Exploring automatic hate speech detection on social media: a focus on content-based analysis

NASCIMENTO, Francimaria R. S., CAVALCANTI, George D. C. and DA COSTA-ABREU, Márjory <http://orcid.org/0000-0001-7461-7570>

Published version

NASCIMENTO, Francimaria R. S., CAVALCANTI, George D. C. and DA COSTA-ABREU, Márjory (2023). Exploring automatic hate speech detection on social media: a focus on content-based analysis. SAGE Open, 13 (2).

Original Research

SAGE Open
April-June 2023: 1–19
© The Author(s) 2023
DOI: 10.1177/21582440231181311
journals.sagepub.com/home/sgo
Sage

Exploring Automatic Hate Speech Detection on Social Media: A Focus on Content-Based Analysis

Francimaria R. S. Nascimento[1], George D. C. Cavalcanti[1], and Márjory Da Costa-Abreu[2]

Abstract

Hate speech is a challenging problem, and its dissemination can cause potential harm to individuals and society by creating a sense of general unwelcoming to the marginalized groups, which usually are targeted. Therefore, it is essential to understand this issue and which techniques are useful for automatic detection. This paper presents a survey on automatic hate speech detection on social media, providing a structured overview of theoretical aspects and practical resources. Thus, we review different definitions of the term "hate speech" from social network platforms and the scientific community. We also present an overview of the methodologies used for hate speech detection, and we describe the main approaches currently explored in this context, including popular features, datasets, and algorithms. Furthermore, we discuss some challenges and opportunities for better solving this issue.

Keywords

hate speech detection, social media, survey, metadata, text features, natural language processing

Introduction

Social media platforms allow users to publish content about different subjects quickly and easily. Easy content dissemination and anonymity on social media platforms can increase the published harmful content. Different information types can intentionally or unintentionally harm (Giachanou & Rosso, 2020), including misinformation, disinformation, and mal-information. *Misinformation* (Aswani et al., 2019; Kar & Aswani, 2021), often defined as satirical, is incorrect or fictional information created and spread, disregarding the proper intention. *Disinformation* (Nasir et al., 2021), for example, fake news is deliberately created to mislead the target users. *Mal-information* (Davidson et al., 2017; Giachanou & Rosso, 2020), for example, hate speech is created to incite or cause harm. In this survey, we particularly investigate the hate speech detection task.

Hate speech is a challenging problem that demonstrates a clear intention to incite harm or promote hatred against others. This issue is considered a worldwide problem faced by many countries and organizations. With the growth of online social media, millions of users can spread much information every second, and the problem has become quite significant. There is a general understanding that when a person feels physically safe, the person's speech tends to be more aggressive (Watanabe et al., 2018). Moreover, there is a real movement from hate groups to recruit people to create and diffuse hate speech messages (Del Vigna et al., 2017).

The easy spread of hate speech on online platforms is a serious concern for our society, considering that the dissemination of hate speech can cause potential harm to individual victims and society, for example, raising hostility between groups (Miškolci et al., 2020; Teh et al., 2018). Particularly, repetitive exposure to hate speech can lead to desensitization to this form of violence, thus lowering the victims' evaluations and increasing the bias against the target groups (Mathew et al., 2019).

[1]Universidade Federal de Pernambuco (UFPE), Recife, Brazil
[2]Sheffield Hallam University, UK

Corresponding Author:
Márjory Da Costa-Abreu, Department of Computing, Sheffield Hallam University, Cantor, Sheffield S1 1WB, UK.
Email: md0948@exchange.shu.ac.uk

Social media platforms, such as Facebook, Twitter, and YouTube, have claimed they have intended to solve this problem, which they present in policies on hate behavior and attempts to combat hate speech (Facebook, 2020; Twitter, 2020; YouTube, 2020). Much of this content moderation currently requires manual review of questionable documents (Waseem & Hovy, 2016). However, the speed with which such messages are transmitted (shared) makes manual control over message content labor-intensive, time-consuming, expensive, and not scalable (Cao et al., 2020; Zhang et al., 2018).

Furthermore, the hate speech detection task suffers from several weaknesses related to specific nuances of this subject and the complexity of this classification task (Poletto et al., 2021). A relevant issue consists of clearly defining hate speech to understand the problem better and avoid strong subjective interpretations. As we will present in this survey, several disciplines have different definitions for the term "hate speech," which are complementary.

All the listed issues and limitations of the manual approaches have motivated considerable research. This survey also aims to provide an overview of better aspects of the problem, such as its definition, different features used in this problem, datasets, and methods. Furthermore, we highlight challenges and draw future work directions, obtaining a theoretical starting ground for new scientists on the topic.

Understanding the better aspects of hate speech detection is relevant to dealing with this issue. As a general basis for this area, we found some surveys proposed in this field exploring different questions. In Schmidt and Wiegand (2017) and Fortuna and Nunes (2019), the researchers also survey critical tasks employed for hate speech detection. Nevertheless, it is relevant to note that this field has received increasing attention from the scientific community, and different resources included in the present survey had not been released when these surveys were published or at least when the researchers performed the search. Other works have focused on survey-specific characteristics of hate speech detection, such as multilingual corpus (Al-Hassan & Al-Dossari, 2019), annotated corpora (Poletto et al., 2021), and hate speech on the social media platform Twitter (Ayo et al., 2020).

This contribution aims to complement these works and present a critical analysis of theoretical aspects and practical resources since this field has constantly grown. (i) We overview a general methodology for hate speech detection on social media, focusing on textual data. (ii) Besides, we present a comprehensive overview of recent resources from different social media and languages, such as the datasets, features used, and algorithms. (iii) We describe the advantages and limitations of several feature extraction techniques currently used in the literature. (iv) We point out different open challenges and opportunities in this field.

This paper is organized as follows: We first present an analysis of different definitions for the term "hate speech" based on several sources; Then, we explain the methodology used to select the works for this review; Next, we discuss a general methodology for hate speech detection; Then an overview of the related datasets; After, we summarize several feature extraction approaches and present the advantages and limitations of the features explored; Then, we discuss several classification methods used in the literature; Furthermore, we present different challenges highlighted in the literature and opportunities in this field; finally, we conclude this survey with the final remarks.

What is Hate Speech?

Hate speech is a complex phenomenon, and detecting whether a text contains hate speech is not a trivial task, even for humans. Therefore, a precise definition of hate speech is crucial to automatically distinguish hate speech from other content (Ross et al., 2016). We have seen an increasing number of studies that have addressed hate speech detection with different definitions of the term. There is probably because of the fog limits between hate speech and appropriate freedom of expression (MacAvaney et al., 2019).

Thus, we have decided to analyze different sources' definitions, considering the wide range of origins. We have analyzed the description of hate speech presented by social media in their "terms and conditions" contracts (Twitter, Facebook, YouTube) because hate speech often occurs on those platforms and some related studies, to include the perspective of the scientific community. Since Cohen-Almagor (2013) proposed one popular definition in the communication literature, Fortuna and Nunes (2019) analyzed several sources and considered distinct aspects, and Davidson et al. (2017) annotated a dataset used in several works. Thus, we will be considering those three aspects in our work.

1. Facebook: "We define hate speech as a direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We define attack as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, or calls for exclusion or segregation." Facebook (2020)
2. Twitter: "Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity,

national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories." Twitter (2020)

3. YouTube: "Hate speech is not allowed on YouTube. We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin, veteran status."YouTube (2020)
4. Cohen-Almagor: "Hate speech is defined as a bias-motivated, hostile, malicious speech aimed at a person or a group of people because of some of their actual or perceived innate characteristics."Cohen-Almagor (2013)
5. Fortuna and Nunes: "Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used."Fortuna and Nunes (2019)
6. Davidson et al: "Language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group." Davidson et al. (2017)

In some aspects, these definitions can be considered similar. A common theme is that hate speech is used against a specific targeted group or group members. Besides, it has been seen by different sources as an attack or incitement to violence. Davidson et al. (2017) define as a language that intended to be abusive, derogatory, humiliating, or insulting. While Cohen-Almagor (2013) considers hostile and malicious speech based on innate characteristics. In general, these definitions have complementary nuances to each other. In particular, Fortuna and Nunes (2019) specifically considers that hate speech can occur even in subtle forms. The authors argue that subtle forms of discrimination can use humor to reinforce stereotypes and racial discrimination, causing adverse effects for some people.

Considering these definitions, we can point out four main characteristics of hate speech described: (1) promotes attack or incites violence; (2) used against a specific target group or members of the group based on any characteristics such as gender, race, sexual orientation, religion, ethnicity or other aspects; (3) may or may not use "abusive language" and derogatory terms; (4) can occur in subtle forms, for example, subtle metaphors "*expecting gender equality is the same as genocide*," this example of hate tweet does not contain explicit hateful lexical (Zhang & Luo, 2019).

Research Methodology

We have surveyed to understand hate speech detection on social media better, focusing on textual data. Our goal is to investigate the most recent studies developed in this field. To limit this research's scope, we have decided to restrict our search to documents published starting in 2015. The reason for this decision is the fact that in Fortuna and Nunes (2019), it was shown that before 2014 this theme received little attention in computer science and engineering research, which is highlighted by the fact that many resources had not been released when previous surveys were published (Poletto et al., 2021).

We searched the documents in different sources, such as ACM digital library, IEEE, Elsevier, and Springer. The keywords selected were "hate speech detection," "hate speech classification," besides also considered the search for "Abusive language," considering that abusive language is a sub-category of hate speech. The keywords selected were searched in the publication title, abstract and keywords. We also used Google Scholar to search for references that cited the original work. We check on these sets and search for the keyword "hate speech detection" on the titles of the documents. Several entries appeared as results of more than one search string.

