information

Review

A Literature Review of Textual Hate Speech Detection Methods and Datasets

Fatimah Alkomah [1,2,*] and Xiaogang Ma [1]

[1] Department of Computer Science, University of Idaho, Moscow, ID 83844-1010, USA; max@uidaho.edu
[2] Department of Information Systems, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia
* Correspondence: alko3916@vandals.uidaho.edu

Abstract: Online toxic discourses could result in conflicts between groups or harm to online communities. Hate speech is complex and multifaceted harmful or offensive content targeting individuals or groups. Existing literature reviews have generally focused on a particular category of hate speech, and to the best of our knowledge, no review has been dedicated to hate speech datasets. This paper systematically reviews textual hate speech detection systems and highlights their primary datasets, textual features, and machine learning models. The results of this literature review are integrated with content analysis, resulting in several themes for 138 relevant papers. This study shows several approaches that do not provide consistent results in various hate speech categories. The most dominant sets of methods combine more than one deep learning model. Moreover, the analysis of several hate speech datasets shows that many datasets are small in size and are not reliable for various tasks of hate speech detection. Therefore, this study provides the research community with insights and empirical evidence on the intrinsic properties of hate speech and helps communities identify topics for future work.

Keywords: hate speech detection; literature review; hate speech datasets; hate speech methods

Citation: Alkomah, F.; Ma, X. A Literature Review of Textual Hate Speech Detection Methods and Datasets. *Information* **2022**, *13*, 273. https://doi.org/10.3390/info13060273

Academic Editors: Diego Reforgiato Recupero and José J. Pazos Arias

Received: 21 April 2022
Accepted: 24 May 2022
Published: 26 May 2022

1. Introduction

Hate speech is a poisonous discourse that can swiftly spread on social media or due to prejudices or disputes between different groups within and across countries [1]. A hate crime refers to crimes committed against a person due to their actual or perceived affiliation with a specific group (https://www.ilga-europe.org/what-we-do/our-advocacy-work/hate-crime-hate-speech, accessed on 1 March 2022) The protected characteristics of Facebook define hate speech as an attack on an individual's dignity, including their race, origin, or ethnicity. According to Twitter policies, tweets should not be used to threaten or harass others due to their ethnicity, gender, religion, or any other factor. In addition to age, caste, and handicap, YouTube also censors content that promotes violence or hatred toward certain persons or groups. Often, hate speech regarding online radicalization or criminal activities is studied [2]; however, hate speech has also been discussed in other contexts [3].

There is strong motivation to study the automatic detection of hate speech due to the overwhelming online spread of information [4]. The detection of hate speech is crucial to reducing crime and protecting people's beliefs. This study is especially important in the face of ongoing wars, distorting reality and dehumanizing the attacked Ukrainian nation. Studies have shown an increase in hate speech against China on online social media, especially racist and abusive content accusing people of causing the COVID-19 outbreak. On the other hand, a lower rate of hate speech reduces crime, such as cyberbullying, which significantly affects social tranquility [5], leading to minimal cyber-attacks [6].

Despite the many studies in the field, hate speech is still problematic and challenging. The literature reports that both humans and machine learning models have difficulty detecting hate speech due to the complexity and variety of hate speech categories. Hate

is characterized by more extreme behaviors associated with prejudice. Figure 1 shows numerous aspects of hate speech. Hate speech is seen as a layer between aggressive and abusive text; however, all of these share the offensive aspect. On the other hand, sexist, homophobic, and religious hate are relatively different as they target a group of people or a gender. The figure shows that the separation of the concept is complicated, a challenge that has been previously discussed [7]. Specific hate speech definitions are contentious [7]; racist and homophobic tweets, for example, are more likely to be labeled as hate speech than other types of offensive or abusive content. Therefore, there is no way to generalize whether an inflammatory text is hate speech [8]. The expansion of the representation of short documents, such as on Twitter, is a significant issue; they cause additional challenges to the traditional bag-of-words model, resulting in data sparsity due to insufficient contextual information [9]. Furthermore, many datasets [10] may differ due to these classifications and definitions, making it challenging to compare machine learning models. In other words, hate speech notions are all within the umbrella of abusive text, according to Poletto et al. [1].

