Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

MetaHate: A Dataset for Unifying Efforts on Hate Speech Detection

Paloma Piot, Patricia Martín-Rodilla, Javier Parapar

IRLab, CITIC Research Centre, Universidade da Coruña
paloma.piot@udc.es, patricia.martin.rodilla@udc.es, javier.parapar@udc.es

Abstract

Hate speech represents a pervasive and detrimental form of online discourse, often manifested through an array of slurs, from hateful tweets to defamatory posts. As such speech proliferates, it connects people globally and poses significant social, psychological, and occasionally physical threats to targeted individuals and communities. Current computational linguistic approaches for tackling this phenomenon rely on labelled social media datasets for training. For unifying efforts, our study advances in the critical need for a comprehensive meta-collection, advocating for an extensive dataset to help counteract this problem effectively. We scrutinized over 60 datasets, selectively integrating those pertinent into MetaHate. This paper offers a detailed examination of existing collections, highlighting their strengths and limitations. Our findings contribute to a deeper understanding of the existing datasets, paving the way for training more robust and adaptable models. These enhanced models are essential for effectively combating the dynamic and complex nature of hate speech in the digital realm.

Introduction and Motivation

In recent years, the pervasive influence of social media and online platforms has facilitated unprecedented connectivity and communication on a global scale. However, this interconnectedness has also brought to the forefront a concerning rise in the prevalence of hate speech—an issue that transcends geographical boundaries and cultural differences (Vogels 2021; Hickey et al. 2023). Different studies report that more than 30% of young people have been victims of cyber hate by their peers (Kansok-Dusche et al. 2023). As society navigates this digital age, the urgency to address hate speech and its detrimental consequences has never been more critical. For identifying different types of offensive messages, machine learning models were suggested in 1997 (Spertus 1997). Nowadays, with the rise of Large Language Models (LLMs), the need for vast datasets is crucial (Kurrek, Saleem, and Ruths 2020). In this paper, we scrutinize state-of-the-art hate speech datasets and assess them to organize a large-scale meta-collection for hate speech detection on social media.

There is no formal definition of hate speech, but previous works (Davidson et al. 2017; Founta et al. 2018; Mathew et al. 2020; ElSherief et al. 2018a,b; Silva et al. 2021; Das et al. 2023) deepened on this topic, defining it as "*language characterized by offensive, derogatory, humiliating, or insulting discourse (Founta et al. 2018) that promotes violence, discrimination, or hostility towards individuals or groups (Davidson et al. 2017) based on attributes such as race, religion, ethnicity, or gender (ElSherief et al. 2018a,b; Das et al. 2023)*". Under this definition, which aligns very well with the United Nations one (Nations 2023), we frame our work by differentiating hate speech from non-hate and offensive speech.

One of the challenges faced in hate speech detection is the lack of standardized datasets (ElSherief et al. 2018a; Poletto et al. 2020; Toraman, Şahinuç, and Yilmaz 2022), evaluation metrics (Röttger et al. 2021), and benchmark models (Poletto et al. 2020). The motivation for creating a meta-collection lies in the recognition that individual efforts to combat hate speech are essential but often limited in scope. By consolidating diverse datasets, methodologies, and models from various contributors, a meta-collection serves as a centralized resource that empowers researchers and practitioners to collaborate, learn from each other's experiences, and collectively advance the field of hate speech detection. This collaborative approach not only accelerates progress but also ensures the development of more robust and generalizable models that can adapt to the ever-evolving nature of online hate speech. In this paper, we advocate for the necessity of a meta-collection on hate speech detection as a pivotal step towards fostering generalizable machine learning detection models. By harnessing collective knowledge and resources, the meta-collection endeavours propel the field forward, providing a united front against the proliferation of online hate speech in our interconnected world.

Data Acquisition and Preparation

Over the past few years, numerous efforts have been made to create datasets for hate speech analysis (Davidson et al. 2017; Golbeck et al. 2017; Founta et al. 2018). The community has a widely recognized list that aims to collect all available hate speech corpora: hatespeechdata. While this repository is a valuable source, it only provides a list of various dataset publications and their links. Nevertheless, the

website also studied 63 datasets, of which 25 are in English, focusing on the best practices for creating datasets for detecting hate speech (Vidgen and Derczynski 2020). Studies such as Poletto et al. (2020) have concentrated on studying all available corpora resources to detect hate speech. This has resulted in a comprehensive survey that highlights the numerous benchmark datasets available for evaluating abusive language. Nevertheless, none of these studies has released any data.

Our study analyzes over 60 hate speech detection datasets, selecting those relevant to our topic for our meta-collection. To gather data on hate speech, we conducted a thorough exploration in search engines and repositories. Additionally, we undertook a comprehensive review of academic literature, including works authored by Vidgen and Derczynski (2020); Poletto et al. (2020); Mody, Huang, and de Oliveira (2023), to enhance the scope and depth of our data collection efforts. All accessed datasets were devoid of personal information, ensuring that users' personally identifiable information remained uncompromised. We meticulously filtered all the examined related works to assess their compatibility with MetaHate. Our selection adheres to specific criteria: (1) inclusion of only social media texts authored by humans, excluding datasets derived from alternative sources or synthetic data; (2) incorporation of datasets aligned with a hate speech definition similar to ours, not considering offensive content as hate speech as it doesn't correspond to our specified hate speech criteria; (3) focusing on English datasets, given the high volume of data available and the inherent coherence within our internally collected information.

We attempted to incorporate additional hate speech datasets that are pertinent to our study, such as the constructed in the ElSherief et al. (2018a) study. Unfortunately, these datasets are not publicly accessible, and our attempts to reach out to the authors went unanswered. Details about these datasets, along with other significant hate speech studies that did not align with our work's definition, such as the research conducted by Das et al. (2023), are compiled in MetaHate website[1]. We gathered the data and reported the actual dataset sizes after removing duplicated entries. Table 2 summarized the datasets integrated into MetaHate, together with other abusive datasets. Next, we present the analysed datasets.

Online Harassment 2017 (Golbeck et al. 2017): Golbeck et al. conducted an exploratory analysis of offensive terms to extract hashtags, lexicon and word structures, which they used to scrape data from Twitter for their dataset. They focused on hate speech and followed a binary classification approach. Contributed: 19 838 posts.

OLID 2019 (Zampieri et al. 2019): Zampieri et al. compiled a list of keywords and constructions that are often included in offensive content to scrape tweets and create their dataset. They propose three classification levels, and we've used the first one (A) - the binary level of offensive speech. We found that their definition of offensive content and hierarchy was similar to our definition of hate speech. Contributed: 14 052 posts.

[1]https://irlab.org/metahate.html

HASOC 2019-2021 (Mandl et al. 2019, 2020; Modha et al. 2021): HASOC is a track featured at the FIRE conference that creates resources for identifying hate speech. It was first introduced in 2019, where participants were provided with a dataset of tweets classified as hate and no hate. In the following editions, the same task was proposed for hate classification in English. The 2019 dataset is publicly available, but the 2020 and 2021 datasets require a password that we were unable to obtain despite reaching out to the authors. Contributed: 6981 posts.

A Curated Hate Speech Dataset 2023 (Mody, Huang, and de Oliveira 2023): Mody et al. attempted to construct a comprehensive large dataset. When we initially downloaded their dataset, we discovered 842 334 posts, but after removing duplicates, we were left with 560 385 samples. Their published curated dataset doesn't provide many details and lacks any experimental analysis of the data. Their sources are derived from 18 datasets, but some of them are no longer available, and there are instances of duplicated datasets in their list. Contributed: 560 385.