We have focused on the field of computer science and engineering research. Also, we only included papers with at least four pages and peer-reviewed scientific resources. Furthermore, we restricted the works as automatic hate speech detection to the only ones performed on social media platforms, particularly from textual data. The text published on these platforms have specific characteristics (e.g., a limited number of characters, URLs, emojis, mentions, and so on). Thus, we have selected a total of 83 papers in the search period. Figure 1 presents the distribution of papers over the selected time interval.

It is quite clear the scientific community's recent efforts toward dealing with automatic hate speech detection relate to the processing and analysis of textual data. The following sections present several automatic hate speech detection techniques that explore this aspect.

Automatic Hate Speech Detection

The automatic hate speech detection process includes tasks such as data collection and processing, feature extraction, detection, and classification. We analyze and

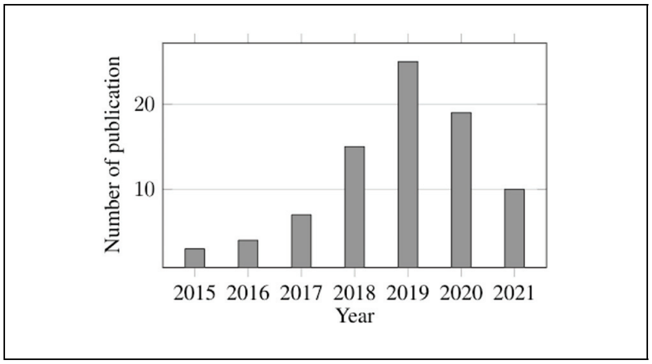

Figure 1. Number of publication toward the years for hate speech detection from January, 1st 2015 to July, 31st 2021.

summarize the main tasks typically employed in automatic hate speech detection on social media platforms. Figure 2 presents an overview of the architecture for hate speech detection.

Social media platforms provide a wide variety of information that can be collected using the programing libraries known as Application Programming Interface (API). The researchers have adopted different strategies to crawl data related to hate speech, such as derogatory words, common slurs, hashtags, specific profiles, following "trigger events, and so on (Burnap & Williams, 2015; Davidson et al., 2017; Fortuna et al., 2019; Founta et al., 2018; Waseem & Hovy, 2016). Moreover, several works have used pre-filtering to exclude spam, samples with no content, and samples not in English (Founta et al., 2018; Pratiwi et al., 2018). According to Founta et al. (2018), abusive tweets are relatively rare, and the percentage can range between 0.1% and 3% of the samples collected.

The methodology employed to collect and annotate the dataset should be carefully chosen to avoid bias in the dataset (Wiegand et al., 2019). The annotation task in different studies used CrowdFlower (CF) workers (Chatzakou et al., 2017; Davidson et al., 2017; Founta et al., 2018; Kumar et al., 2019; Waseem, 2016), but this approach can be expensive. The authors Chatzakou et al. (2017), Founta et al. (2018) used a default payment scheme for batch (each with 10 tweets) to minimize costs without compromising the annotation quality. Moreover, the authors also performed the annotation task (Waseem & Hovy, 2016) or used non-experts and experts annotated (Basile et al., 2019; Fortuna et al., 2019; Waseem, 2016). Another approach employed is active learning annotation (Charitidis et al., 2020) for further annotation and dataset expansion. Several authors (Alsafari et al., 2020; Golbeck et al., 2017; Mossie & Wang, 2020; Waseem & Hovy, 2016) developed a coding guideline to help human annotators classify the content due to the subjectivity of the human interpretation of hate speech. The following section presents a further overview of hate speech detection datasets.

In the context of social media platforms, the text used frequently has specific characteristics, such as abbreviations, incorrect spelling, slang, acronyms, URLs,

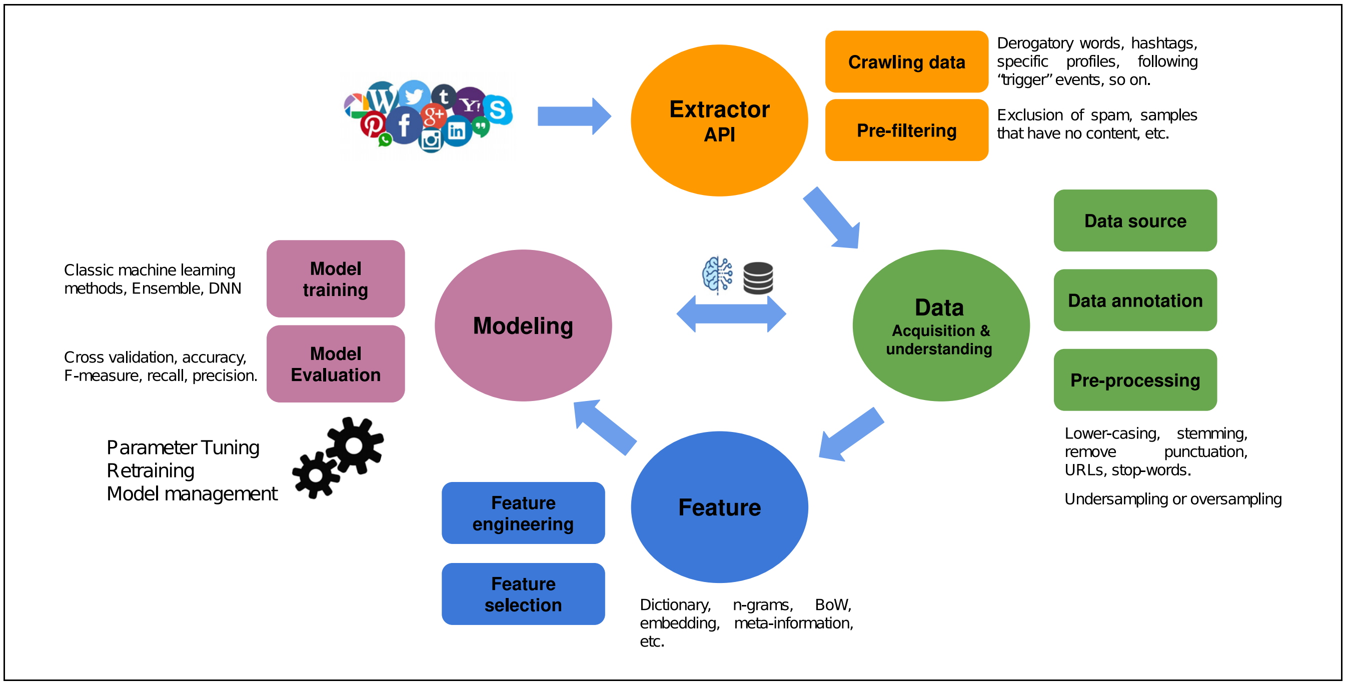
Extractor
API
Crawling data
Derogatory words, hashtags, specific profiles, following "trigger" events, so on.
Pre-filtering
Exclusion of spam, samples that have no content, etc.
Data source
Data annotation
Pre-processing
Data
Acquisition & understanding
Lower-casing, stemming, remove punctuation, URLs, stop-words.
Undersampling or oversampling
Feature
Feature engineering
Feature selection
Dictionary, n-grams, BoW, embedding, meta-information, etc.
Modeling
Model training
Classic machine learning methods, Ensemble, DNN
Model Evaluation
Cross validation, accuracy, F-measure, recall, precision.
Parameter Tuning
Retraining
Model management

Figure 2. Overview of architecture for hate speech detection on social media platforms.

hashtags, emojis, mentions, and so on. The unstructured text and, at times, the informal language can introduce noise in the classification task (Naseem et al., 2021). Several pre-processing methods are explored before the feature extraction task in order to reduce noise in the dataset, such as lower-casing of words, stemming, removing punctuation, URLs, stop-words, replacing emoticons and emojis, elongated characters (Dorris et al., 2020; Nugroho et al., 2019; Pratiwi et al., 2018; Sohn & Lee, 2019; Watanabe et al., 2018; Zhang et al., 2018). Naseem et al. (2021) evaluated 12 different pre-processing techniques and the combination of them in three datasets of hate speech (proposed in Davidson et al., 2017; Golbeck et al., 2017; Waseem & Hovy, 2016). The authors concluded that the lemmatisation and lower casing of words presented a high performance in most cases. On the other hand, removing punctuation and URLs, user mentions, and Hashtags symbols presented a low performance in most cases. Moreover, some studies focused on techniques to deal with the class imbalance problem, such as oversampling and undersampling. The oversampling technique is applied in the training data to increase the minority class (Chatzakou et al., 2017; Elisabeth et al., 2020), while the undersampling technique reduces the majority class (Miok et al., 2019). However, most of the works did not deal with class imbalance.

Feature extraction is an important task in text analysis. Several approaches are explored in hate speech detection and related subjects. Among these, dictionary or lexical resources (Burnap & Williams, 2015; Gitari et al., 2015; Mathew et al., 2019; Nobata et al., 2016; Teh et al., 2018), distance metric (Mossie & Wang, 2020; Nandhini & Sheeba, 2015), bag-of-word (Burnap & Williams, 2016; Senarath & Purohit, 2020; Waseem et al., 2018), *n*-grams (Corazza et al., 2020; Mossie & Wang, 2020; Santosh & Aravind, 2019; Senarath & Purohit, 2020; Wulczyn et al., 2017), term frequency (Almatarneh et al., 2019; Mossie & Wang, 2020; Salminen et al., 2020), text embedding and deep learning (Cao et al., 2020; Miok et al., 2019; Senarath & Purohit, 2020; Zimmerman et al., 2018), meta-information (Founta et al., 2019; Pitsilis et al., 2018; Waseem & Hovy, 2016), and so on. Different studies addressed hate speech detection on social media present better results when combining a set of features (Salminen et al., 2020; Senarath & Purohit, 2020). In this study, we highlight several methods used to feature extraction and their advantages and limitations.