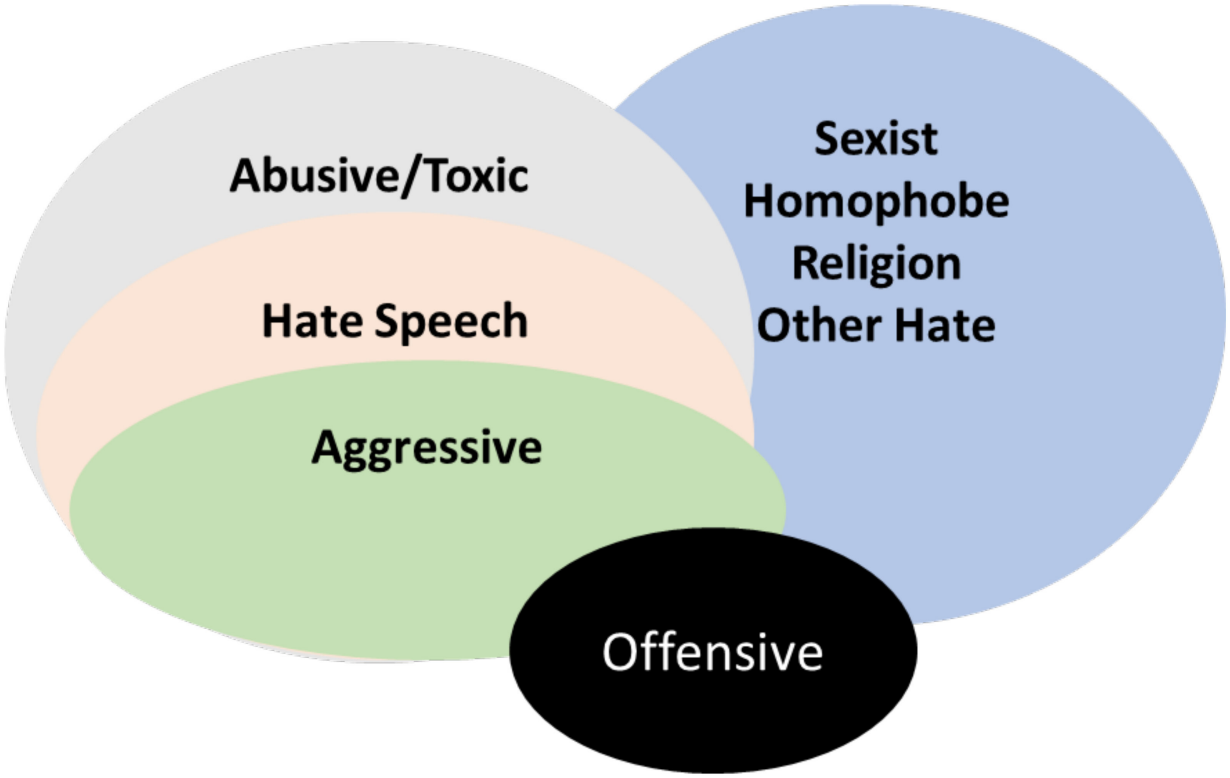

Figure 1. Hierarchy of hate speech concepts.

The literature showed several literature review articles; however, most of them targeted one area of the literature or are relatively old [7,11–14]. The literature study in [7] was devoted to building a generic metadata architecture for hate speech classification based on predefined score groups using semantic analysis and fuzzy logic analysis. The study in [11] targeted hate speech concerning gender, religion, and race related to cyberterrorism and international legal frameworks; however, the study did not focus on Twitter or datasets for machine learning. Most of the papers cited in [12] are related to the legal literature that defines hate speech for criminal sanctions. The study in [13] was devoted to hate speech geographical aspects, social media platform diversity, and the generic qualitative or quantitative methods used by researchers. To the best of our knowledge, no review has been dedicated to English hate speech dataset analysis. This paper aimed to deeply review hate speech concepts, methods, and datasets to provide researchers with an insight into the latest state-of-the-art studies in hate speech detection.