Measuring Hate Speech 2020 & 2022 (Kennedy et al. 2020; Sachdeva et al. 2022): Collaborative endeavours ended up in the compilation of a dataset sourced from three diverse platforms: Twitter, Reddit, and YouTube. To assess the hatefulness of the content, the authors opted for a linear hate speech scale, employing Rasch item response theory (IRT). Annotator ratings were transformed into this hate scale, where high values ($>$0.5 approx.) indicated more hateful texts. Values between -1 and 0.5 were assigned to texts perceived as neutral or ambiguous, while those below -1 denoted counter or supportive speech. Contributed: 39 565 posts.

Intervene Hate 2019 (Qian et al. 2019): Qian et al. directed their attention to the Reddit and Gab platforms. Their objective was to automatically generate responses for intervention during online conversations containing hate speech. For Reddit, they gathered the top 200 hottest submissions from toxic subreddits and reconstructed the conversation. On Gab, they employed hate keywords to retrieve the original posts and reconstruct the conversation context. Contributed: 45 170 posts.

ETHOS 2022 (Mollas et al. 2022): Mollas et al. meticulously curated two distinct collections, from toxic subreddits and hatebusters: the first comprising 998 comments labelled for the presence or absence of hate speech content, and the second consisting of 443 hate speech messages categorized through multiclass and multilabel classification. Both datasets were meticulously assembled using data from Reddit and YouTube. Contributed: 998 posts.

Hate in Online News Media 2018 (Salminen et al. 2018): Salminen et al. meticulously labelled comments extracted from YouTube videos and Facebook posts associated with an online news and media company that maintained a highly active presence on social media platforms. The labeling process involved categorizing comments into hateful and neutral ones, and additionally specifying the target of the hate. Contributed: 3214 posts.

Supremacist 2018 (de Gibert et al. 2018): The authors randomly sampled posts from Stormfront, a neo-Nazi Inter-

net forum, as scraping data from this forum, had an intrinsic nature of being hateful, and labelled the data on hate and non-hate speech. Contributed: 10 534 posts.

The Gab Hate Corpus 2022 (Kennedy et al. 2022): In an effort to mitigate potential biases introduced by keyword-based strategies, Kennedy et al. opted to build a hate speech dataset from a less conventional social network, Gab. Gab is known as a right-wing social platform where hate or abusive comments are more prevalent, and the researchers chose to randomly sample diverse publications and classify them as an assault on human dignity or not. Contributed: 27 434 posts.

HateComments 2023 (Gupta, Priyadarshi, and Gupta 2023): Gupta et al. have recently created a hate speech dataset from YouTube and BitChute video comments. The comments were tagged on a binary level and the dataset also includes some video context. Contributed: 2070 posts.

Toxic Spans 2021 (Pavlopoulos et al. 2021): Various initiatives addressing the issue of hate speech were undertaken within the framework of the "SemEval-2021" task. The primary goal was to anticipate the specific spans within posts that contributed to their classification as toxic. To accomplish this, a dataset sourced from Civil Comments was meticulously curated and labelled with spans to indicate the precise portions responsible for the toxic labels. Contributed: 10 621 posts.

Ex Machina 2016 (Wulczyn, Thain, and Dixon 2016): In their early endeavours, Wulczyn et al. dedicated their efforts to constructing a sizable dataset by randomly sampling posts and incorporating comments from blocked accounts. Their collection comprises over 100 000 Wikipedia comments, focusing on personal attacks on a binary level. Contributed: 115 705 posts.

Context Toxicity 2020 (Pavlopoulos et al. 2020): Pavlopoulos et al. delved beyond individual messages, investigating the impact of context on hate speech detection. Their research unfolded in two parts: firstly, they constructed a small dataset of 250 comments from Wikipedia to assess how annotation outcomes were influenced by the provision of context. Secondly, they expanded their dataset to include nearly 20 000 Wikipedia comments. Half of the comments were annotated without context, and the remaining half were annotated with context, on toxic or non-toxic speech. Contributed: 19 842 posts.

BullyDetect 2018 (Bin Abdur Rakib and Soon 2018): Bin Abdur Rakib and Soon conducted research on cyberbullying detection by building a binary corpus of posts from Reddit. However, it is not clear how they created the dataset. Their main focus was on training a word embedding model to then train a Random Forest Classifier in order to evaluate their dataset. Contribution: 6562 posts.

US 2020 Elections (Grimminger and Klinger 2021): Grimminger and Klinger directed their hate detection efforts toward the political domain, specifically targeting the 2020 United States Elections. They sampled data from Twitter using search terms related to mentions of presidential candidates, various hashtags reflecting voter alignment, campaign slogans, and even nicknames of the candidates. In their comprehensive analysis, the dataset was meticulously annotated following a binary classification. Contributed: 2999 posts.

"Call me sexist but" 2021 (Samory et al. 2021): The "Call me sexist, but..." dataset was meticulously derived from an extensive pool of over 13 000 tweets. These texts underwent a detailed binary classification process, discerning between instances of sexism and those devoid of such content. Four distinct creation approaches were deployed, including (1) the utilization of sexism scales, (2) incorporating (Waseem and Hovy 2016)'s sexism tweets, (3) integrating benevolent sexism tweets from (Jha and Mamidi 2017), and (4) collecting tweets containing the phrase "call me sexist, but". Contributed: 3058 posts.

Hateval 2019 (Basile et al. 2019): SemEval is a series of workshops that focus on the evaluation and comparison of computational semantic analysis systems. They provided the participants with a dataset, of which the construction strategy focused mainly on fetching tweets by keywords and monitoring haters and victims. In 2019, task 5 topic was hate speech, where the first task was about a binary classification task. Contributed: 12 747 posts.

Hate Offensive 2017 (Davidson et al. 2017): This dataset, curated from hate lexicon, consists of 24 783 entries from Twitter, which have been meticulously categorized into three distinct classes, namely hate speech, offensive language, and neutral language. Contributed: 24 783 posts.

TRAC1 2018 (Kumar et al. 2018): Kumar et al. collaborate with a multilingual dataset, including samples from Twitter and Facebook in English. They fetched data from more than 40 pages discussing controversial topics on Facebook and used popular hashtags around different topics to retrieve data from Twitter. They annotated the data into three different levels of aggressiveness. Contributed: 14 587 posts.

ENCASE 2018 (Founta et al. 2018): This dataset was built by collecting random samples of tweets and boosting a sample that represents 12.5% of the dataset. This sample shows a strong negative polarity and contains at least one offensive word from hate speech lexicons. The tweets were tagged in one of these four different categories: abusive, normal, spam and hateful. Contributed: 91 950 posts.

MLMA 2019 (Ousidhoum et al. 2019): This multilingual work, which includes English posts from Twitter, shares that the authors were able to build a multilabel and multiclass dataset of 5593 entries by looking to tweets that contained keywords and phrases related to hate. Contributed: 5593 posts.

HateXplain 2020 (Mathew et al. 2020): Mathew et al. directed their attention towards examining the bias and interpretability facets of hate speech. They compiled a dataset from Twitter by extracting content using a lexicon and incorporating the Mathew et al. (2019) dataset sourced from the Gab platform. Each post was evaluated by three annotators who assigned labels based on three distinct categories: hate, offensive, and normal. Contributed: 20 109 posts.