Although the feature engineering process's effective for text representation, the feature space can present a high dimensionality. However, in the context of hate speech detection, few studies (Robinson et al., 2018; Zhang et al., 2018) have evaluated the feature selection process's impact. The automatic feature selection algorithms can reduce the original feature space by 90% and improve machine learning algorithms' performance for hate speech detection (Robinson et al., 2018; Zhang et al., 2018).

Classic supervised machine learning methods have been explored for automated hate speech detection. Among these, Support Vector Machines (SVM) (Burnap & Williams, 2015; Salminen et al., 2020), Logistic Regression (LR) (Davidson et al., 2017; Khan et al., 2021; Waseem & Hovy, 2016), Naive Bayes (NB) (Ibrohim & Budi, 2019; Salminen et al., 2020), Random Forest (RF) (Almatarneh et al., 2019), C4.5 decision tree learning (Watanabe et al., 2018). Although more expensive, ensemble approaches have presented robust results of the different classification task (Burnap & Williams, 2015; Markov et al., 2021; Nugroho et al., 2019; Paschalides et al., 2020; Zimmerman et al., 2018). Another approach explored is the DNN, which has been used for feature extraction and classifiers' training. The most used approaches are CNN, LSTM, and GRU (Al-Makhadmeh & Tolba, 2020; Alsafari et al., 2020; Cao et al., 2020; Dorris et al., 2020; Marpaung et al., 2021; Mossie & Wang, 2020; Pitsilis et al., 2018; Rizos et al., 2019; Santosh & Aravind, 2019; Zhang & Luo, 2019). This work discusses several methods used for hate speech detection on social media platforms in the following section.

The following sections present an overview of the datasets, feature extraction techniques, and classification methods employed for automatic hate speech detection.

Datasets for Hate Speech Classification

Representative publicly available datasets are essential for developing automatic hate speech detection approaches. However, collecting and annotating data in the context of hateful messages is challenging, especially, as previously mentioned, no universal definition is adopted. The most common way of labeling this type of content is using social media platforms' definitions. Besides, the number of hate speech texts compared to non-hate on social media platforms is significantly smaller. The studies adopted some strategies to collect the dataset, such as using terms and phrases related to hate content from dictionaries like HateBase, specific profiles, hashtags and keywords (Davidson et al., 2017; Fortuna et al., 2019; Founta et al., 2018; Waseem & Hovy, 2016).

Table 1 summarizes the main information from several datasets proposed in the literature. These datasets vary considerably in their labels, number of instances, characteristics of hate speech, etc. The most popular data source is Twitter, which has attracted a significant part of the research due to the increasingly available data and

Table 1. Summary of Datasets for Hate Speech Classification.

Dataset	Year	Distribution	Number of instances	Labels (%)	Annotators	Origin source	Language
WH Waseem and Hovy (2016)	2016	GitHub repository	16,914	sexism (20%), racism (12%) and none (68%)	authors	Twitter	English
WS Waseem (2016)	2016	GitHub repository	6,909	sexism (13%), racism (11%), both (1%), neither(84%)	3 or more	Twitter	English
DV Davidson et al. (2017)	2017	GitHub repository	24,802	hate (5%) offensive(76%) neither(17%)	3 or more	Twitter	English
GB Golbeck et al. (2017)	2017	Need request access	35k	Harassing (15.7%) Non-Harassing (74.3%)	2–3	Twitter	English
FT Founta et al. (2018)	2018	Dataverse	80k	hateful(7.5%), abusive(11%), spam(22.5%), normal (59%)	5–20	Twitter	English
PR Pratiwi et al. (2018)	2018	GitHub repository	835	Hate speech 34.24% not hate speech 65. 75%	3	Instagram	Indonesian
SE Basile et al. (2019)	2019	GitHub repository	19,600 (6,600 -Spanish; 13,000 -English)	Hate (43%)/Not Hate (57%)	3	Twitter	English, Spanish
FO Fortuna et al. (2019)	2019	GitHub repository	5,668	hate speech (22%), not hate speech(78%)	3	Twitter	Portuguese
IB Ibrohim and Budi (2019)	2019	GitHub repository	13,169	hate speech (42.2%) not hate speech (57%)	3	Twitter	Indonesian
YT Philipp and Roman (2019)	2019	Zenodo platform	1k	hate speech (13.8%) not hate speech (86.2%)	—	YouTube	English
KU Kumar et al. (2019)	2019	Need request access	18k tweets (T) 21k facebook (F)	Overtly Aggressive (T −6.0% F 27.5%), Covertly Aggressive (T-44.1% F 29.9%), Non-aggressive (T-49.9% F 42.6%)	3	Facebook and Twitter	Code-mixed (Hindi-English)
AL Alsafari et al. (2020)	2020	GitHub repository	5,360	Hate 26.65% Offensive 8.18% Clean 65,17%	3	Twitter	Arabic
CH Charitidis et al. (2020)	2020	Zenodo platform	EN 92,022 DE 43,735 ES 37,688 FR 29,109 GR 61,481	Hate speech EN p 7.78% n 92.22% DE p 3.9% n 96.1% ES p 2.64% n 97.36% FR p 9.3% n 90.7% GR p 1.86% n 98.14%	1	Twitter	English, German, Spanish, French and Greek

free APIs (Davidson et al., 2017; Waseem & Hovy, 2016; Watanabe et al., 2018). English has been the most popular language analyzed, but we can also find works exploring other languages, such as Arabic, Spanish, Indonesian, Portuguese, German, French, and Greek.

Overall, the publicly available datasets for hate speech detection in different languages and social media platforms are scarce, with few studies publishing their datasets. In most cases, the datasets are not available for external researchers, such as a large annotated dataset of abusive language detection from the "Yahoo! Finance and News" (Nobata et al., 2016); Facebook, Italian language corpus of hate speech (Del Vigna et al., 2017), Amharic language corpus for hate speech detection approach to vulnerable community identification (Mossie & Wang, 2020). Poletto et al. (2021) performed a further analysis in several datasets for hate speech detection, including methodology, topical focus, language, and other factors. The results presented different data sources and highlighted some issues and improvements.

Feature Extraction Approaches

An essential task in text analysis is the meaningful feature extraction from data. The approaches selected often have a significant impact on the data analysis itself. However, extracting insights and patterns from a text can be challenging, especially in the context of social media, where there is the issue of unstructured text. Table 2 presents the advantages and limitations of the most widespread techniques for feature extraction used in the context of hate speech detection and related subjects. In this section, we analyze features used in hate speech detection and related subjects.

Dictionaries or Lexical Resources

Dictionary is a relevant approach used in natural language processing (NLP) based on keywords. This strategy lists potential keywords and counts the number of occurrences in the text or context.

These frequencies can be used as features or to compute scores. For hate speech context, different dictionaries have been available:

- Hatebase is a multilingual dataset of derogatory terms with data across 95 + languages and 175 + countries. This resource offers constants updates in the terminology and a broad vocabulary (https://hatebase.org/);
- Dictionary of general swear words and insults in English (https://www.noswearing.com/);
- Urban dictionary of colloquial language and slang words in English (https://www.urbandictionary.com).

Previous works used this approach, in general, considering negative or derogatory words (Burnap & Williams, 2015; Gitari et al., 2015; Mathew et al., 2019; Nobata et al., 2016; Teh et al., 2018). Gitari et al. (2015) built a lexicon of hate-related verbs which encourage violent acts (such as to discriminate, loot, riot, beat, kill, and evict). Mathew et al. (2019) created a lexicon with 45 hate words selected from the Hatebase and Urban dictionary for further analysis of hateful and non-hateful users on Gab. Teh et al. (2018) constructed a lexical of profane words frequently used in different types of hate speech from comments on YouTube which showed that 35% of profane words are related to sexual orientation, based on 500 comments. Burnap and Williams (2016) focused on specialized lists toward particular subtypes of hate, such as LGBT slang terms, ethnic slurs, and negative connotation against disabled people. Hayaty et al. (2020) focused on local languages in Indonesia for hate speech detection and created a dictionary of abusive words containing of 250 terms.

Despite their general effectiveness, a limitation of this approach is the dependency oon hateful keywords (MacAvaney et al., 2019). Thus, lexical features can be employed as an additional step of feature extraction (Schmidt & Wiegand, 2017).

Distance Metric

The presence of noise and conjugations often makes it difficult to perform automatic detection of hateful content. Once derogatory words are intentionally used in text messages (Nobata et al., 2016), it is possible to identify such words with characters substitution such as "*ni99er*," "*@ss*," "*sh1t*" which can make the whole process even more challenging for automatic detection. Approaches to compute the minimum number of edit operations of individual characters like Levenshtein distance can also be used for this end (Nandhini & Sheeba, 2015). There is no lexicon for hate speech detection in some languages, such as the Amharic language. Thus, one approach employed was translating the text into English using the Google translator tool. In this approach, the researchers used the cosine distance to evaluate the semantic similarity between each input word

Table 2. Overview of the Features Used in the Context of Hate Speech Detection. Where *n* is the number of different words/tokens/string in the document.