This study carries a systematic literature review based on the methodology of Tranfield et al. [15]. The method synthesizes and extracts results on evidence-based systematic literature extracted based on four aspects: the research question, review criteria, final literature review, and data extraction and synthesis. The studied topic is heterogeneous in its wide range of methods and homogenous in terms of using textual Twitter hate speech content, but is bounded by a specific research question. Therefore, the methodology of Tranfield et al. is applicable [16].

The central research question is as follows.

RQ: What are the dominant hate speech concepts, datasets, and machine learning methods?

The paper is outlined as follows. Section 2 discusses the research methodology. Results and content analysis are discussed in Section 3. Section 4 discusses research challenges and future directions. Finally, conclusions are illustrated in Section 5.

2. Methodology

This study focuses on machine learning classification models, hate speech datasets, and the most significant features of hate speech models in light of such a widespread phenomenon. We applied a survey and content analysis to retrieve and analyze relevant studies based on the methodology of Tranfield et al. [15]. The methodology of Tranfield et al. [15] is based on four phases: research question formulation, review criteria definition, final literature review criteria, and data extraction and combination. The primary research question in this paper is to identify the dominant approaches to hate speech and its significant datasets that are commonly used by the research community. Following the research question, the review criteria include: (1) selecting a set of hate speech terms combined with machine learning methods. The selection includes terms such as hate, cyberhate, racist language, offensive language, online harm, poisonous speech, religion hate, ethnicity, nationality, race, gender hate, aggressive, abusive, misogyny, sexist speech, homophones, and homophobic language; (2) the following data repositories were consulted: Science Direct, Scopus, Emerald, IEEE Xplore, and Google Scholar; (3) the study was conducted between 2016 and 2021; and (4) papers must be in English with a minimum length of 5 pages. During the data extraction and synthesis phase, the collection of literature was analyzed and then grouped according to the review criteria that were initially put in place.

Figure 2 shows the proposed methodology. The initial analysis process followed the approach of Tranfield et al.; papers that passed the RQ and the initial inclusion criteria were included in this review. The initial analysis is based on the analysis of the paper title, abstract, and introduction. Consequently, papers that did not pass the inclusion criteria were eliminated. Then, the data extraction and synthesis phase included papers from the previous step and comprehensively analyzed them by reading and grouping them into meaningful categories. During this step, the thematic analysis technique was used [17]. The thematic approach is based on the recurring patterns of hate speech detection. With careful reading and analysis, the critical information related to the RQ was coded using the content analysis in each paper. The codes were then converted and grouped into higher-order themes. Finally, after deep understanding and linking knowledge embodied in analyzed papers, a set of crucial challenges was highlighted.