Hateful Tweets 2022 (Albanyan and Blanco 2022): This research aimed to construct a corpus for analyzing the context and counter-narratives of hate speech texts. Utilizing works by Waseem and Hovy (2016), Founta et al. (2018), and Davidson et al. (2017), the authors selectively curated entries featuring racist, sexist, abusive, hateful, and offen-

sive content. Subsequently, they sought to retrieve replies and contextual information for these targeted tweets. Contributed: 1141 posts.

Multiclass Hate Speech 2022 (Toraman, Şahinuç, and Yilmaz 2022): The authors created this dataset by using a diverse set of keywords and hashtags spanning various topics, including religion, gender, racism, politics, and sports. They categorized the tweets into hate, offensive, and normal texts. Contributed: 68 597 posts.

Slur Corpus 2020 (Kurrek, Saleem, and Ruths 2020): Almost 40 000 posts sourced from Reddit were sampled using three slurs targeting discrimination across sexuality, ethnicity, and gender, from the Pushshift Reddit dataset (Baumgartner et al. 2020). Kurrek et al. established a comprehensive taxonomy and categorized their dataset into derogatory, non-derogatory, homonym, appropriation, and noise labels. Contributed: 39 960 posts.

TRAC2 2020 (Bhattacharya et al. 2020): In endeavours related to TRAC shared tasks, Bhattacharya et al. constructed a dataset of comments extracted from YouTube videos. Adhering to a methodology akin to that of Kumar et al. (2018), they distinguished between three types of aggression. Contributed: 5329 posts.

CAD 2021 (Vidgen et al. 2021): Introducing a taxonomy comprising six conceptually distinct categories, Vidgen et al. undertook a sampling of posts from various subreddits anticipated to contain abusive content. The texts were meticulously labelled across these categories (directed abuse, counter-speech, etc.). Contributed: 23 060 posts.

Hate Speech A 2016 (Waseem and Hovy 2016): Waseem and Hovy built this dataset by manually searching on Twitter for common slurs, terms, and phrases related to religious, sexual, gender, and ethnic minorities. They categorized the tweets into three groups: racism, sexism, and none. Contributed: 16 849 posts.

Hate Speech B 2016 (Waseem 2016): Waseem's subsequent works continued to centre around hate speech detection, with a particular emphasis on refining the annotation aspect of this task. In this endeavour, they meticulously annotated almost 7000 tweets across four distinct categories—expanding by one category compared to their previous work. The new categories included racism, sexism, both, and none. Contributed: 6909 posts.

#MeTooMA 2020 (Gautam et al. 2020): This paper focuses on a specific type of hate speech, about the topic #MeToo. They narrowed their search where the #MeToo movement and identified keywords and phrases to create a 75-keyword lexicon and then queried tweets from these countries using this lexicon. The entries were labelled on hate speech, sarcasm, allegation, justification, refutation, support and opposition. Contributed: 9889 posts.

Hate Speech Data 2017 (Mondal, Silva, and Benevenuto 2017): In their study, Mondal et al. incorporated data from both Whisper and Twitter, curating a dataset comprising 28 309 entries—6157 from Whisper. Distinguishing themselves from conventional approaches, their method involved devising diverse hate speech sentence structures aimed at isolating purely hate speech comments. These structures were then enriched with a hate lexicon, enabling the identification of hate based on various categories (race, class, gender, ethnicity, disability, etc). Another study by Silva et al. (2021) analyzed this dataset to explore the targets of online hate speech. Contributed: 6157 posts from Whisper (we were unable to access the Twitter data).

Several other studies on hate speech detection in social media have shifted their focus to low-resource languages and languages beyond English. Noteworthy efforts have been observed in languages like Spanish (Fersini, Nozza, and Rosso 2018; Basile et al. 2019; Pereira-Kohatsu et al. 2019), Portuguese (Fortuna et al. 2019), Italian (Sanguinetti et al. 2018, 2020), French and Arabic (Ousidhoum et al. 2019), Turkish (Toraman, Şahinuç, and Yilmaz 2022), Slovene (Ljubešić, Fišer, and Erjavec 2019), and many more. While our work is centred on constructing a meta-collection in English due to space constraints, we acknowledge the importance of exploring other languages. Future works from our team will delve into addressing this gap.

Furthermore, we were able to find different hate speech datasets that were not linked to any publication but were uploaded to different platforms like Kaggle. Table 3 shows the summary of these datasets.

Meta-Collection Overview

MetaHate represents a comprehensive compilation of recent advancements in hate speech datasets, with a dual focus: (1) detecting hate speech, toxic behaviour, cyberbullying, aggression, and related terminologies, all falling under the umbrella of the defined harmful online content, and (2) analyzing text authored by humans across various social media platforms. Our meticulous review encompassed over 60 datasets dedicated to hate speech detection, ultimately incorporating 36 datasets into our meta-collection. The compilation resulted in 1 667 496 entries, streamlined to 1 226 202 non-duplicated comments. In shaping the scope of our dataset, we prioritized social media comments from platforms like Twitter or Facebook, while excluding synthetic data and sources outside the realm of social media, such as news comments or video game chats.

In an effort to streamline the dataset for broader applicability, we adopted a binary classification: hate or no hate. The primary rationale for this choice is that approximately half of the base datasets use a binary classification approach, while the remaining half adopt a multiclass approach. It is relatively straightforward to convert from a multiclass classification type to a binary one. However, the reverse process would be time-consuming and demanding. We are committed to fostering collaborative research by making our meta-collection available[2] to the research community.

Dataset Analysis

As previously mentioned, MetaHate encompasses 1 226 202 comments gathered from various social media networks: Twitter, Facebook, Reddit, Stormfront, Gab, Whisper, Wikipedia, Civil Comments, YouTube, and BitChute. The strategies employed for the datasets creation include: (1) utilizing lexicons, keywords, hashtags, and phrase structures,

[2]https://irlab.org/metahate.html

and (2) randomly sampling from sites with a likelihood of containing hate content. In terms of conceptualization, the majority of works adopt a binary strategy. However, a vast of them take a multiclass approach, distinguishing between abusive, hate, offensive, or normal speech, among other terms. A few efforts explore a probabilistic approach, assigning a numerical value between 0 and 1 to gauge the degree of hatefulness in a comment. Additionally, some studies opt for a multiclass and multilabel approach and others go a step further, attempting to extract the specific span that contains hate speech within a larger sentence.

Among these entries, 253 145 instances are labelled as hate, while the remaining 973 057 are identified as non-hateful. This signifies that 20.64% of our dataset entries are categorized as hate comments. This observation aligns with the consensus of some existing works, some of them emphasize that while hate speech is a genuine concern online, the overwhelming majority of posts fall within the non-hateful category (Waseem and Hovy 2016). The average post length is 261.80 characters and 43.39 words, with the longest post containing 2832 words. While this aligns with the character limit of 280 in a tweet, our dataset includes posts from platforms like Reddit, where the maximum length can reach 40 000 characters.

Lexical Analysis: First, we conducted a simple term frequency analysis. We observed that within the 20 most frequent keywords in our posts, terms such as *f*ck*, *user*, *people*, *n*gger*, *f*ggot*, *b*tch*, *hate*, *article*, and *woman* emerged. While certain words like *f*ggot* and *b*tch* could be associated with hate speech, others like *article* seem more neutral. Terms such as *people*, *f*ck*, and *women* could be present in both hate and non-hate posts. Additionally, the term *user* is common, as some of our dataset sources replaced user mentions with the word `user` (Zampieri et al. 2019; Ousidhoum et al. 2019; Mathew et al. 2020). Additionally, in Figure 1, we produced a Named Entity Recognition (NER) analysis. When comparing hate and non-hate posts, it's notable that hate comments often include a substantial percentage of references to **individuals** (almost 25%), followed by more than a 14% mentioning **nationalities, and religious or political groups** (NORP). This suggests a discernible pattern where attacks are frequently targeted at individuals and extensively involve references to nationality, religion, or political affiliation. In contrast, non-hate posts contain fewer references to **persons**, although they still make up more than 21%. References to **nationalities or religious or political** groups hover around 8%, and we observe that non-hateful posts include more references to **dates**, **companies, agencies, institutions** and **geopolitical entities** (i.e. countries, cities, states).