Method	Advantages	Limitations	Average vector size
Dictionaries or lexical resources	It is a simple method and effective to detect hate speech with derogatory terms.	The dependency of hateful keywords	$\sim 5 - 250$ words
Distance Metric	It captures the number of edit operations and semantic similarity.	It is few explored in the context of hate speech, and it is used as a complementary metric.	*n* strings
Bag-of-words (BoW)	The corpus is collected from the training data.	It ignores word sequences and its semantic and syntactic content, may lead misclassification to words used in various contexts.	*n* different words in the sentence
N-grams	Overcome the limitation of BoW. The subclass POS captures information about the syntactic structure of the text.	It can suffer from a high level of distance between related words. Besides, the POS technique can promote confusion between the classes due to the abundance of similar patterns.	items sequences (with *n* in range between $1 - 5$)
Term frequency	It provides good classification performance for hate speech detection, simple method.	It did not help the model generalize well across different dataset domains.	*n* tokens
Template Based Strategy	Structures predefined.	It can generate false positives, besides it is useful often to the specific context.	Template length
Typed Dependencies	It extracts a subset of dependency relationship labels.	It is often used as a complementary metric and can increase the number of false-negative instances.	Number of sentences extracted
Text embedding and Deep learning approaches	The pre-trained word embeddings have proved useful for abusive text classification, besides it required fewer training samples to obtain a good performance. The DNN technique learns abstract feature representations for hate speech detection; It can be used for feature extraction as well as a machine learning classifier.	A problem faced with pre-trained word embedding is out-of-vocabulary (OOV) words. Moreover, a limitation of DNN techniques is the high cost computational and explainability.	$25 - 300$ dimensions
Sentiment analysis	Usually, negative sentiment belongs to the hate speech message, besides several automatic tools available.	It needs using other techniques to improve results.	Number of sentiment polarity (usually “positive,” “negative,” “neutral,” and “compound”)
Meta-information	It provides additional information about the context of the message.	It is scarce and often not readily available for external researchers; it might introduce bias in the model.	Amount of metadata

and the corresponding vectors in the model (Mossie & Wang, 2020).

Bag-of-Words (BoW)

Bag-of-Words (BoW) is another technique used to detect hateful speech (Burnap & Williams, 2016; Nobata et al., 2016; Senarath & Purohit, 2020; Waseem et al., 2018). Similarly to the dictionary, this technique uses keywords, the main difference being that it creates a corpus from the collected training data, while the dictionary uses predefined words. After the data collection stage, word frequencies are used as a feature for training a classifier. A limitation of this approach is ignoring word sequence and its semantic and syntactic content. Hence, it may lead to the mistaken classification of words used in various contexts. Another technique that can be adopted to overcome this limitation is *n*-grams.

A statistical analysis conducted using BoW with all typed dependencies and with only hateful and derogatory terms to investigate its influence in the classification task is presented in Burnap and Williams (2015), which follows the assumption that BoW can confuse the classification task when the same word is frequently in non-hateful and hateful scenarios. The study showed that using only hateful and derogatory terms can potentially increase the number of false negatives because the hateful content does not necessarily uses derogatory or hateful terms.

N-grams

The *n*-grams is one of the most used techniques in automatic hate speech detection and related tasks (Chakraborty & Seddiqui, 2019; Corazza et al., 2020; Mossie & Wang, 2020; Santosh & Aravind, 2019; Senarath & Purohit, 2020; Wulczyn et al., 2017). It combines a sequence of *n* adjacent items into a list with size *N*, where the items can be words (most common), syllables, or characters (Fortuna & Nunes, 2019). However, for the problem of hate speech detection, "character *n*-grams" provided better performance than "word *n*-grams," because it captures the changes in the words associated with hate (Del Vigna et al., 2017; Unsvåg & Gambäck, 2018; Waseem & Hovy, 2016).

Its main disadvantage is that it suffers from a high level of distance between related words (Burnap & Williams, 2016), which is closely associated with the selection of the *n* value. Since *n*-grams may not be able to capture long-range dependencies between words, for example: "*Jews are lower class pigs*," the words "*Jews*" and "*pigs*," similarities would not be connected using only *n*-grams, depending on the *n* selected (Nobata et al., 2016).

These features often are combined with other features to improve the hate speech classification. For instance, in Watanabe et al. (2018), the authors explored different features for hate speech detection, such as the most common word unigrams, pattern features, sentimental, and semantic features. They believed that unigrams features could help identify explicit forms of hate speech. Overall, unigrams features presented high accuracy, but all features combined performed better.

Part-of-speech (POS) is a subclass of the *n*-gram approach that detects the role of the word in the context of the sentence, which tags capture the syntactic function of the word, for instance, personal pronoun (PRP), verbs (VB), nouns (NN), adjectives (JJ). These approaches have been used for hate speech detection to capture information about the syntactic structure of the text to extract frequencies from unigrams, bigrams, and trigrams (Davidson et al., 2017).

Furthermore, it was also used to collect unigrams with a specific syntactic function (e.g., noun, verb, adjective or adverb) from the training set to investigate occurrences in hateful and offensive tweets (Watanabe et al., 2018). However, POS, when used as a feature, can promote confusion between the classes due to the abundance of similar patterns (Burnap & Williams, 2015; Fortuna & Nunes, 2019).

Term Frequency

The word or term frequency indicates the relevance of the word in the document that contains it. The most common types of word frequency are Term Frequency (TF), Term Relative Frequency (TFR), Inverse Document Frequency (IDF), and Term Frequency-Inverse Document Frequency (TF-IDF) (Liu et al., 2019). In Plaza-Del-Arco et al. (2020) used TF weighting to represent unigrams and bigrams as vectors of numerical features to misogyny and xenophobia detected in Spanish tweets. Several works used TF-IDF weighting features for hate speech detection (Almatarneh et al., 2019; Elisabeth et al., 2020; Mossie & Wang, 2020; Salminen et al., 2020). The TF-IDF provided good classification performance for hate speech detection with the same dataset to train and test the models. However, it did not help the model generalize well when used across different dataset domains (Senarath & Purohit, 2020).

Typed Dependencies

Typed dependencies have been widely used for hate speech detection (Alorainy et al., 2019; Burnap & Williams, 2015, 2016). The probabilistic parse trees, provided by Stanford Typed Dependency Parser (De Marneffe & Manning, 2008, can be used to extract a subset of dependency relationship labels and provide a description of the grammatical relationships in a sentence (Alorainy et al., 2019). The introduction of typed dependency features for hate speech detection can reduce the false positive rate, but this can lead to an increase in false-negative instances. This approach performed better when combined with other features (Burnap & Williams, 2016).

Template Based Strategy

In this strategy, the main idea is to build a corpus of structured sentences. Mondal et al. (2017) proposed the follow sentence structure "*I < intensity> < userintent*

> < *hatetarget* >," to search hate speech post. Thus, they additionally designed two templates, focusing on exploring hate against groups of people. The first was simply "< one word> people" for scenarios when hate was directed toward a group, and the second template used words collected on Hatebase for < hate target> tokens.

Text Embedding and Deep Learning Approaches

The embedding technique is aimed at training a model to provide a vector representation of sentence/word, which captures the semantic and the syntactic relationship between the words (Indurthi et al., 2019). Word embedding methods have improved prediction accuracy for hate speech classification (Liu et al., 2019), which can be illustrated by several studies using pre-trained word embedding approaches, such as Word2vec, GloVe, FastText, ELMo, LASER, XLM, BETO (Cao et al., 2020; del Arco et al., 2021; Miok et al., 2019; Senarath & Purohit, 2020; Sreelakshmi et al., 2020; Vitiugin et al., 2021). The pre-trained word embedding had been proven effective for abusive text classification. Besides, it required fewer training samples to obtain a good performance (Founta et al., 2019). Another approach is sentence embedding which represents sentences as vectors. Miok et al. (2019) proposed a model for hate speech detection in three datasets (from Twitter and YouTube) using word and sentences embedding. The approach used the LSTM model with Monte Carlo dropout obtained better performance by using pre-trained sentence embedding than word embedding and state-of-the-art features.

However, an issue faced with pre-trained word embedding is out-of-vocabulary (OOV) words. Particularly, present on social media data because of its colloquial nature, users often perform intentional obfuscation of words which can be mitigated by performing pre-processing before feature extraction for noise reduction (Zhang & Luo, 2019). Corazza et al. (2020) investigated the impact of word embedding and emoji embedding on the specific domain and compared it with pre-trained embeddings, such as FastText. Specific embedding improved the results but needed a large amount of data. On the other hand, pre-trained embedding using binary models could mitigate the issue of OOV word, since this approach provided sub-words information.

Deep neural network (DNN) techniques have been recently explored to learn abstract feature representations for hate speech detection. The most popular approaches are the Convolutional Neural Network (CNN) and the Long Short-Term Memory network (LSTM). In the context of hate speech classification, CNN was applied as a feature extractor, and LSTM was used for modeling sequences of word or character dependencies (Bouazizi et al., 2021; Kapil & Ekbal, 2020; Sajjad et al., 2019; Santosh & Aravind, 2019; Zhang et al., 2018).

Even though very expensive, another approach explored was deep learning ensembles that used CNN for feature extraction (Zhou et al., 2020; Zimmerman et al., 2018). These techniques are robust and improve the results of the different classification tasks. In a study conducted in seven datasets from Twitter in the English language (Zhang & Luo, 2019), CNN showed more effectiveness for specific types of hate (racism and sexism) than polarized data (hate and non-hate).

Other approaches have investigated the language model pre-training BERT (Bidirectional Encoder Representation from Transformers) (Calabrese et al., 2021; Wich et al., 2021). BERT was designed to pre-train deep bidirectional representation. In Hendrawan et al. (2020), analyzed the BiLSTM and the BiLSTM with BERT multilingual trained with Wikipedia from 104 languages. However, the BiLSTM with BERT was less effective than the BiLSTM and the Random Forest Decision Tree.

Sentiment Analysis

Sentiment analysis is often considered synonymous with "opinion mining," a field of study that aims to analyze a person's feelings, opinions, and emotions toward "elements" (Serrano-Guerrero et al., 2015). The "elements" in this context can represent individuals, events, services, products, and topics. Sentiment analysis and hate speech are related, and often negative sentiments are associated with hate speech messages (Schmidt & Wiegand, 2017).

