a correlation test between the dataset features and the classification F1-score. This analysis helps to identify the most informative features and optimize the dataset for classifier performance. Fig 6 presents the results of our correlation analysis for the mono dataset test, showing the Pearson correlation between each feature and the F1-score of the classifier. This information can be used to identify the most informative features of the dataset.

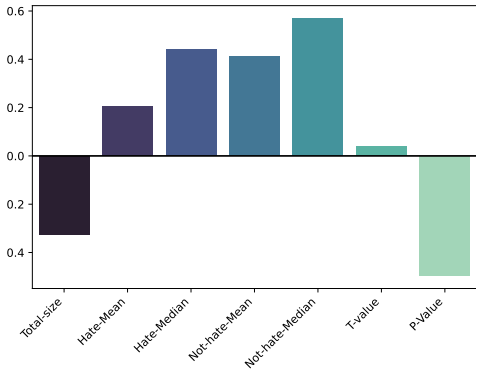



Fig. 6: Pearson correlation between each feature and the F1-score of the hatespeech classifier for the mono dataset

TABLE XI: The features for the ML model. WC = Word Count.

| | Features (X) | | | | | | | Target (y) |
|---|---|---|---|---|---|---|---|---|
| | | Hate Word Count | | Not Hate WC | | Hate Terms | | Mono |
| Model | Total-size | Mean | Median | Mean | Median | T-value | P-Value | F1 |
| Vidgen et al. B | 8186 | 39.5 | 19 | 27.2 | 14 | -3.4119 | 0.0007 | 0.740 |
| Davidson | 2860 | 13.9 | 13 | 17.8 | 15 | -1.6402 | 0.1011 | 0.930 |
| Qian et al. | 4696 | 27.8 | 23 | 19.9 | 15 | -3.9461 | 0.0001 | 0.816 |
| Kennedy | 92042 | 25.8 | 19 | 28 | 21 | -3.4692 | 0.0005 | 0.840 |
| Gomez et al. | 50526 | 11.7 | 11 | 11.4 | 11 | 0.0067 | 0.9947 | 0.697 |
| Vidgen | 37938 | 23.8 | 15 | 25.1 | 17 | -2.3177 | 0.0205 | 0.789 |
| Waseem and Havoy | 5384 | 16.9 | 17 | 14 | 14 | -1.7845 | 0.0744 | 0.879 |
| Suryawanshi et al. | 606 | 45 | 33 | 44.8 | 32 | -1.6805 | 0.093 | 0.902 |
| Salminen | 1716 | 43.6 | 34 | 38.2 | 30 | -1.749 | 0.0804 | 0.884 |
| Gibert et al. | 2874 | 22 | 20 | 17.3 | 15 | -3.5345 | 0.0004 | 0.777 |

Based on our analysis, we observed that the Median word count for the not-hate part of the datasets and the P-value are highly correlated, indicating that these features have a significant impact on the performance of the hatespeech classifier. On the other hand, the T-value is the least correlated feature, suggesting that it may not have a significant impact on the classifier's effectiveness.

This finding highlights the importance of statistical characteristics of the dataset used in the hatespeech classifier, as some characteristics may have a more significant impact on the model's performance than others. By understanding the

TABLE XII: The regression model predicted F1-score and the actual achieved scores

| Dataset | LR | F1-score |
|---|---|---|
| Gomez et al. | 0.700 | 0.697 |
| Vidgen et al. B | 0.724 | 0.740 |
| Gibert et al. | 0.816 | 0.777 |
| Vidgen | 0.847 | 0.789 |
| Qian et al. | 0.800 | 0.816 |
| Kennedy | 0.817 | 0.840 |
| Waseem and Havoy | 0.877 | 0.879 |
| Salminen | 0.883 | 0.884 |
| Suryawanshi et al. | 0.904 | 0.902 |
| Davidson | 0.886 | 0.930 |

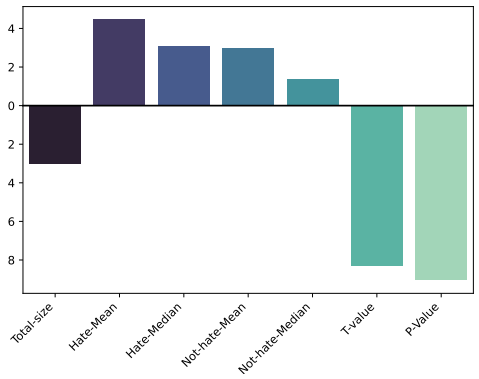



Fig. 7: Pearson correlation between each feature and the F1-score of the hatespeech classifier for the Generalized Learning Experiment.

correlation between the characteristic and the model's effectiveness, we can optimize the dataset balancing and improve the accuracy of the classifier.

To ensure the validity of our correlation test, we implemented a linear regression model that was trained on the statistical characteristics of the dataset and used to predict the F1-score of the hatespeech classifier. Table XI shows the features used in the regression model. The regression model produced accurate predictions of the classification model, with a coefficient of determination of 0.84. This high level of accuracy demonstrates the reliability of our correlation test and supports the conclusion that the identified features have a significant impact on the effectiveness of the classifier. The predicted values and the achieved F1-scores are presented in Table XII.