As a third experiment, we explore the realm of hate and non-hateful data employing topic modelling techniques. The focal point of our investigation was the optimization of hyperparameters, namely (1) the number of topics, (2) the alpha parameter, and (3) the beta parameter, within the LDA topic model. We aimed to maximize topic coherence, ensuring that the identified topics encapsulate meaningful and coherent thematic elements. As a result, we uncovered eight prominent topics within each category (hate and non-hate),

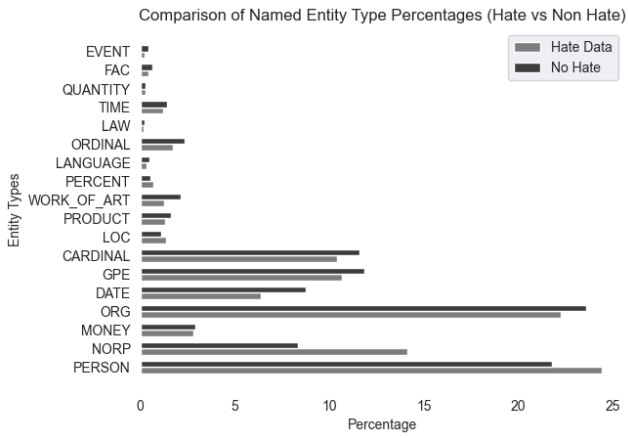

Figure 1: Comparison of Named Entity Type percentages between hate and non-hate posts.

each offering valuable insights into distinct thematic elements present in the dataset. For the purpose of this paper, we focus on presenting six out of the eight identified topics.

We examined the different topics through (1) word clouds, showcasing the top 10 words in each topic, providing a qualitative glimpse into the most frequent terms associated with each theme; (2) word count and importance metrics for these top words, offering quantitative insights, allowing us to discern the significance of each term within its respective topic; (3) distribution of document word counts by topic, providing a comprehensive overview of the length and complexity of discussions within each identified theme. Moreover, we employed T-distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton 2008) clustering, where we visually represented the relationships and proximity of documents within each of the six hate topics and the six non-hate topics. This technique allowed us to explore the underlying patterns and similarities between documents, providing a more nuanced understanding of the structural dynamics within the hate data.

In Figure 2, we present the word clouds showcasing the top 10 words for each of the six topics, distinguishing between hateful posts and non-hateful comments. Notably, the word clouds reveal distinctive lexical patterns characterizing each category. Hate topics prominently feature words such as *hate*, *f*ggot* and *bastard*, indicating explicit negativity and derogatory language. In contrast, the vocabulary associated with non-hateful posts includes terms like *article*, *good*, *think*, *country*, and *image*, reflecting a more positive and constructive linguistic orientation. This divergence in vocabulary underscores the semantic contrast between hate and non-hate content, revealing the distinct linguistic markers that contribute to the characterization of each category.

Furthermore, it is evident that certain topics within the analyzed content are associated with various forms of hate speech. In Topic 0, the presence of words such as *black*, *white*, and *people* is indicative of content aligned with racism themes. Topic 2 features terms like *kids d*ck* and *suck*, suggesting elements of child abuse. Similar associa-

tions can be observed in Topics 3 and 4: the former includes expressions like *kill*, *die* and *hole*, pointing to suicide, while the latter hints at fatphobia through the inclusion of words such as *fat* and *b*tch*. Topic 5 top words include *f*ggot*, *gay* and *loser*, inciting hatred of homophobia.

Figure 2: Word clouds of hate topics (up) and non-hate topics (down).

t-SNE is a dimensionality reduction algorithm used to visualize high-dimensional data. The diagram in Figure 3 (left) effectively illustrates the clustering of hate speech from social media LDA topics. Each distinct colour indicates a unique cluster of hate speech, the closer two topics are in the visualization, the more similar they are semantically. We use the topic weights to ensure that the most important topics are given more weight in the visualization. The cluster on the right (topic 1), being the largest, represents the most prevalent hate speech type on social media (including words like *rape*, *motherf*cker* and *ignorant*). The cluster on the left (topic 0) denotes the second most common hate speech, followed by the cluster on the top (topic 2), and so forth. The uneven distribution of the clusters suggests a hierarchical structure of hate speech, with some types being more common than others. While the clusters are somehow separated, they are not randomly distributed, indicating some degree of overlap between them. The t-SNE visualization effectively highlights the semantic relationships between different topics. Topics 1 (right) and 3 (top centre), for instance, share the common term *motherf*cker* and are positioned adjacent to each other, suggesting a close thematic connection. Similarly, topics 0 (left) and 4 (bottom centre), which include the derogatory terms *wh*re* and *b*tch* respectively, exhibit a discernible link in the visualized space.

On the other hand, for non-hateful texts, the t-SNE diagram can be seen in Figure 3 (right). The graph shows that the different types of non-hateful content are relatively well-

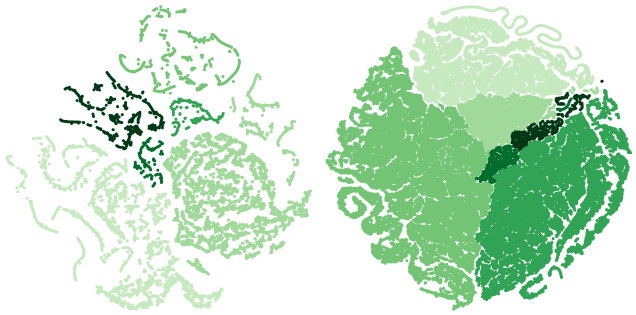

Figure 3: t-SNE diagram illustrating the clustering of hate speech (left) and non-hate speech (right).

separated, but not as well-separated as the clusters of hate topics. This could be because hate speech is more diverse than non-hate speech. The cluster on the left (topic 2) is the largest cluster, which suggests that it represents the most common type of non-hateful content on social media (including words such as *really*, *think*, *people* and *know*). As the different clusters are not completely separate, this suggests that there is some overlap between the different types of non-hateful content. For example, topics 2 and 3 (left and right), are adjacent as topic 2 terms are quite general and can be related to communication or decision-making, and topic 3 terms are associated with topics related to reading or writing and could be linked to discussions about articles, argumentation or any form of written communication.

Psycholinguistic Analysis: We employed the Plutchik set of emotions (Plutchik 1980), encompassing eight primary emotions (anger, fear, sadness, disgust, surprise, anticipation, trust, and joy) and two basic sentiments (positive and negative). To quantify emotion levels, we utilized the NRC emotion lexicon (Mohammad and Turney 2013), which comprises words linked to the Plutchik emotions. In our analysis, we computed the percentage of posts within each group (hate, no hate) that included at least one word associated with the primary emotions and sentiments. When comparing the results exposed in Figure 4, it is evident that **negative** emotions are strongly correlated with hate speech posts, while **positive** emotions are more prevalent in non-hateful content. Additionally, **fear**, **disgust**, and **sadness** exhibit a higher frequency in hate messages than in regular ones. Conversely, emotions like **trust** are associated with non-hate posts. The predominant emotion in our collection leans towards the negative spectrum, possibly attributed to the construction of hate datasets using keywords and lexicons containing negative terms. Furthermore, there is a significant contrast in the expression of emotions like **disgust** and **anger** between hate and non-hate publications.