Several works have used the sentiment as a feature for hate speech detection (Cao et al., 2020; Corazza et al., 2020; Gitari et al., 2015; Rodriguez et al., 2019). Features based on emotions and sentiments are relevant approaches and can improve classification tasks on hate speech detection (Corazza et al., 2020; Markov et al., 2021). However, supervised methods required labels for sentiment classification and hate speech datasets often did not have this information. Different automatic tools were explored to overcome this limitation of supervised methods for sentiment analysis, such as JAMMIN, an emotion analysis tool, and VADER (Valence Aware Dictionary for sEntiment Reasoning), a sentiment analysis tool (Cao et al., 2020; Rodriguez et al., 2019).

Although related, it is arguable that hate speech detection is a different task requiring more sophisticated techniques (Watanabe et al., 2018). In sentiment analysis, the presence of positive/negative words or expressions can be considered helpful in this process. The presence of negative words or expressions, even in such sentences using the word "hate," depending on the context, does not make them related to hate speech. Thus, this approach for feature extraction is usually used with other techniques to improve results (Cao et al., 2020; Corazza et al., 2020; Watanabe et al., 2018).

Meta-information

Additional information from social media can help better understand the characteristics of the post-context and provide valuable data for hate speech detection. Social media platforms offer a wide variety of information that can be collected through APIs, such as user gender, demographics data, timestamp, user profiles, and network structures (Ayo et al., 2020; DeSouza & Da-Costa-Abreu, 2020).

Background information about the user can improve the predictably of hateful messages since hateful users are densely connected (Ribeiro et al., 2018). In a study about the impact of information like user gender and demographic information in tweets (Waseem & Hovy, 2016), these features brought slight improvement, but this could be because of the lack of coverage. Information about user gender was also explored in Unsvåg and Gambäck (2018), which used a similar approach performed in Waseem and Hovy (2016), to identify the user gender based on username or profile names as well as the user description in messages. However, a limitation of this approach is names used for both female and male. Another approach investigated the metadata based on text content to analyze specific attributes in tweets, such as the number of hashtags and mentions of other users, emoticons in the tweet, words with only uppercase letters, URLs included, and frequency of punctuation marks (Al-Makhadmeh & Tolba, 2020; Chatzakou et al., 2017; Del Vigna et al., 2017; Founta et al., 2019).

Furthermore, another meta-information relevant is the user network, such as user friends and followers. These features are beneficial in classifying aggressive user behavior (Chatzakou et al., 2017). Features about user behavioral are also useful for detecting racist and sexist messages (Pitsilis et al., 2018). These features can help describe the user's tendency toward the class based on their tweets history, post content, and subsets of those tweets with labeled messages. This information is scarce and often not readily available for external research (Cao et al., 2020; MacAvaney et al., 2019). Since these data have sensitive information about users and publishing raises privacy issues. Moreover, user information can introduce bias in the model against particular users or groups (MacAvaney et al., 2019).

Other Techniques

Other features used in the classification task are based on Flesch-Kincaid Grade Level (FKGL) and Flesch Reading Ease (FRE) scores to measure the quality of a document (Şahi et al., 2018); *Pattern features (*Watanabe et al., 2018); *Latent Dirichlet Allocation (LDA), typically used for topic modelling.*Cao et al. (2020)*used LDA to determine the posts' topic distribution in each dataset, considering each post as a single document.*

Texts extracted from social media platforms often contain URLs, punctuation, symbols, username, and tags such as "@," RT and < >. Some studies, before the feature extraction stage, have used stemming and removed special characters and stop-words (Zhang et al., 2018). However, using stemming, some words in the Indonesian language can be converted into words with different meanings, such as "*dadakan*" which means *all of sudden* to "*dada*" which means *chest*. Besides, stop-word removal can reduce the information from the sentence (Hendrawan et al., 2020).

Classification Methods

Automated hate speech detection on social media is a complex problem. Several approaches have been explored to deal with this problem, such as classic supervised machine learning methods, ensemble, and DNN techniques. Table 3 summarizes several studies with the results for the best model for each work.

Classic supervised machine learning methods have been explored for automated hate speech detection. Nobata et al. (2016) developed a machine learning based method to detect hate speech from the "Yahoo! Finance and News" dataset that outperformed a deep learning approach. The decision tree classifiers were also explored for hate speech detection and related subjects (Chatzakou et al., 2017; Watanabe et al., 2018). In the study Chatzakou et al. (2017), the Random Forest classifier presented a better performance in classifying bullying and aggressive behavior from a Twitter dataset than other tree classifiers experimented (J48, LADTree, LMT, NBTree, and Functional Tree), with 90% AUC (Area Under Curve). The authors in Watanabe et al. (2018) also analyzed datasets from Twitter. The data was collected and combined from three different datasets labeled as hateful, offensive, or clean. They selected the C4.5 decision tree to classify the data in two explored approaches binary and ternary. The binary classification (polarized the tweets as offensive and clean) obtained an accuracy of 87.4%, and the ternary classification

Table 3. Summary of Studies for Hate Speech Detection on Social Media.

Ref.	Year	Feature	Model	Social Media	Dataset	Acc	AUC	R	P	F
Nobata et al. (2016)	2016	Token and char N-grams (3-5), POS tags, word2vec, comment2vec, length of comment in tokens, number of puctuations, so on.	Vowpal Wabbit's regression model	Yahoo	collected	—	—	—	—	Fin. 0.79 News 0.81
Chatzakou et al. (2017)	2017	Meta-information, Word embedding, sentiment, dictionary	RF	Twitter	collected	—	0.90	0.91	—	0.89
Del Vigna et al. (2017)	2017	POS, sentiment, word2vec, char- lemma- and word- n-grams, repetition of n-grams char, punctuation	SVM, LSTM	Facebook	collected	0.80	—	~0.79	~0.83	~0.78
Wulczyn et al. (2017)	2017	n-gram (word, char)	LR, MLP	Wikipedia	collected	—	0.96	—	—	—
Pitsilis et al. (2018)	2018	User features	ensemble of LSTM	Twitter	WH	—	—	~0.87	~0.9	~0.89
Watanabe et al. (2018)	2018	sentiment, punctuation marks, all-capitalized words, POS, word unigram, pattern features	C4.5	Twitter	All combined (available on https://www.crowdflower.com/data-for-everyone/Crowdflower, DV, WH)	0.87	—	0.87	0.88	0.87
Zhang and Luo (2019)	2019	skipped CNN (sCNN), word2vec, CNN	CNN + GRU and CNN + sCNN	Twitter	WS (WS-S.amt, WS-S.exp, WS-S.gb, WS.pj), DV Zhang et al. (2018)	*—	—	—	—	0.83–0.94
Rizos et al. (2019)	2019	word2vec, GloVe, FastText, POS-tags	LSTM, CNN	Twitter	DV, WS, WH	—	—	DDV 0.49	—	DDV 0.74, WH 0.82, WS 0.83
Almatarneh et al. (2019)]	2019	n-grams, TF-IDF and CountVectorizer	SVM, GNB, CNB, DT, K-NN, RF, and NN	Twitter	SE	—	—	—	—	EN 0.76 ES 0.77
Liu et al. (2019)	2019	embedding; LDA	fuzzy ensemble	Twitter	Burnap and Williams (2015)	0.93	—	—	—	—
Santosh and Aravind (2019)	2019	char and word n-grams, negation words, punctuation marks	SVM, RF, Sub-word level LSTM, Hierarchical LSTM	Twitter	Bohra et al. (2018)	0.66	—	0.45	—	0.48
Al-Makhadmeh and Tolba (2020)	2019	sentiment, semantic,unigram and pattern features	ensemble deep learning	Twitter	collected	0.98	—	—	—	—
Senarath and Purohit (2020)	2020	BoW, tf-idf, n-grams, dictionary, FrameNet, word2vec	SVM	Twitter	DV, FT	DV 0.94, FT 0.94	DV 0.96, FT 0.78	DV 0.97 FT 0.90	—	DV 0.96, FT 0.70
Senarath and Purohit (2020)	2020	BoW, TF-IDF, GloVe, BERT, and all combined	LR, NB, SVM, XGBoost, and Neural Networks	YouTube, Reddit, Wikipedia, and Twitter	Salminen et al. (2018), Almerekhi et al. (2019), Wulczyn et al. (2017) DV	—	—	—	—	D1 0.91, D2 0.77, D3 0.86, DV 0.98
Cao et al. (2020)	2020	GloVe, word2vec, Paragram, sentiment and LDA.	LSTM, C-LSTM-Att	Twitter	(WH and WS combined), DV, FT, All combined	—	—	—	—	D1 0.78, DV 0.89, FT 0.79, D4 0.92
Alsafari et al. (2020)	2020	Unigram, word and char n-grams, word embedding	NB, SVM, LR, CNN, LSTM, GRU	Twitter	collected	—	—	0.87	0.86	0.87
Mossie and Wang (2020)	2020	word n-grams, TF-IDF and word2vec	GBT, RF, LSTM, GRU	Facebook	collected	0.92	0.97	—	—	—

(polarized the tweets as hateful, offensive, and clean) had an accuracy of 78.4%.

Del Vigna et al. (2017) analyzed the SVM classifier and a recurrent neural network LSTM on a dataset from Facebook in the Italian language. The classifiers presented a similar performance for hate speech detection. The LR and MLP are used in Wulczyn et al. (2017) both classifiers obtained 96% AUC. Several classifiers are explored in Almatarneh et al. (2019). In the study, the Complement Naive Bayes (CNB), SVM, and RF presented the best performances to identify specific hate speech against women and immigrants in English and Spanish languages. The SVM was also used in Senarath and Purohit (2020) to evaluate semantic features of social media messages for hate speech detection.