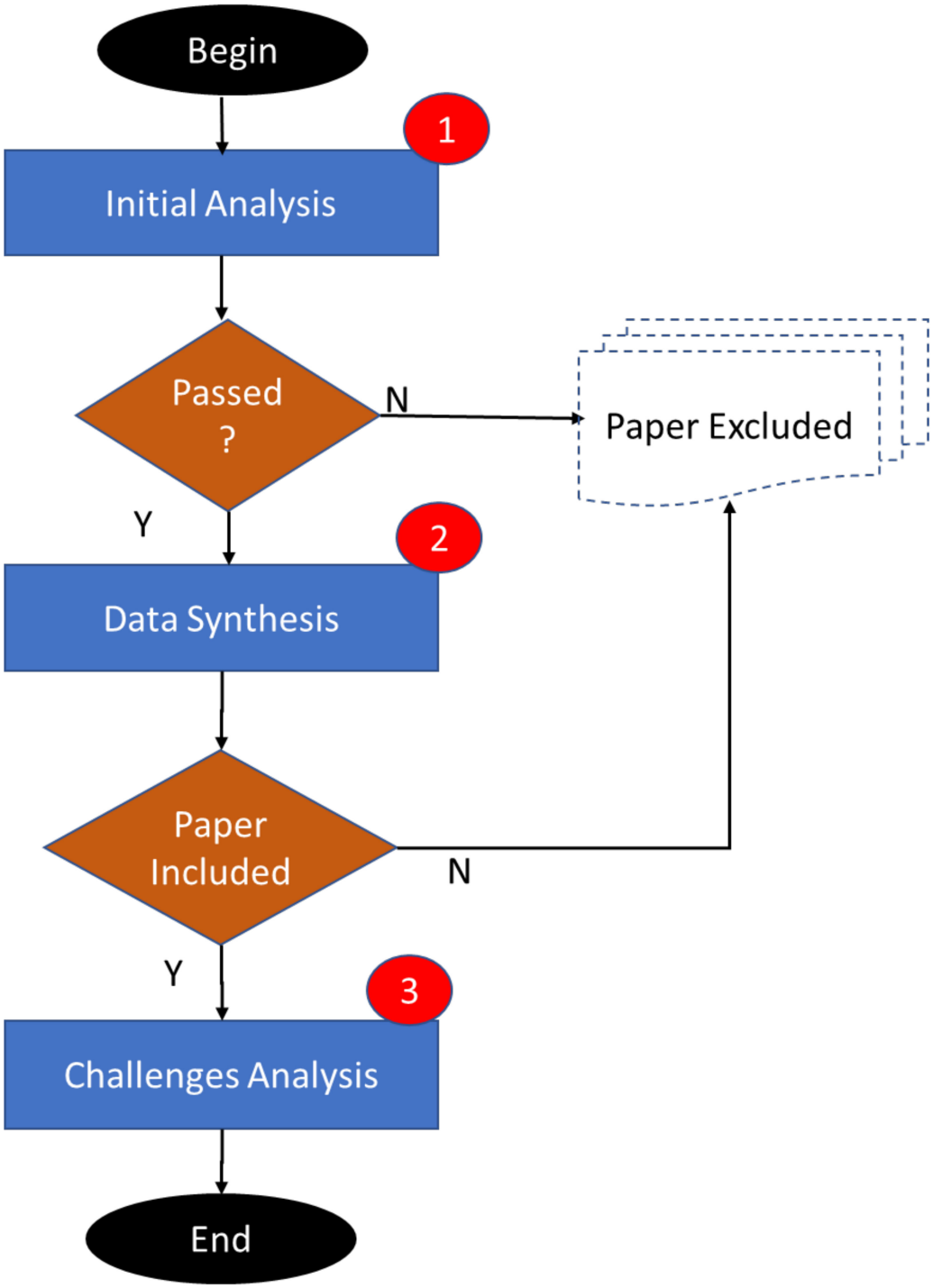

Figure 2. Proposed methodology.

3. Results and Analysis

Following the proposed approach in Figure 2, the following steps were executed. In Step 1, the papers that passed the initial inclusion criteria were analyzed (512 papers). In Step 2, the thematic analysis resulted in the final set of papers (138 papers) analyzed in accordance with the approach of Tranfield et al. Step 3 is discussed in Section 4. The most relevant papers are shown in as shown in Appendix A.

The distribution of relevant papers over the years is illustrated in Figure 3. The figure shows an increase in hate speech detection over the years. Among the leading research conferences that are dedicated to hate speech are the Sem-Eval, HASCOC, and FIRE confer-

ence series. The Sem-Eval series conference is one of the most specialist conferences on hate speech. In the Sem-Eval series, Task 6 of SemEval-2019 comprises three sub-tasks [18]: A: detecting offensive/non-offensive speech; B: determining offensive speech (insult/nontarget insult); and C: determining the targeted individual or group demonstrating offensive speech. The Offensive Language Identification Dataset (OLID) (14,100 tweets) [18] is part of SemEval-2019 Task 6. In another conference series, for three languages, namely Hindi, German, and English, the HASOC 2020 track of FIRE2020 [19] contained two goals for detecting hate speech: binary (hate/not hate) and multilabel (hate, offensive, and profanity). The initial challenge of FIRE2020 is to distinguish between non-hate–offensive and hate–offensive information.