Additionally, as we can see in Figure 5 (left), we observe a higher prevalence of second-person pronouns in hate speech posts (36.23%), while first-person singular and third-person singular pronouns are more common in non-hate posts (33.87% and 28.31%, respectively). Moreover, as reflected in Figure 5 (right), hate speech posts predominantly use the present tense (70.01%), with 17.50% in the past

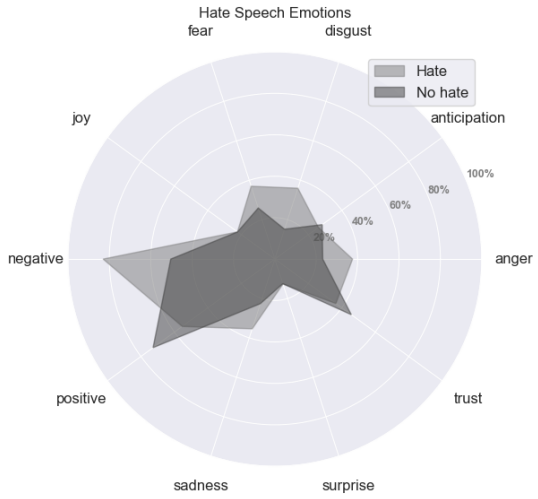

Figure 4: Radar plot showing the percentage of posts that contain a word associated with the Plutchik emotions for hate and non-hate data.

tense. In contrast, non-hate posts favour the present tense (64.67%), with 20.49% in the past tense. This implies that, although present tenses are predominant in both types of posts, hate speech comments exhibit a higher proportion of present tenses.

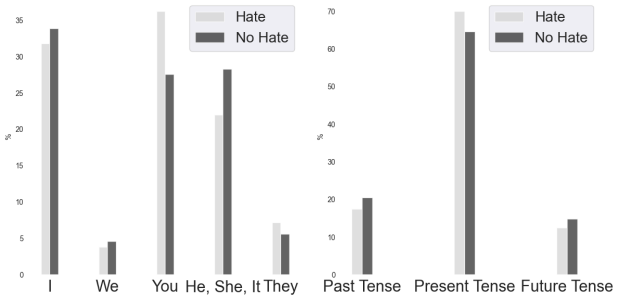

Figure 5: Distribution of pronouns in hate and non-hate posts (left). Hate speech tends to mention `YOU` more, possibly indicating attacks. On the other hand, non-hate speech more commonly includes references to `I` and `HE`, `SHE`, `IT`. And distribution of verb tenses in hate and non-hate speech posts (right). Note that `PRESENT` tenses are more prevalent in hate posts, suggesting possible attacks.

Baselines

In this section, we present our efforts to establish a common baseline for hate speech detection on social media. Our objective is to introduce basic text classification methods that can serve as foundational benchmarks for upcoming hate speech classification experiments. To achieve this, we've opted for conventional models like the Support Vector Machine (SVM), utilizing TF-IDF features for text features. Additionally, inspired by recent works (Mathew et al. 2020; Samory et al. 2021; Kennedy et al. 2022), we explored the inclusion of CNNs and BERT in our experimental framework.

We employed MetaHate to train the SVM and CNN models and to fine-tune the BERT model. In all our experiments, the dataset was divided into training and test sets, with a test size of 20%.

Support Vector Machine (SVM): We employed a linear SVM model with a feature limit set at 1M, excluding stop words, and implementing TF-IDF unigrams.

Convolutional Neural Network (CNN): We established a neural network and conducted training for one epoch on our tokenized sentences, ensuring padding to a maximum length of 512 tokens. The network architecture consists of an embedding layer, flattening, a dense layer with ReLU activation, and a final dense layer with a sigmoid activation for binary classification. The model was compiled using binary cross-entropy loss and the Adam optimizer.

BERT: We used the BERT base uncased model for text classification by loading the pre-trained implementation from the HuggingFace library. We did not do any additional hyperparameter tuning, but we used a learning rate of $5e^{-5}$, a sequence length of 512 during 3 epochs and a batch size of 32.

All the implementation details can be found in our GitHub repository[3].

Method	ACC	F1	$F1_{MICRO}$	$F1_{MACRO}$
SVM	0.85	0.84	0.85	0.74
CNN	0.86	0.84	0.86	0.74
BERT	**0.89**	**0.88**	**0.89**	**0.80**

Table 1: Hate speech detection results run on MetaHate.

The results, as shown in Table 1, reveal notable performance variations among the models. The traditional SVM and CNN models demonstrated competitive accuracies of 0.85 and 0.86, respectively, with corresponding F1 scores of 0.84. In contrast, the **BERT** model outperformed both, achieving an accuracy of **0.89** and higher F1 scores of 0.88, 0.89 (micro), and 0.80 (macro). This suggests that BERT, as a pre-trained language model, exhibits superior discriminatory power in capturing intricate patterns within hate speech data, leading to enhanced classification accuracy and nuanced performance across micro and macro F1 metrics. This outcome serves as a robust benchmark, showcasing the model's effectiveness in discerning hate speech within English-language posts on social media platforms using the MetaHate dataset.

To gain a clearer insight into these outcomes, Figure 6 offers a visual representation illustrating the distribution of accurate and inaccurate predictions generated by our best classification model BERT and SVM. CNN confusion matrix is similar to the SVM one, but due to space limits, we only included the visual representation of SVM. It is important to

[3]https://github.com/palomapiot/metahate/

note that for this classification task, false negatives are more important than false positives. This prioritization stems from the urgency to promptly identify instances where individuals are exposed to harmful content, with the associated benefits outweighing the potential drawbacks.

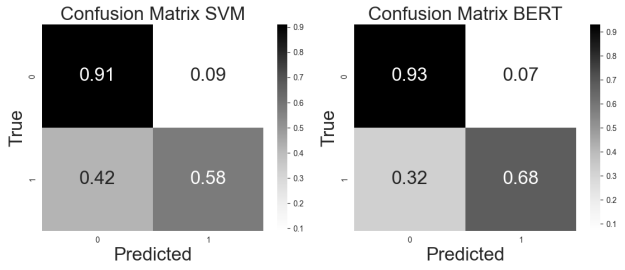

Figure 6: Confusion matrix showing the prediction accuracy of the different models on hate speech detection.

Discussion and Conclusions

Online social networks serve as a valuable repository of digital communication, occasionally becoming a breeding ground for hate speech discourse. It is imperative to address this issue with different models, that rely on large amounts of data to successfully perform. In this work, we presented **MetaHate**[4]: a meta-collection encompassing 36 hate speech datasets, representing the first large benchmark corpus on hate speech detection on social media. Additionally, we established a robust baseline, offering future researchers a clear reference for comparison. While we have conformed to the hate speech definitions put forth by the United Nations and previous works, it is important to note that, currently, there is no universally accepted definition of hate speech. Nevertheless, our objective is to move in the direction of establishing such a consensus. Moreover, in enhancing hate speech detection systems, the importance of context and multilingualism has become evident. Current methods often focus on individual English messages, lacking a broader and inclusive view. Future research should aim to capture complete conversation contexts, including different languages, for more accurate and nuanced detection of hate speech online. This work presents another limitation as we adopted a binary classification for hate speech data.