To optimize the hatespeech classifier, we conducted an analysis using the average F1-score in the generalized learning Experiment as the metric for evaluating the classifier's performance. Our analysis revealed that the P-value is still one of the most highly correlated dataset characteristics for the classifier's effectiveness, as shown in Fig 7.

However, we also observed that the Median Not-Hate word count, which was previously highly correlated in our initial analysis, has now dropped to become the least correlated characteristic of the training dataset in the generalized Learning Experiment. This finding suggests that the importance of certain dataset characteristics may vary depending on the specific experimental conditions, highlighting the importance of comprehensive analysis to identify the optimal dataset for training the hatespeech classifier.

By using these insights to fine-tune the dataset, we can improve the accuracy of the hatespeech classifier in real-world applications, making it a more effective tool for identifying and combating hatespeech.

## VI. Conclusions and Future Work

Despite myriad benefits of social network platforms for users and businesses, malicious users abuse them by targeting specific users based on their identity. This phenomenon is referred to as Hatespeech, where malicious users target vulnerable people with abusive text, or graphics, to degrade and cause them harm. The severity of hatespeech on victims has forced researcher, social media platforms, and governments to take action to eliminate it. The task of eliminating hatespeech require automation, since manual intervention is time-consuming and expensive for humans to perform. Thus, large numbers of annotated dataset are developed for training machine learning algorithms to automate the detection and classification of hatespeech. However, the available datasets have been annotated with different types of hatespeech in a number of different ways. This introduces various constraints that affect the training of machine learning algorithms on them. In this study, we have conducted extensive empirical evaluation of multiple hatespeech datasets to examine their suitability for training machine learning algorithms for the task of automated hatespeech classification. We proposed a baseline deep learning model that appears to have reasonable generalization capability across multiple datasets, and it outperforms some of the existing models. Our main contributions in this study are as followings: (i) We present the extensive empirical evaluation of many hatespeech datasets, (ii) We empirically demonstrate that the quality of dataset content has a greater positive impact on AI hatespeech classification than factors such as content volume, context diversity, and data modalities, iii) We propose novel approach to extract and use statistical features from hatespeech dataset and use certain machine learning algorithms to predict deep learning algorithms performance on hatespeech classification, and (iv) We offer a baseline deep learning architecture for automated hatespeech classification.

To complement the dataset evaluation, we have conducted statistical data analyses on the dataset to examine their features. Moreover, we have analyzed the system output to highlight and compare systems' error rates and error types based on each dataset.

We have identified several future works. The binary classification of hatespeech could be the first step in automated hatespeech processing. In this project, we demonstrated the strengths and weaknesses of several public hatespeech datasets using a binary classification approach. Multi-label classification, which is helpful for identifying specific types of hatespeech, would be helpful in tasks that require identifying specific hatespeech content, such as hatespeech based on race, gender, sexuality, religion, etc. Most public datasets contain different types of hatespeech, but they lack consistency. A

second area for future research involves enhancing the content quality of datasets that the baseline classifier struggles to effectively learn from them. Our plan is to explore automated methods for relabeling the content of selected datasets. Additionally, in our future work, we intend to evaluate datasets that encompass a common set of hatespeech types. By addressing these aspects, we aim to improve the overall performance and robustness of hatespeech classification models.

## REFERENCES


[1] A. J. Soto, C. Ryan, F. P. Silva, T. Das, J. Wolkowicz, E. E. Milios, and S. Brooks, "Data quality challenges in twitter content analysis for informing policy making in health care," in *Hawaii International Conference on System Sciences*, 2018.

[2] C. J. Kennedy, G. Bacon, A. Sahn, and C. von Vacano, "Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application," 2020.

[3] R. Gomez, J. Gibert, L. Gómez, and D. Karatzas, "Exploring hate speech detection in multimodal publications," 2019.

[4] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, (New York, NY, USA), p. 29–30, Association for Computing Machinery, 2015.

[5] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, pp. 512–515, May 2017.

[6] J. Salminen, H. Almerekhi, M. Milenković, S.-g. Jung, J. An, H. Kwak, and B. J. Jansen, "Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media," in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[7] S. Suryawanshi, B. R. Chakravarthi, M. Arcan, and P. Buitelaar, "Multimodal meme dataset (multioff) for identifying offensive content in image and text," in *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pp. 32–41, 2020.

[8] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on Twitter," in *Proceedings of the NAACL Student Research Workshop*, (San Diego, California), pp. 88–93, Association for Computational Linguistics, jun 2016.

[9] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, (Valencia, Spain), pp. 1–10, Association for Computational Linguistics, Apr. 2017.

[10] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, (Republic and Canton of Geneva, CHE), p. 145–153, International World Wide Web Conferences Steering Committee, 2016.

[11] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in *Proceedings of the Second Workshop on Language in Social Media*, (Montréal, Canada), pp. 19–26, Association for Computational Linguistics, jun 2012.

[12] B. Vidgen, D. Nguyen, H. Margetts, P. Rossini, R. Tromble, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, *et al.*, "Introducing cad: the contextual abuse dataset," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2289–2303, Association for Computational Linguistics, 2021.