Salminen et al. (2020) analyzed hate speech as a problem of multiple social media platforms (YouTube, Reddit, Wikipedia, and Twitter). They investigated multiple algorithms and individual features as well as combined features. The ensemble algorithm XGBoost (Extreme Gradient Boosted Decision Trees) presented a more significant performance than the other algorithms analyzed (*F1* = 0.92). In the analysis of the features, the models show the best performance with BERT features.

Another approach explored is the Deep Neural Network (DNN), which has been used for feature extraction and classifier training. The most used classifiers are LSTM, CNN, and GRU (Al-Makhadmeh & Tolba, 2020; Alsafari et al., 2020; Cao et al., 2020; Mossie & Wang, 2020; Pitsilis et al., 2018; Rizos et al., 2019; Santosh & Aravind, 2019; Zhang & Luo, 2019). Cao et al. (2020) proposed a framework for hate speech detection on social media, namely DeepHate. They evaluated the DeepHate using three public datasets and the combination of the three datasets. The DeepHate outperformed different CNN models.

An ensemble of recurrent neural networks is also investigated for hate speech detection (Pitsilis et al., 2018). The authors proposed an ensemble of LSTM with the user's tendency toward each class as a feature method. Their model proposed has obtained more effective results than state-of-the-art with the detection of sexist messages (about F1-score = 0.99), neutral (about F1-score = 0.95), and racism (about F1-score = 0.70).

The ensemble deep learning method was also explored in Al-Makhadmeh and Tolba (2020). The authors proposed a hybrid approach, namely Killer Natural Language Processing Optimisation Ensemble Deep Learning (KNLPEDNN), which combines NLP and machine learning techniques. They used Stormfront (a neo-Nazi website) and CrowdFlower Twitter datasets. The ensemble method was used to minimize the weak features and to improve the prediction of hate. The system obtained 98.71% accuracy.

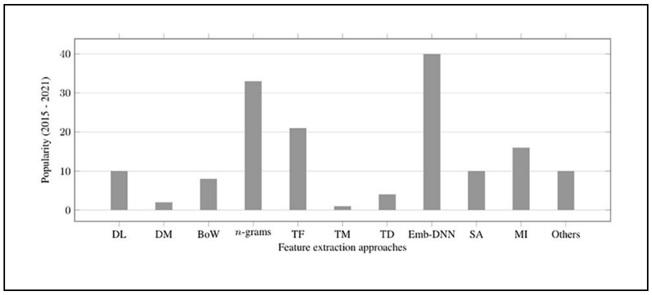

Figure 3. The frequency of feature extraction techniques from 2015 to July 2021.

Note. DL = Dictionary or Lexical; DM = Distance Metrics; BoW = Bag-of-Words; n-grams; TF = Term Frequency ; TM = Template Method; TD = Typed Dependencies; Emb-DNN = Text Embedding and DNN ; SA = Sentiment Analysis; MI = Meta-information.

The models used different metrics to evaluate the performance of the models, such as Accuracy (Acc), AUC, Precision (P), Recall (R), and F-measure (F). Accuracy measures the number of correctly predicted samples among all predicted samples. The AUC computes the area under the ROC Curve. Precision measures the percentage of true positives among the true and false positives predicted. Recall measures the percentage of true positive cases that are correctly predicted positive. The F-measure calculates the harmonic average of precision and recall. Despite the results obtained in the studies evaluated, it needs to be clarified which model performed better. Furthermore, several works evaluate only the dataset collected by itself without evaluating whether the model generalizes well to other domains.

Research Directions and Gaps for Hate Speech Detection on Social Media

This section aims at presenting challenges and points out automatic hate speech detection opportunities on social media platforms. As our previous sections suggested, the community has developed several resources to benefit from benchmark datasets for hate speech detection on social media platforms. Several feature extraction techniques and classification methods are employed on hate speech detection and related subjects. Figure 3 presents information about the popularity of the approaches used for feature extraction, and Figure 4 presents the classification method's popularity. The feature extraction techniques more used are embedding and DNN, and the *n*-grams. The classification method more used is SVM. Most works use more than one approach or a

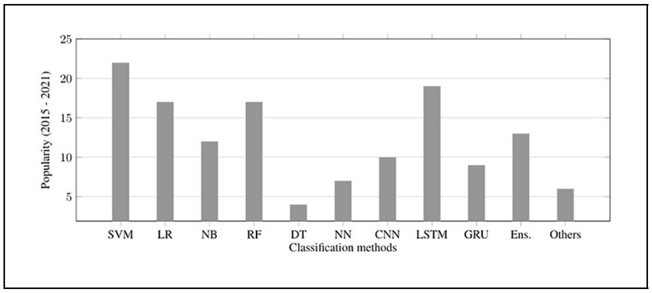

Figure 4. The frequency of classification methods from 2015 to July 2021.
Note. SVM = Support Vector Machines; LR = Logistic Regression; NB = Naive Bayes; RF = Random Forest ; DT = Decision Tree; NN = classical Neural Network; CNN = Convolutional neural network; LSTM = Long Short-Term Memory; GRU = Gated Recurrent Unit; Ens. = Ensemble . The "Others" are techniques less used, such as *K*-Nearest Neighbors (*K*-NN), DeGroot's model, and so on.

combination of them. In the following sections, we highlighted the challenges and opportunities.

Challenges and Opportunities

Hate speech detection is a complex phenomenon and difficult to recognize, both by humans and machines. Despite the efforts of the scientific community, different open challenges can be highlighted:

- Issues with datasets: include bias because, in many cases, most data belong to the same user. Thus, dataset bias can overestimate the current state-of-the-art (Arango et al., 2019; Calabrese et al., 2021). In particular, one of the most widely used datasets, proposed in Waseem and Hovy (2016), most of the data are generated by a few users. The dataset has more than 16k tweets annotated as racist, sexist, and neither sexist nor racist, where only nine users sent the 1,972 for racist content;
- Context-dependent: transfers poorly across datasets, different approaches present high performance, however only within specific datasets, in which training and test sets were taken from the same dataset (Arango et al., 2019; Gröndahl et al., 2018; Senarath & Purohit, 2020). This issue can be motivated by the influence of the social-demographic and cultural context of the dataset collection that can affect the data sampling and annotation methodology (Waseem et al., 2018);
- Polysemy words: when the word has many different meanings, hidden the actual text interpretation (Senarath & Purohit, 2020);
- Imbalanced dataset: detection methods should not be vulnerable to imbalanced classes. Usually, hate speech datasets are highly imbalanced, with a small percentage of hate content, while most data are non-hate content. Practical resources often need to focus on the minority class (hate content). Therefore, the results evaluated using micro-average metrics on the entire dataset can hide the real performance of minority classes (Charitidis et al., 2020; Zhang & Luo, 2019);
- Despite the efforts to automatically identify hate speech, a limitation is classifying messages without explicitly hateful words (Alorainy et al., 2019; MacAvaney et al., 2019);

Despite the challenges, we also can point out some opportunities in this field.

Feature selection: There is a clear lack of investigation on the impact of the feature selection process since text representation can deal with high dimensionality. In a study performed in Robinson et al. (2018), the authors stated that automatic feature selection algorithms reduced about 90% of the feature space but only selected generic features. Therefore, to understand the contribution of distinct features to hate speech detection, there must be a focus on the existing feature selection techniques, which have proven to affect classification performance significantly.

Metadata: It is relevant that we can transpose our exploitative research into different languages. However, the study of features or indeed approaches for feature representation or metadata that works for more than one language is lacking since online social media platforms can offer a wide variety of information that improve the predictability of hateful (abusive) content (Chatzakou et al., 2017; Founta et al., 2019; Pitsilis et al., 2018), regardless of the text. Furthermore, in the study performed in Ribeiro et al. (2018), the authors have shown that users who produce hate speech are strongly linked. Therefore, metadata features can be helpful in this context.

Hate type: Better defining the specific characteristics of each type of hate speech (racism, gender hate, LGBT hate, religion, ethnicity, political view, etc.) can be potentially a significant advancement in this area.

Comparative studies: As we have pointed out, studies across datasets can help the analysis of the resulting generalization models. In addition, different studies explored only the proposed dataset that often is not publicly available (Del Vigna et al., 2017; Nobata et al., 2016). Comparative studies using different models, features, and datasets are also necessary to understand better what is more effective for hate speech detection on online social media.

Multilingual research: Many researchers have explored datasets in only one language, the majority in English, which creates a lack of work focusing on cross-lingual scenarios. Few works use multilingual or bilingual content on social media platforms from the different Indian dialects (code-mixed language) (Kumar et al., 2019; Santosh & Aravind, 2019). Different particularities, such as distinct grammatical constructions and spelling variations, make the hate speech detection task in this context more difficult (Sreelakshmi et al., 2020). In order to deal with this, classification models and the datasets need to be more robust to lead to better classification performance on the code-mixed scenarios.

Ensemble learning: This approach has received relatively little attention in the context of hate speech detection. Moreover, ensemble methods have improved the results in different classification tasks.

"Memes" analysis: In certain cultures, there is heavy use of image-based with text dissemination of hate-related content. Such analysis has not yet been explored in this field, even though the distribution of such material is mainly done via social media sharing.

Free speech: There is a lack of comparative analysis of samples of free speech text and hate speech. For instance, in a study performed in Casula et al. (2021), the authors discussed the effects of the moderation policies to avoid a toxic online environment in free speech. The researchers affirmed that even though online social media platforms state that they have developed a more inclusive online discourse environment, the moderation policies on online social media platforms can inhibit free speech and precipitates self-censorship. Since the preservation of free speech is essential in a democratic world, there is a need to create a mathematical analysis and definition of the main differences between those two models.