Ethical Statement

Addressing online hate speech involves navigating ethical considerations, particularly in the context of free speech controversies. Our data collection methods encompassed accessing publicly available datasets, obtaining information through submitted forms, and contacting authors via email. Importantly, the collected data strictly lacked any personally identifiable information. The significance of our work extends to potential societal benefits, offering a meticulously curated meta-collection of hate speech data that can enhance the identification of harmful comments on social media platforms. In light of the existence of offensive elements within the meta-collection, it is essential to exercise caution to prevent any potential misapplication for negative purposes, including the propagation of animosity or the targeting of individuals or communities.

Despite our contributions, it is crucial to acknowledge certain limitations. While some datasets are publicly accessible without specific terms and conditions, others necessitate explicit permission from the authors, often requiring individual dataset contracts for MetaHate availability.

We emphasize that any unintentional biases in the dataset are not intended to cause harm to individuals or target communities; rather, they may be inherent in the original publications. To promote further research in hate speech detection, we share our data and code. However, it is imperative to note that our dataset is released exclusively for research purposes and is not licensed for commercial or malicious use.

[4]Code available here: https://github.com/palomapiot/metahate/

Acknowledgments

The authors thank the funding from the Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them. The authors also thank the financial support supplied by the Consellería de Cultura, Educación, Formación Profesional e Universidades (accreditation 2019-2022 ED431G/01, ED431B 2022/33) and the European Regional Development Fund, which acknowledges the CITIC Research Center in ICT as a Research Center of the Galician University System and the project PID2022-137061OB-C21 (MCIN/AEI/10.13039/501100011033, Ministerio de Ciencia e Innovación supported by ERDF "A way of making Europe"). The authors also thank the funding of project PLEC2021-007662 (MCIN/AEI/10.13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next Generation EU).

Resources

Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.432 kgCO_2eq/kWh. A cumulative of 98 hours of computation was performed on hardware of type RTX A6000 (TDP of 300W). Total emissions are estimated to be 12.7 kgCO_2eq. Estimations were conducted using the MachineLearning Impact calculator presented in (Lacoste et al. 2019).

References

Albanyan, A.; and Blanco, E. 2022. Pinpointing Fine-Grained Relationships between Hateful Tweets and Replies. *Proceedings of the AAAI 2022*, 36(10): 10418–10426.

Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel Pardo, F. M.; Rosso, P.; and Sanguinetti, M. 2019.

SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 54–63. ACL.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. *Proceedings of the ICWSM 2020*, 14: 830–839.

Bhattacharya, S.; Singh, S.; Kumar, R.; Bansal, A.; Bhagat, A.; Dawer, Y.; Lahiri, B.; and Ojha, A. K. 2020. Developing a Multilingual Annotated Corpus of Misogyny and Aggression. In *Proceedings of the TRAC 2020*, 158–168. ELRA.

Bin Abdur Rakib, T.; and Soon, L.-K. 2018. *Using the Reddit Corpus for Cyberbully Detection*, 180–189. Springer. ISBN 9783319754178.

Das, M.; Raj, R.; Saha, P.; Mathew, B.; Gupta, M.; and Mukherjee, A. 2023. HateMM: A Multi-Modal Dataset for Hate Video Classification. *Proceedings of the ICWSM 2023*, 17: 1014–1023.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the ICWSM 2017*, 11(1): 512–515.

de Gibert, O.; Perez, N.; García-Pablos, A.; and Cuadros, M. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the ALW2 2018*. ACL.

ElSherief, M.; Kulkarni, V.; Nguyen, D.; Yang Wang, W.; and Belding, E. 2018a. Hate Lingo: A Target-Based Linguistic Analysis of Hate Speech in Social Media. *Proceedings of the ICWSM 2018*, 12(1).

ElSherief, M.; Nilizadeh, S.; Nguyen, D.; Vigna, G.; and Belding, E. 2018b. Peer to Peer Hate: Hate Speech Instigators and Their Targets. *Proceedings of the ICWSM 2018*, 12(1).

Fersini, E.; Nozza, D.; and Rosso, P. 2018. *Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI)*, 59–66. Accademia University Press. ISBN 9788831978699.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Fortuna, P.; Rocha da Silva, J.; Soler-Company, J.; Wanner, L.; and Nunes, S. 2019. A Hierarchically-Labeled Portuguese Hate Speech Dataset. In *Proceedings of the ALW 2019*, 94–104. ACL.

Founta, A.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; ...; and Kourtellis, N. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *Proceedings of the ICWSM 2018*, 12(1).

Gautam, A.; Mathur, P.; Gosangi, R.; Mahata, D.; Sawhney, R.; and Shah, R. R. 2020. MeTooMA: Multi-Aspect Annotations of Tweets Related to the MeToo Movement. *Proceedings of the ICWSM 2020*, 14: 209–216.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Golbeck, J.; Ashktorab, Z.; Banjo, R. O.; Berlinger, A.; Bhagwan, S.; Buntain, C.; ...; and Wu, D. M. 2017. A Large Labeled Corpus for Online Harassment Research. In *Proceedings of the ACM WebSci 2017*. ACM.

Grimminger, L.; and Klinger, R. 2021. Hate Towards the Political Opponent: A Twitter Corpus Study of the 2020 US Elections on the Basis of Offensive Speech and Stance Detection. In *Proceedings of the WASSA 2021*, 171–180. ACL.

Gupta, S.; Priyadarshi, P.; and Gupta, M. 2023. Hateful Comment Detection and Hate Target Type Prediction for Video Comments. In *Proceedings of the CIKM 2023*, CIKM '23, 3923–3927. ACM. ISBN 9798400701245.

Hickey, D.; Schmitz, M.; Fessler, D.; Smaldino, P. E.; Muric, G.; and Burghardt, K. 2023. Auditing Elon Musk's Impact on Hate Speech and Bots. In *Proceedings of the ICWSM 2023*, 1133–1137. AAAI.

Jha, A.; and Mamidi, R. 2017. When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP + CSS*, 7–16. ACL.

Kansok-Dusche, J.; Ballaschk, C.; Krause, N.; Zeißig, A.; Seemann-Herz, L.; Wachs, S.; and Bilz, L. 2023. A systematic review on hate speech among children and adolescents: definitions, prevalence, and overlap with related phenomena. *Trauma, violence, & abuse*, 24(4): 2598–2615.

Kennedy, B.; Atari, M.; Davani, A. M.; Yeh, L.; Omrani, A.; Kim, Y.; ...; and Dehghani, M. 2022. Introducing the Gab Hate Corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1): 79–108.

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

Kumar, R.; Reganti, A. N.; Bhatia, A.; and Maheshwari, T. 2018. Aggression-annotated Corpus of Hindi-English Code-mixed Data. In *Proceedings of the LREC 2018*. Miyazaki, Japan: ELRA.

Kurrek, J.; Saleem, H. M.; and Ruths, D. 2020. Towards a Comprehensive Taxonomy and Large-Scale Annotated Corpus for Online Slur Usage. In *Proceedings of the WOAH 2020*, 138–149. Online: ACL.

Lacoste, A.; Luccioni, A.; Schmidt, V.; and Dandres, T. 2019. Quantifying the Carbon Emissions of Machine Learning. *arXiv preprint arXiv:1910.09700*.

Ljubešić, N.; Fišer, D.; and Erjavec, T. 2019. The FRENK Datasets of Socially Unacceptable Discourse in Slovene and English. In *Text, Speech, and Dialogue*, 103–114. Springer. ISBN 978-3-030-27947-9.