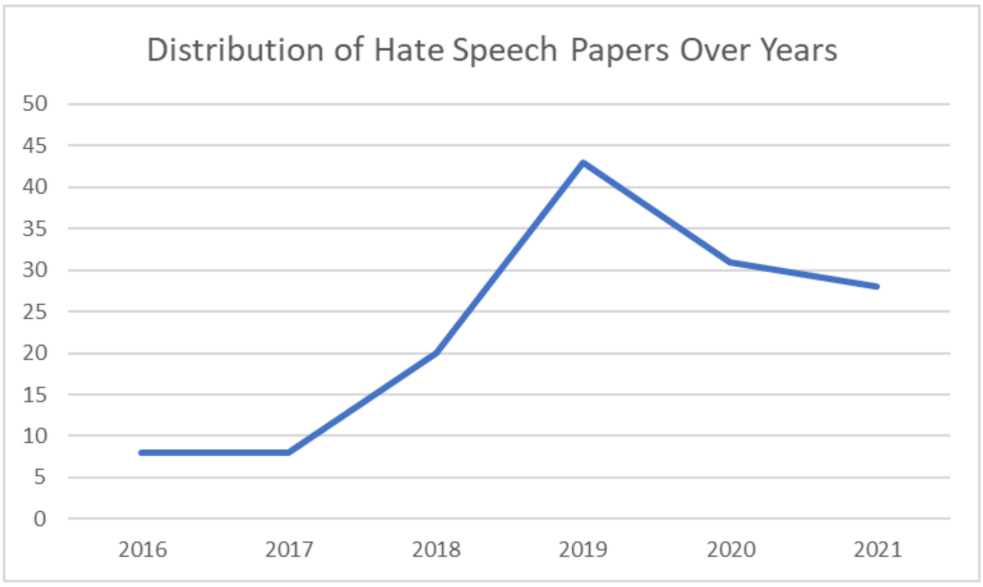

Figure 3. Papers over the years.

The content analysis of this study (Figure 4) shows the distribution of hate speech detection methods. According to the figure and our analysis, the most successful hate speech detection methods use more than one approach for detection (hybrid models). The figure also shows a wide range of shallow TFIDF methods to complex BERT models.

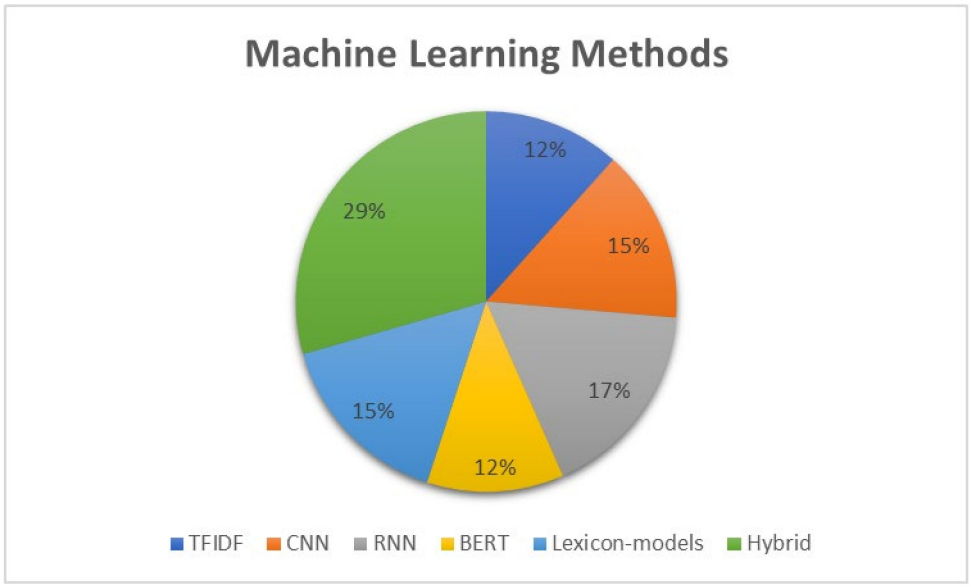

Figure 4. Machine learning models in the studied literature.