Conclusion

In this paper, we have presented a critical overview of automatic hate speech detection in text from the period between 2015 and 2021. So far, this task has been designed as a supervised learning problem and has used different techniques for feature extraction. Several works have applied simple features and feature extraction techniques, such as BOW, *n*-grams, or Term frequency, which provided a reasonable classification performance. Lexical resources are often used considering negative or derogatory words and have been employed as features or strategies for dataset collection. The pre-trained text embedding has been shown useful for abusive text classification. Features such as sentiment, meta-information, and extracted using DNN are relevant approaches and can improve the result when used to learn additional information. Other less frequently used features are FKGL and FRE scores, pattern features, LDA, so on.

Judging which approaches are the best is a complex issue because several studies evaluate only one dataset, and many are private. Hate speech detection is a recent subject, and different weaknesses still need to be explored.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article: This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and Scholarship no. 88887.484211/2020-00.

ORCID iDs

George D. C. Cavalcanti https://orcid.org/0000-0001-7714-2283
Márjory Da Costa-Abreu https://orcid.org/0000-0001-7461-7570

References

Al-Hassan, A., & Al-Dossari, H. (2019). *Detection of hate speech in social networks: A survey on multilingual corpus* [Conference session]. *6th International Conference on Computer Science and Information Technology*, *9*(2), 83–100.

Al-Makhadmeh, Z., & Tolba, A. (2020). Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. *Computing*, *102*(2), 501–522.

Almatarneh, S., Gamallo, P., Pena, F. J. R., & Alexeev, A. (2019). *Supervised classifiers to identify hate speech on english and spanish tweets* [Conference session]. International Conference on Asian Digital Libraries. Springer (pp. 23–30).

Almerekhi, H., Kwak, H., Jansen, B. J., & Salminen, J. (2019). *Detecting toxicity triggers in online discussions* [Conference session]. Proceedings of the 30th ACM Conference on Hypertext and Social Media (pp. 291–292).

Alorainy, W., Burnap, P., Liu, H., & Williams, M. L. (2019). The enemy among us: Detecting cyber hate speech with threats-based othering language embeddings. *ACM Transactions on the Web*, *13*(3), 1–26.

Alsafari, S., Sadaoui, S., & Mouhoub, M. (2020). Hate and offensive speech detection on arabic social media. *Online Social Networks and Media*, *19*, 100096.

Arango, A., Pèrez, J., & Poblete, B. (2019). *Hate speech detection is not as easy as you may think: A closer look at model validation. SIGIR'19* (p. 4554). ACM.

Aswani, R., Kar, A. K., & Ilavarasan, P. V. (2019). Experience: managing misinformation in social mediainsights for policymakers from twitter analytics. *Journal of Data and Information Quality*, *12*(1), 1–18.

Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (2020). Machine learning techniques for hate speech classification of Twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, *38*, 100311.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., Rosso, P., & Sanguinetti, M. (2019). *Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter* [Conference session]. Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 54–63).

Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018). *A dataset of Hindi-English code-mixed social media text for hate speech detection* [Conference session]. Proceedings of the Second Workshop on Computational Modeling of Peoples Opinions, Personality, and Emotions in Social Media (pp. 36–41).

Bouazizi, M., Niida, N., & Ohtsuki, T. (2021). *All-in-one hate speech detectors may not be what you want* [Conference session]. 2021 The 4th International Conference on Software Engineering and Information Management, ICSIM 2021. ACM (p. 165170). https://doi.org/10.1145/3451471.3451498.

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, *7*(2), 223–242.

Burnap, P., & Williams, M. L. (2016). Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, *5*(1), 11.

Calabrese, A., Bevilacqua, M., Ross, B., Tripodi, R., & Navigli, R. (2021). *AAA: Fair evaluation for abuse detection systems wanted* [Conference session]. 13th ACM Web Science Conference 2021, WebSci '21. ACM (p. 243252). https://doi.org/10.1145/3447535.3462484.

Cao, R., Lee, R. K. W., & Hoang, T. A. (2020). *Deephate: Hate speech detection via multi-faceted text representations* [Conference session]. In: 12th ACM Conference on Web Science, WebSci '20. ACM (p. 1120). https://doi.org/10.1145/3394231

Casula, P., Anupam, A., & Parvin, N. (2021). *We found no violation!: Twitter's violent threats policy and toxicity in online discourse* [Conference session]. C&T '21: Proceedings of the 10th International Conference on Communities & Technologies Wicked Problems in the Age of Tech, C&T '21. ACM (p. 151159). https://doi.org/10.1145/3461564.3461589.

Chakraborty, P., & Seddiqui, M. H. (2019). *Threat and abusive language detection on social media in bengali language* [Conference session]. In: 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT). IEEE (pp. 1–6).

Charitidis, P., Doropoulos, S., Vologiannidis, S., Papastergiou, I., & Karakeva, S. (2020). Towards countering hate speech against journalists on social media. *Online Social Networks and Media*, *17*, 100071.

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). *Mean birds: Detecting aggression and bullying on twitter* [Conference session]. In: Proceedings of the 2017 ACM on Web Science Conference (pp. 13–22).

Cohen-Almagor, R. (2013). Freedom of expression v. Social responsibility: Holocaust denial in canada. *Journal of Mass Media Ethics*, *28*(1), 42–56.

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S. (2020). A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology*, *20*(2), 1–22.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). *Automated hate speech detection and the problem of offensive language* [Conference session]. Eleventh International AAAI Conference on Web and Social Media.

del Arco, F. M. P., Molina-Gonzlez, M. D., Urea-Lpez, L. A., & Martn-Valdivia, M. T. (2021). Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, *166*, 114120. https://doi.org/10.1016/j.eswa.2020.114120

Del Vigna, F., Cimino, A., DellOrletta, F., Petrocchi, M., & Tesconi, M. (2017). *Hate me, hate me not: Hate speech detection on Facebook* [Conference session]. Proceedings of the First Italian Conference on Cybersecurity (ITASEC17) (pp. 86–95).

De Marneffe, M. C., & Manning, C. D. (2008). *Stanford typed dependencies manual* (Technical report). Stanford University.

DeSouza, G., & Da-Costa-Abreu, M. (2020). *Automatic offensive language detection from twitter data using machine learning and feature selection of metadata* [Conference session]. 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1–6). https://doi.org/10.1109/IJCNN48605.2020.9207652

Dorris, W., Hu, R. R., Vishwamitra, N., Luo, F., & Costello, M. (2020). *Towards automatic detection and explanation of hate speech and offensive language* [Conference session]. Proceedings of the Sixth International Workshop on Security and Privacy Analytics, IWSPA '20. ACM (p. 2329). https://doi.org/10.1145/3375708.3380312.

Elisabeth, D., Budi, I., & Ibrohim, M. O. (2020). *Hate code detection in indonesian tweets using machine learning approach: A dataset and preliminary study* [Conference session]. 2020 8th International Conference on Information and Communication Technology (ICoICT). IEEE (pp. 1–6).

Facebook, C. S. (2020). *Hate speech*. Retrieved September 9, 2020, from https://www.facebook.com/communitystandards/hate_speech

Fortuna, P., da Silva, J. R., Soler-Company, J., Wanner, L., & Nunes, S. (2019). *A hierarchically-labeled portuguese hate speech dataset* [Conference session]. Proceedings of the Third Workshop on Abusive Language Online (pp. 94–104).

Fortuna, P., & Nunes, S. (2019). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4), 1–30.

Founta, A. M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., & Leontiadis, I. (2019). *A unified deep learning* *architecture for abuse detection* [Conference session]. Proceedings of the 10th ACM Conference on Web Science (pp. 105–114).

Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). *Large scale crowdsourcing and characterization of twitter abusive behavior* [Conference session]. Twelfth International AAAI Conference on Web and Social Media.

Giachanou, A., & Rosso, P. (2020). *The battle against online harmful information: The cases of fake news and hate speech* [Conference session]. Proceedings of the 29th ACM International Conference on Information & Knowledge Management. ACM (pp. 3503–3504).

Gitari, N. D., Zhang, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215–230.

Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A. A., Gergory, Q., Gnanasekaran, R. K., Gunasekaran, R. R., Hoffman, K. M., Hottle, J., Jienjitlert, V., Khare, S., Lau, R., Martindale, M. J., Naik, S., Nixon, H. L., & Wu, D. M. (2017). *A large labeled corpus for online harassment research* [Conference session]. Proceedings of the 2017 ACM on Web Science Conference (pp. 229–233).

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). *All you need is "love": Evading hate speech detection* [Conference session]. Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security, AISec '18. ACM (p. 212). https://doi.org/10.1145/3270101

Hayaty, M., Adi, S., & Hartanto, A. D. (2020). Lexicon-based indonesian local language abusive words dictionary to detect hate speech in social media. *Journal of Information Systems Engineering and Business Intelligence*, *6*(1), 9–17.

Hendrawan, R., Adiwijaya, Al, & Faraby, S. (2020). *Multilabel classification of hate speech and abusive words on indonesian twitter social media* [Conference session]. 2020 International Conference on Data Science and Its Applications (ICoDSA) (pp. 1–7). https://doi.org/10.1109/ICoDSA50139.2020.9212962.

Ibrohim, M. O., & Budi, I. (2019). *Multi-label hate speech and abusive language detection in Indonesian Twitter* [Conference session]. Proceedings of the Third Workshop on Abusive Language Online. ACL (pp. 46–57).

Indurthi, V., Syed, B., Shrivastava, M., Chakravartula, N., Gupta, M., & Varma, V. (2019). *Fermi at semeval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women on Twitter* [Conference session]. Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 70–74).