Mandl, T.; Modha, S.; Kumar M, A.; and Chakravarthi, B. R. 2020. Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In *Proceedings of the FIRE 2020*, FIRE 2020. ACM.

Mandl, T.; Modha, S.; Majumder, P.; Patel, D.; Dave, M.; Mandlia, C.; and Patel, A. 2019. Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In *Proceedings of the FIRE 2019*, FIRE '19. ACM.

Mathew, B.; Dutt, R.; Goyal, P.; and Mukherjee, A. 2019. Spread of Hate Speech in Online Social Media. In *Proceedings of the ACM WebSci 2019*, WebSci '19. ACM.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2020. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *Proceedings of the AAAI 2020*.

Modha, S.; Mandl, T.; Shahi, G. K.; Madhu, H.; Satapara, S.; Ranasinghe, T.; and Zampieri, M. 2021. Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages and Conversational Hate Speech. In *Proceedings of the FIRE 2021*, FIRE 2021. ACM.

Mody, D.; Huang, Y.; and de Oliveira, T. E. A. 2023. A curated dataset for hate speech detection on social media text. *Data in Brief*, 46: 108832.

Mohammad, S. M.; and Turney, P. D. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3): 436–465.

Mollas, I.; Chrysopoulou, Z.; Karlos, S.; and Tsoumakas, G. 2022. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6): 4663–4678.

Mondal, M.; Silva, L. A.; and Benevenuto, F. 2017. A Measurement Study of Hate Speech in Social Media. In *Proceedings of the ACM HT*, HT '17, 85–94. ACM. ISBN 9781450347082.

Nations, U. 2023. What is hate speech? Accessed: 15/11/2023.

Ousidhoum, N.; Lin, Z.; Zhang, H.; Song, Y.; and Yeung, D.-Y. 2019. Multilingual and Multi-Aspect Hate Speech Analysis. In *Proceedings of the EMNLP-IJCNLP 2019*, 4675–4684. ACL.

Pavlopoulos, J.; Sorensen, J.; Dixon, L.; Thain, N.; and Androutsopoulos, I. 2020. Toxicity Detection: Does Context Really Matter?

Pavlopoulos, J.; Sorensen, J.; Laugier, L.; and Androutsopoulos, I. 2021. SemEval-2021 Task 5: Toxic Spans Detection. In *Proceedings of the SemEval 2021*, 59–69. ACL.

Pereira-Kohatsu, J. C.; Quijano-Sánchez, L.; Liberatore, F.; and Camacho-Collados, M. 2019. Detecting and Monitoring Hate Speech in Twitter. *Sensors*, 19(21).

Plutchik, R. 1980. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1: 3–31.

Poletto, F.; Basile, V.; Sanguinetti, M.; Bosco, C.; and Patti, V. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2): 477–523.

Qian, J.; Bethke, A.; Liu, Y.; Belding, E.; and Wang, W. Y. 2019. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. In *Proceedings of the EMNLP-IJCNLP 2019*, 4755–4764. ACL.

Röttger, P.; Vidgen, B.; Nguyen, D.; Waseem, Z.; Margetts, H.; and Pierrehumbert, J. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the ACL-IJCNLP*, 41–58. ACL.

Sachdeva, P.; Barreto, R.; Bacon, G.; Sahn, A.; von Vacano, C.; and Kennedy, C. 2022. The Measuring Hate Speech Corpus: Leveraging Rasch Measurement Theory for Data Perspectivism. In *Proceedings of the LREC 2022*, 83–94. ELRA.

Salminen, J.; Almerekhi, H.; Milenković, M.; Jung, S.-g.; An, J.; Kwak, H.; and Jansen, B. 2018. Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. *Proceedings of the ICWSM 2018*, 12(1).

Samory, M.; Sen, I.; Kohne, J.; Flöck, F.; and Wagner, C. 2021. "Call me sexist, but..." : Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. *Proceedings of the ICWSM 2021*, 15: 573–584.

Sanguinetti, M.; Comandini, G.; di Nuovo, E.; Frenda, S.; Stranisci, M.; Bosco, C.; Caselli, T.; Patti, V.; and Russo, I. 2020. *HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task*, 93–101. Accademia University Press.

Sanguinetti, M.; Poletto, F.; Bosco, C.; Patti, V.; and Stranisci, M. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the LREC 2018*. ELRA.

Silva, L.; Mondal, M.; Correa, D.; Benevenuto, F.; and Weber, I. 2021. Analyzing the Targets of Hate in Online Social Media. *Proceedings of the ICWSM 2021*, 10(1): 687–690.

Spertus, E. 1997. Smokey: Automatic Recognition of Hostile Messages. In *AAAI/IAAI*.

Toraman, C.; Şahinuç, F.; and Yilmaz, E. 2022. Large-Scale Hate Speech Detection with Cross-Domain Transfer. In *Proceedings of the LREC 2022*, 2215–2225. ELRA.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

Vidgen, B.; and Derczynski, L. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12): e0243300.

Vidgen, B.; Nguyen, D.; Margetts, H.; Rossini, P.; and Tromble, R. 2021. Introducing CAD: the Contextual Abuse Dataset. In *Proceedings of the NAACL 2021*, 2289–2303. ACL.

Vogels, E. A. 2021. The state of online harassment.

Waseem, Z. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the NLP + CSS 2016*, 138–142. ACL.

Waseem, Z.; and Hovy, D. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL 2016*. ACL.

Wulczyn, E.; Thain, N.; and Dixon, L. 2016. Ex Machina: Personal Attacks Seen at Scale.

Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; and Kumar, R. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the NAACL 2019*, 1415–1420. ACL.

AAAI ICWSM Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, in section *Data Acquisition and Preparation*.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, because we opted for non-synthetic data, the heterogeneity in data distribution and origin is apparent, stemming from diverse sampling across various social media platforms. It is important to note, however, that the inherent limitations of each artifact have been explicitly outlined in their respective original publications.
   (e) Did you describe the limitations of your work? Yes, in section *Discussion and Conclusions*.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, in section *Ethical Statement*.
   (g) Did you discuss any potential misuse of your work? Yes, in section *Ethical Statement*.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, in section *Ethical Statement*.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, in subsection *Baselines*.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, in subsection *Baselines*.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, because we run simple experiments in order to set clear baselines.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, in the GitHub repository (https://github.com/palomapiot/metahate)
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, in subsection *Baselines*.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, in subsection *Baselines*.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, in section *Data Acquisition and Preparation*.
   (b) Did you mention the license of the assets? Yes, in the *Appendix* each dataset is referenced with its availability. Each specific license is referenced in its original publication.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, in section *Meta-Collection Overview*.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, in the *Appendix*, table 2, *Available* column.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, in section *Data Acquisition and Preparation* and, as a hate speech dataset, offensive content is present.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes, because we are sharing our dataset via Hugginface. Where the dataset is assigned a DOI (doi:10.57967/hf/1572), is described with rich metadata, is indexed, is retrievable by their identifier, is acessible, is interoperable and is reusable.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes, it is linked with our

sourcecode (https://github.com/palomapiot/metahate/blob/main/DATASHEET.md).