3.1. Machine Learning Hate Speech Models

The thematic structures identified in this study are discussed in these major themes: TFIDF, lexicon-based, traditional, and deep learning methods, and hybrid models.

3.1.1. TFIDF Methods

Figure 4 shows that only 12% of the analyzed papers used traditional machine learning methods based on the TFIDF. The term frequency (TF), paired with Inverse Document Frequency (IDF), is a regularly utilized feature in hate speech identification (TFIDF). The TFIDF considers the statistical aspect of input text and the specificity of terms (or chars or POS tags) in the corpus, giving less frequently used words greater weight. For example, in hate speech tweets, the extensive use of hashtags that signify objectionable material, such as "#BanIslam", "#whoriental", and "#whitegenocide", could indicate hostile speech. Therefore, lexicons, hashtags, and POS tags were used in an association rule system [20] to classify hate speech. Several works have been used in this category, such as the author profiling of the Hate Speech Spreader Detection shared task organized by PAN 2021 [21], char/word-IDF [22,23], and Vader/RoBERTa word embeddings [24], trigram features and POS tags [25–27], n-grams [28–32], and emotions [26,33].

3.1.2. Lexicon-Based Methods

Lexicon-based methods depend on using existing keywords (lexicons) fixed by authors or extracted from the literature for hate speech detection. The work of Frenda et al. [34] preprocessed two datasets of misogynistic and sexist tweets over SVM classifiers. Frenda et al. used three datasets: Automatic Misogyny Identification (AMI) IberEval [35], AMI EvalIta [36], and the SRW [28]. A list of 690 lexicon items was added to the TFIDF for characters and words, which was utilized to weight lexical features using Information Gain [34]. The following categories are included in their lexicon: hashtags, vulgarity, abbreviations, femininity, sexuality, and human body. Their model has a precision of 0.76. Their model adopts sentence embeddings, the TFIDF of tweets, Glove Bow over Logistic Regression, and XGboost over the English dataset of the AMI shared problem of EVALITA2018. Their highest performing method was LR, which obtained an accuracy of 0.704. Several other studies have used lexicons in hate speech detection such as the Hate Speech and Offensive Content in Social Media Text (HASOC) detection models [25,37,38]; Jewish and Muslim lexicons [39,40]; hate speech-related migrants and religious minorities of Twitter [41–49]; and InferSent and Google [46,50,51].

3.1.3. Deep Learning Methods

The convolutional neural network (CNN) is a class of artificial neural networks which is most commonly applied to analyze visual imagery. However, it has also been used in textual classification. CNNs use filters and a set of pooling layers for hate speech detection tasks [49,52–69]. Zimmerman et al. [67] tested Waseem [28] and SemEval 2013 sentiment analysis datasets [68]. Zimmerman et al. employed a CNN structure to represent 50 tokens based on CNN parameters (epochs, weights, and batch size). The best-reported model had three epochs and a batch size of ten for positive and negative classifications, with an F1 average score of 75.98%. Gambäck and Sikdar [69] created a dataset of 6655 tweets and used 4 CNN structures to train it. They used n-grams, Word2Vec, and random vectors created at random. The results (using Word2Vec) showed an F1-score of 78.3%. Pre-trained Glove and FastText models were also employed against women and immigrants at an individual and group level [49]. Similarly, in [53,54], the authors used convolutional neural networks to recognize the profiles of hate speech-based word n-grams. Many alternative CNN architectures were used for various content languages [49,55]. In the study which used the SemEval-2019 dataset, for example, word embedding was utilized to detect hate speech in Spanish and English tweets [55].