[13] J. Qian, A. Bethke, Y. Liu, E. Belding, and W. Y. Wang, "A benchmark dataset for learning to intervene in online hate speech," 2019.

[14] O. De Gibert, N. Perez, A. García-Pablos, and M. Cuadros, "Hate speech dataset from a white supremacy forum," *arXiv preprint arXiv:1809.04444*, 2018.

[15] B. Vidgen, T. Thrush, Z. Waseem, and D. Kiela, "Learning from the worst: Dynamically generated datasets to improve online hate detection," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Online), pp. 1667–1682, Association for Computational Linguistics, Aug. 2021.

[16] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, (New York, NY, USA), p. 1980–1984, Association for Computing Machinery, 2012.

[17] G. Njagi, Dennis, Z. Zhang, Z. Zuping, D. Hanyurwimfura, and L. Jun, "A lexicon-based approach for hate speech detection," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, pp. 215–230, 2015.

[18] A. Dmonte, T. Arya, T. Ranasinghe, and M. Zampieri, "Towards generalized offensive language identification," 2024.

[19] A. Mazari, N. Boudoukhani, and A. Djeffal, "Bert-based ensemble learning for multi-aspect hate speech detection," *Cluster Computing*, vol. 27, p. 325–339, 2024.

[20] A. A. Hind Saleh and K. Moria, "Detection of hate speech using bert and hate speech word embedding with deep model," *Applied Artificial Intelligence*, vol. 37, no. 1, p. 2166719, 2023.

[21] T. Wullach, A. Adler, and E. Minkov, "Character-level hypernetworks for hate speech detection," *Expert Systems with Applications*, vol. 205, p. 117571, 2022.

[22] K. Mnassri, P. Rajapaksha, R. Farahbakhsh, and N. Crespi, "Bert-based ensemble approaches for hate speech detection," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, pp. 4649–4654, 2022.

[23] R. Ali, U. Farooq, U. Arshad, W. Shahzad, and M. O. Beg, "Hate speech detection on twitter using transfer learning," *Computer Speech & Language*, vol. 74, p. 101365, 2022.

[24] S. Veerasamy, Y. Khare, A. Ramesh, A. S, P. Singh, and A. T, "Hate speech detection using mono bert model in custom content-management-system," in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 1681–1686, 2022.

[25] M. Bilal, A. Khan, S. Jan, S. Musa, and S. Ali, "Roman urdu hate speech detection using transformer-based model for cyber security applications," *Sensors*, vol. 23, no. 8, 2023.

[26] A. J. Keya, M. M. Kabir, N. J. Shammey, M. F. Mridha, M. R. Islam, and Y. Watanobe, "G-bert: An efficient method for identifying hate speech in bengali texts on social media," *IEEE Access*, vol. 11, pp. 79697–79709, 2023.

[27] J. A. Maha, "Bert-based approach to arabic hate speech and offensive language detection in twitter: Exploiting emojis and sentiment analysis," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 5, 2022. Copyright - © 2022. This work is licensed under https://creativecommons.org/licenses/by/4.0/ (the "License"). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2023-12-04.

[28] F. M. P. del Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia, "Comparing pre-trained language models for spanish hate speech detection," *Expert Systems with Applications*, vol. 166, p. 114120, 2021.

[29] SciPy 1.0 Contributors, "SciPy 1.0: Fundamental algorithms for scientific computing in python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[30] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.


**Dr. Sardar Jaf** was awarded PhD in natural language processing from the University of Manchester (UK) in 2015. He has been an active academic and researcher since 2012. He has published and presented many high quality conferences and journal articles in many international venues. He has conducted multidisciplinary research with Social Science and Psychology academics and researchers. He has also collaborated with industry partners and artists, showcasing his research impact beyond academia. Dr. Jaf is a regular reviewer for several international journals (IEEE Access, International Journal of Molecular Diversity Preservation, International and Multidisciplinary Digital Publishing Institute (MDPI) - International Journal of Human-Computer. Springer Nature etc.). He is also grant application reviewer for several funding organizations (e.g., EPSRC. ESRC. UKRI, Swiss National Science Foundation (SNSF), etc.).

**Dr. Basel Barakat** is a Senior Lecturer in the Faculty of Business and Technology at the University of Sunderland (UK). He is a Fellow of the Advanced Higher Education Academy (FHEA) and has received several awards for his teaching and research, including the 'Top10 2021 MDPI Future Internet High Cited Series Paper' for his work on 6G, and the Champion of Champions award from the Royal Academy of Engineering.

Dr Barakat is a member of the University of Sunderland Academic Board (the highest level academic authority). He is leading the university Research Excellence Framework for $11^{th}$ unit of assessment. He is also a member of the EPSRC funding peer review college.

Prior to joining the University of Sunderland, Dr. Barakat worked as a Lecturer at Edinburgh Napier University and as a Research Fellow at the University of Greenwich. He received his MSc and PhD from the University of Greenwich (UK) in 2014 and 2019, respectively. In 2018, he was a visiting scholar at the University of Cambridge.