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA

Appendix

Dataset	Actual Size	Source	Classification Type	Labels	Baseline	Creation Strategy	Available
Online Harassment 2017	19,838	Twitter	Binary	hate, no hate	-	Lexicon, word structures, hashtags	Via email
AMI 2018	*3,251*	*Twitter*	*Binary*	*misogyny, no misogyny*	*SVM*	*Keywords, account monitoring*	*Protected with password*
Cyberbullying Personality	*3,987*	*Twitter*	*Binary*	*cyberbullying, no cyberbullying*	*Random Forest*	*Hashtag*	*No*
Hate Lingo 2018	*148,256*	*Twitter*	*Binary*	*directed hate, generalized hate*	*-*	*Lexicon, hashtags, dataset sampling, random sampling*	*Only Tweet IDs*
Hateval 2019	12,747	Twitter	Binary	hate, no hate	SVM	Lexicon, hashtags, user monitoring	Yes
OLID 2019	14,052	Twitter	Binary	offensive, not offensive	CNN	Keywords, phrase structures	Yes
SWAD 2020	*2,569*	*Twitter*	*Binary*	*hate, no hate*	*LSVC*	*OLID subset using keywords*	*Yes*
US 2020 Elections	2,999	Twitter	Binary	hate, no hate	BERT	Keywords, hashtags	Yes
"Call me sexist but" 2021	3,058	Twitter	Binary	sexism, no sexism	BERT	Sexism scales, phrase "call me sexism, but", datasets sampling	Via registration
HASOC 2019-2021	6,981	Twitter, Facebook	Binary	hate, no hate	LSTM	Heuristics, topics, keywords and hashtags	Only 2019
A Curated Hate Speech Dataset 2023	560,385	Twitter, Facebook, Wikipedia, etc.	Binary	hate, no hate	-	Not explained	Yes
Hate Speech B 2016	6,909	Twitter	Multiclass	racism, sexism, both, none	n-grams	(Waseem and Hovy 2016) sample	Via email
Hate Speech A 2016	16,849	Twitter	Multiclass	racism, sexism, none	Logistic Regression	Lexicon	Via email
Mean Birds 2017	*9,484*	*Twitter*	*Multiclass*	*normal, spammer, aggressor, bully*	*Random Forest*	*Hashtags*	*Only Tweet IDs*
Ambivalent Sexism 2017	*22,142*	*Twitter*	*Multiclass*	*hostile, benevolent, others*	*FastText*	*Keywords, hashtags, phrase structures*	*Only Tweet IDs*
Hate Offensive 2017	24,783	Twitter	Multiclass	hate speech, offensive, neither	Logistic Regression	Lexicon	Yes
TRAC1 2018	14,537	Twitter, Facebook	Multiclass	very aggressive, a bit aggressive, not aggressive	-	Hashtags, controversial topics	Via form

Continued on next page

Continued from previous page

Dataset	Actual Size	Source	Classification Type	Labels	Baseline	Creation Strategy	Available
Harassment Corpus 2018	*25,000*	*Twitter*	*Multiclass*	*sexual, racial, appearance-related, intellectual, political*	*-*	*Lexicon*	*No*
ENCASE 2018	91,950	Twitter	Multiclass	abusive, normal, spam, hateful	-	Random sample, lexicon	Yes
MLMA 2019	5,593	Twitter	Multiclass, multilabel	abusive, hateful, offensive, disrespectful, fearful, normal	biLSTM	Keywords	Yes
#MeTooMA 2020	9,889	Twitter	Multiclass	directed hate, generalized hate, sarcasm, allegation, justification, refutation, support, oppose	-	Lexicon, filtering by country	Via email
HateXplain 2020	20,109	Twitter, Gab	Multiclass	hate, offensive, normal	BERT	Lexicon	Yes
Hate Speech Data 2017	6,157	*Twitter*, Whisper	Multiclass	race, behaviour, physical, sexual orientation, class, gender, ethnicity, disability, religion, other	-	Sentence structures	Only Whisper
Hateful Tweets 2022	1,141	Twitter	Multiclass	hate, justification, attacks author, additional hate	NA	Existing datasets extension	Via email
Multiclass Hate Speech 2022	68,597	Twitter	Multiclass	hate, offensive, normal	Megatron (BERT)	Keywords, hashtags	Via email
Measuring Hate Speech 2020/2022	39,565	Twitter, Reddit, YouTube	Probability	higher is more hate, lower less	RoBERTa	Random Sample	Yes
BullyDetect 2018	6,562	Reddit	Binary	cyberbullying, no cyberbullying	Random Forest	Not explained	Yes
Intervene Hate 2019	45,170	Reddit, Gab	Binary	hate, no hate	CNN/RNN	Toxic subreddits, keywords	Yes
Slur Corpus 2020	39,960	Reddit	Multiclass	derogatory, non derogatory, homonym, appropriation, noise	-	(Baumgartner et al. 2020) keywords filtering	Yes
CAD 2021	23,060	Reddit	Multiclass, multilabel	identity directed abuse, affiliation directed abuse, person directed abuse, counter speech, neutral	BERT	Community-based sampling	Yes
ETHOS 2022	998	Reddit, YouTube	Probability	1 hate, 0 no hate	DistilBERT	Subreddits, hatebusters.org	Yes
Hate in Online News Media 2018	3,214	Facebook, YouTube	Binary	hate, neutral	SVM	Filtering by an online news and media company	Yes

Continued on next page

Continued from previous page

Dataset	Actual Size	Source	Classification Type	Labels	Baseline	Creation Strategy	Available
Supremacist 2018	10,534	Stormfront	Binary	hate, no hate	LSTM	Random sample	Yes
The Gab Hate Corpus 2022	27,434	Gab	Binary	assault on human dignity, not assault on human dignity	BERT	Random sample	Yes
HateComments 2023	2,070	YouTube, BitChute	Binary	hate, no hate	BERT	Manual selection from list of categories	Yes
TRAC2 2020	5,329	YouTube	Multiclass	very aggressive, a bit aggressive, not aggressive	-	Selected topics	Via form
Toxic Spans 2021	10,621	Civil Comments	Spans	Span positions	BERT	Civil Comments labelled as toxic	Yes
Dynamic Hate 2021	*41,135*	*Synthetic*	*Binary*	*hate, no hate*	*RoBERTa*	*Data generation*	*Yes*
CONAN 2019-2022	*8,883*	*Semi-Synthetic*	*Multiclass*	*disabled, jews, LGBT+, migrants, muslims, people of color, women*	*-*	*LMs*	*Yes*
Ex Machina 2016	115,705	Wikipedia	Binary	attack, no attack	MLP	Random sample, blocked users	Yes
Context Toxicity 2020	19,842	Wikipedia	Binary	toxic, no toxic	BERT	Not explained	Yes
MetaHate	**1,226,202**	**Social media**	**Binary**	**hate, no hate**	**BERT**	**Meta collection**	**Yes**

Table 2: Hate Speech English datasets. Datasets in italic and gray colour are not included in MetaHate.

Author	Platform	Link	Lang.	Actual Size	Source	Classification Type	Labels	Available
Roshan Sharma / Ali Toosi	Hugging Face	https://t.ly/v9tin	EN	29,530	Twitter	Binary	hate, no hate	Yes
Munki Albright	Kaggle	https://t.ly/ZO_Cx	EN	18,208	Twitter	Multiclass	suspicious, cyberbullying, hate, suicidal	Yes
SR	Kaggle	https://t.ly/XpBNt	EN	159,571	Twitter	Multiclass	malignant, highly malignant, rude, threat, abuse, loathe	Yes
Jigsaw	Kaggle	https://t.ly/v6IkS	EN	223,549	Wikipedia	Multiclass, multilabel	toxic, severe toxic, obscene, threat, insult, identity hate	Yes

Table 3: Hate Speech Datasets from different platforms.