A recurrent neural network (RNN) is a neural network that uses internal memory to recall its input. The algorithm family, for example, is utilized in Apple's Siri and Google's

voice search. Many algorithms, such as LSTMs and GRUs, are included in RNNs. LSTMs were created to solve the problem of vanishing gradients that can occur while training standard RNNs. The GRU is similar to long short-term memory (LSTM) with a forget gate, but lacks an output gate. Hence, it has fewer parameters. Variations of LSTM were also used for many tasks of hate speech detection, such as hate ideologies [70], COVID-19 and the US election [71–73], the hate detection of hybrid CNN and RNN [74–76], HASOC LSTM models [77–79], BiLSTM models [65,80,81], and generative pretrained transformer models [82–89]. Some studies have added additional features to Twitter network analysis such as user profiles for hate speech detection. For example, Founta et al. [90] proposed two-layer RNNs. The unified model was built on tweet characteristics (the Glove technique) and metadata on persons, networks, and content. They used many datasets to test their model, including the cyberbullying dataset [91], the hateful dataset [28], the offensive dataset [25], the sarcasm dataset [92], and the abusive dataset [84]. The model produced variable results depending on the input characteristics and dataset utilized; nevertheless, the RNN and metadata interleaved model were the best, with an average accuracy of 90.2. The dataset of Waseem and Hovy [28] was used by Founta et al. [93] to detect abusive cyberbullying language on Twitter. Their model blends the LSTM architecture with social network analysis to detect hate speech sources. In order to train the LSTM model, they used FastText embeddings. An F1-score of 0.823 was reported. Corazza et al. [94] used the dataset of Waseem and Hovy [28] with the following algorithms: LSTM, GRU, and BiLSTM. N-grams of words; tweet features, such as emojis and emotion lexica; and social network-specific features are among the features employed. They reported an F1-score of 0.823; the best method was LSTM.

Bidirectional Encoder Representations from Transformers (BERTs) is a transformer-based machine learning technique for natural language processing which was pre-trained and developed by Google based on the knowledge extracted from text (vectors) using the surrounding text to establish the context. Several complex models that used BERT are discussed in the literature such as hateful meme challenges [64]; the study of Ron Ahu et al. [95]; ensembles of BERT [96]; Yu's model of knowledge enhanced vision-language representations [97]; Facebook Hateful Memes (FHM) [98,99]; Liu et al. [100] and BERT over Zampieri [101] three tasks; Caselli et al. [102] who re-trained the BERT for hate speech (HATEBERT), the AbusEval [103], and the HatEval [43] models; Nguyen et al. [104] who employed the BERT model of RoBERTa [105] by training on 80 GB of uncompressed texts of 850 M tweets (16B-word tokens), and BERTweet which outperforms strong baselines RoBERTa-base and XLM-R-base [106]. Caselli et al. [102] used the Reddit Abusive Language English dataset, which contains over one million Reddit messages (43 billion tokens). They compared the following datasets from SemEval 2019: Task 6 OffensEval 2019 [18], the AbusEval [103], and the HatEval [43]. They found that the HATEBERT model was more efficient than the datasets they evaluated, with a 5% increase in precision. Nguyen et al. [104] used the RoBERTa [105] BERT model to train 80 GB of raw texts from 850 M tweets (16 B-word tokens). It was found that BERTweet surpasses the RoBERTa-base and XLM-R-base [106], which are both powerful baselines.

Consequently, RNNs and CNNs have commonly used machine learning methods in deep learning methods; however, research has also shown that hybrid and complex models cover approximately 29% of studied models. For example, Jahan and Oussalah [107] showed that the BERT method covers approximately 33% of studies, followed by LSTM and CNN with 20%.