Kapil, P., & Ekbal, A. (2020). A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems*, *210*, 106458. https://doi.org/10.1016/j.knosys.2020.106458

Kar, A. K., & Aswani, R. (2021). How to differentiate propagators of information and misinformation–insights from social media analytics based on bio-inspired computing. *Journal of Information and Optimization Sciences*, *42*(6), 1307–1335.

Khan, M. M., Shahzad, K., & Malik, M. K. (2021). Hate speech detection in roman urdu. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *20*(1), 1–19. https://doi.org/10.1145/3414524

Kumar, R., Reganti, A., Bhatia, A., & Maheshwari, T. (2019). *Aggression-annotated corpus of hindi-english code-mixed data* [Conference session]. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA) (pp. 1425–1431).

Liu, H., Burnap, P., Alorainy, W., & Williams, M. L. (2019). *Fuzzy multi-task learning for hate speech type identification* [Conference session]. The World Wide Web Conference (pp. 3006–3012).

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

Markov, I., LjubešićN,Fišer, D., & Daelemans, W. (2021). *Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection* [Conference session]. Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. ACL (pp. 149–159). https://aclanthology.org/2021.wassa-1.16

Marpaung, A., Rismala, R., & Nurrahmi, H. (2021). *Hate speech detection in Indonesian Twitter texts using bidirectional gated recurrent unit* [Conference session]. 2021 13th International Conference on Knowledge and Smart Technology (KST). IEEE (pp. 186–190). https://doi.org/10.1109/KST51265.2021.9415760

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). *Spread of hate speech in online social media* [Conference session]. Proceedings of the 10th ACM Conference on Web Science (pp. 173–182).

Miok, K., Nguyen-Doan, D., Škrlj, B., Zaharie, D., & Robnik-šikonja, M. (2019). *Prediction uncertainty estimation for hate speech classification* [Conference session]. International Conference on Statistical Language and Speech Processing. Springer (pp. 286–298).

Miškolci, J., Kováčová, L., & Rigová, E. (2020). Countering hate speech on facebook: The case of the roma minority in slovakia. *Social Science Computer Review*, *38*(2), 128–146.

Mondal, M., Silva, L. A., & Benevenuto, F. (2017). *A measurement study of hate speech in social media* [Conference session]. Proceedings of the 28th ACM Conference on Hypertext and Social Media, HT '17. ACM (p. 8594). https://doi.org/10.1145/3078714.3078723

Mossie, Z., & Wang, J. H. (2020). Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, *57*(3), 102087. https://doi.org/10.1016/j.ipm.2019.102087

Nandhini, B. S., & Sheeba, J. I. (2015). *Cyberbullying detection and classification using information retrieval algorithm* [Conference session]. Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015), ICARCSET '15. ACM.

Naseem, U., Razzak, I., & Eklund, P. W. (2021). A survey of preprocessing techniques to improve short-text quality: A case study on hate speech detection on twitter. *Multimedia Tools and Applications*, *80*, 35239–35266. https://doi.org/10.1007/s11042-020-10082-6

Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, *1*(1), 100007.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). *Abusive language detection in online user content* [Conference session]. Proceedings of the 25th International Conference on World Wide Web (pp. 145–153).

Nugroho, K., Noersasongko, E., Purwanto Fanani, AZ, & Basuki, RS. (2019). *Improving random forest method to detect hatespeech and offensive word* [Conference session]. 2019 International Conference on Information and Communications Technology (ICOIACT). IEEE (pp. 514–518).

Paschalides, D., Stephanidis, D., Andreou, A., Orphanou, K., Pallis, G., Dikaiakos, M. D., & Markatos, E. (2020). Mandola: A big-data processing and visualization platform for monitoring and detecting online hate speech. *ACM Transactions on Internet Technology*, *20*(2), 1–21.

Philipp, K., & Roman, K. (2019). *Youtoxic english (version 1.0.0)*. https://zenodo.org/record/2586669#.X4l053VKjeR

Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, *48*(12), 4730–4742.

Plaza-Del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2020). Detecting misogyny and xenophobia in spanish tweets using language technologies. *ACM Transactions on Internet Technology*, *20*(2), 1–19.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, *55*, 477–523.

Pratiwi, N. I., Budi, I., & Alfina, I. (2018). *Hate speech detection on indonesian instagram comments using fasttext approach* [Conference session]. 2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE (pp. 447–450).

Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A., & Meira, W., Jr. (2018). *Characterizing and detecting hateful users on twitter* [Conference session]. Twelfth International AAAI Conference on Web and Social Media.

Rizos, G., Hemker, K., & Schuller, B. (2019). *Augment to prevent: short-text data augmentation in deep learning for hate-speech classification* [Conference session]. Proceedings of the 28th ACM International Conference on Information and Knowledge Management (pp. 991–1000).

Robinson, D., Zhang, Z., & Tepper, J. (2018). *Hate speech detection on Twitter: feature engineering vs feature selection* [Conference session]. European Semantic Web Conference, Springer (pp. 46–49).

Rodriguez, A., Argueta, C., & Chen, Y. L. (2019). *Automatic detection of hate speech on facebook using sentiment and emotion analysis* [Conference session]. 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). IEEE (pp. 169–174).

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *Bochumer Linguistische Arbeitsberichte*, *17*, 6–9.

Şahi, H., Kılıc, Y., & Salam, R. B. (2018). *Automated detection of hate speech towards woman on Twitter* [Conference session]. 2018 3rd International Conference on Computer Science and Engineering (UBMK). IEEE (pp. 533–536).

Sajjad, M., Zulifqar, F., Khan, M. U. G., & Azeem, M. (2019). *Hate speech detection using fusion approach* [Conference session]. 2019 International Conference on Applied and Engineering Mathematics (ICAEM). IEEE (pp. 251–255).

Salminen, J., Almerekhi, H., Milenkovic, M., Jung, S. G., An, J., Kwak, H., & Jansen, B. J. (2018). *Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media*. ICWSM (pp. 330–339).

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

Santosh, T., & Aravind, K. (2019). *Hate speech detection in hindienglish code-mixed social media text* [Conference session]. Proceedings of the ACM India Joint International Conference on Data Science and Management of Data (pp. 310–313).

Schmidt, A., & Wiegand, M. (2017). *A survey on hate speech detection using natural language processing*. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media (pp. 1–10).

Senarath, Y., & Purohit, H. (2020). *Evaluating semantic feature representations to efficiently detect hate intent on social media* [Conference session]. 2020 IEEE 14th International Conference on Semantic Computing (ICSC) (pp. 199–202). https://doi.org/10.1109/ICSC.2020.00041

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, *311*, 18–38.

Sohn, H., & Lee, H. (2019). *Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations* [Conference session]. 2019 International Conference on Data Mining Workshops (ICDMW). IEEE (pp. 551–559).

Sreelakshmi, K., Premjith, B., & Soman, K. P. (2020). Detection of hate speech text in hindi-english code-mixed data. *Procedia Computer Science*, *171*, 737–744.

Teh, P. L., Cheng, C. B., & Chee, W. M. (2018). *Identifying and categorising profane words in hate speech* [Conference session]. Proceedings of the 2nd International Conference on Compute and Data Analysis, ICCDA 2018. ACM (p. 6569). https://doi.org/10.1145/3193077.3193078

Twitter. (2020). *Hateful conduct policy*. Retrieved September 9, 2020, from https://help.twitter.com/en/rules-and-policies/hate

Unsvåg, E. F., & Gambäck, B. (2018). *The effects of user features on twitter hate speech detection*. Proceedings of the 2nd Workshop on Abusive Language Online (ALW2) (pp. 75–85).

Vitiugin, F., Senarath, Y., & Purohit, H. (2021).) *Efficient detection of multilingual hate speech by using interactive attention network with minimal human feedback* [Conference session]. 13th ACM Web Science Conference 2021, WebSci '21. ACM (p. 130138). https://doi.org/10.1145/3447535.3462495

Waseem, Z. (2016). *Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter* [Conference session]. Proceedings of the First Workshop on NLP and Computational Social Science. ACL (pp. 138–142). http://aclweb.org/anthology/W16-5618.

Waseem, Z., & Hovy, D. (2016). *Hateful symbols or hateful people? predictive features for hate speech detection on Twitter* [Conference session]. Proceedings of the NAACL Student Research Workshop. ACL (pp. 88–93). http://www.aclweb.org/anthology/N16-2013.

Waseem, Z., Thorne, J., & Bingel, J. (2018). Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In J. Golbeck (Ed.), *Online harassment* (pp. 29–55). Springer.

Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, *6*, 13825–13835.

Wich, M., Breitinger, M., Strathern, W., Naimarevic, M., Groh, G., & Pfeffer, J. (2021). *Are Your Friends also haters? Identification of Hater Networks on social media: Data Paper* (p. 481485). ACM.

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). *Detection of abusive language: the problem of biased datasets* [Conference session]. Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 602–608).

Wulczyn, E., Thain, N., & Dixon, L. (2017). *Ex machina: Personal attacks seen at scale* [Conference session]. Proceedings of the 26th International Conference on World Wide Web (pp. 1391–1399).

YouTube. (2020). *Hate speech policy*. Retrieved September 9, 2020, from https://support.google.com/youtube/answer/2801939?hl = en

Zhang, Z., & Luo, L. (2019). Hate speech detection: A solved problem? The challenging case of long tail on twitter. *Semantic Web*, *10*(5), 925–945.

Zhang, Z., Robinson, D., & Tepper, J. (2018). *Hate speech detection using a convolution-lstm based deep neural network*. ESWC 2018: The Semantic Web.

Zhou, Y., Yang, Y., Liu, H., Liu, X., & Savage, N. (2020). Deep learning based fusion approach for hate speech detection. *IEEE Access*, *8*, 128923–128929.

Zimmerman, S., Kruschwitz, U., & Fox, C. (2018). *Improving hate speech detection with deep learning ensembles* [Conference session]. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).