3.1.4. Hybrid Methods

Hybrid methods combine more than one machine learning method to improve performance, including ensemble models. In addition, hybrid models are considered robust as they incorporate more than one data source, such as the metadata from Twitter. Pitsilis et al. [108] tested an ensemble of RNN classifiers in the datasets of Waseem and Hovy [28]. They used the word frequency vectors and the user sentiment towards each

class of hate speech (neutral, racism, or sexism) based on user tweet history. The result showed that the model provided an F1-score of 0.93. Joulin et al. [109] reported the best performance was when using a combination of LSTM + Random Embedding + Boosted Decision Trees (GBDTs) with an F1-score of 0.93 in the dataset of Waseem and Hovy [28] and using Glove embeddings. Paschalides et al. [110] used the dataset of Davidson et al. [25] to develop a new online Twitter hate speech detection system called MANDOLA. The system is based on many ensembles of CNNs and RNNs that automatically learn abstract feature representations depending on many features: TF-IDF vectors and word embeddings. MANDOLA allows users to visualize online results. The approach reported a balanced accuracy of 0.770.

Other hybrid models include the RETINA model of Masuad et al. [111]; the multi-channel convolutional bidirectional gated recurrent unit (MCBiGRU) [112]; attention-LSTM of Wang and Ding [113]; bi-directional GRU layers with CNN of Wiedemann et al. [114], Setyadi et al. [115], and Ziqi et al. [87]; variation dataset of Gambäck et al. [69] based on the Tsironi [116] model; CAT boost of Qureshi and Sabih [117]; pre-trained word embeddings and metadata [118–120]; ensemble RNN of Pitsilis et al. [108] and [121]; LSTM + Random Embedding + Boosted Decision Trees (GBDTs) of Joulin et al. [109] and Miok et al. [122]; MANDOLA of Paschalides et al. [110]; MCD + LSTM [122], Sachdeva et al. [123], Sajjad et al. [124], and Liu et al. [125]; the fuzzy rule method [126], Ayo et al. [7]; the KDEHatEval team [121], Ali et al. [127], and J et al. [112]; and Google's Jigsaw on the Kaggle Model (https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/overview, accessed on 10 February 2022), Wang and Ding [113]; attention-LSTM model based on a modified version of HAN [128] and a BiGRU-capsule model using the dataset of [46], Wiedemann et al. [114]; bi-directional GRU layers with CNN layers, Setyadi et al. [115], Ziqi et al. [87], and Qureshi and Sabih [117].

3.2. Datasets

Datasets have targeted numerous hate speech categories, as shown in Appendix B. However, examples could be unclear in several datasets, such as Waseem's dataset [27] or hierarchical datasets including the dataset of Basile et al. [58]. Furthermore, datasets are of poor quality because they are not regularly updated when Twitter users use new phrases or abbreviations. In addition, approximately 60% of dataset creators found inter-annotator agreement [1]. Therefore, a useful predictive hate speech detection model requires relevant and non-obsolete datasets. The maturity of datasets is regarded as a one-of-a-kind task for superior quality systems. According to Koco et al. [129], separating annotator groups has a more significant impact on the performance of hate detection systems. They also stated that group consensus impacts the quality of recognition. It has been demonstrated that the identification group of people who publish tweets introduces bias into the dataset, and negative data are challenging to compile and ensure; hence, implicit hate speech is difficult to measure [130]. In addition, many datasets overlap across class labels, as shown by Waseem [50], who found an overlap of 2876 tweets between the Waseem and Hovy dataset.

Our analysis showed that research requires more robust, trustworthy, and large datasets due to the wide applications of hate speech detection. The dataset of Shibly et al. [65] developed a robust and colossal dataset. Those datasets backed by conference or workshop series, such as Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC) [19,38] and SemEval, are probably among the most popular datasets. HASOC is divided into three subtasks: the first focuses on identifying hate speech and offensive language (sub-task A); the second focuses on identifying the type of hate speech (sub-task B); and the third focuses on identifying the hate speech's target group (or persons). The SemEval Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter continues to focus on numerous academics in the SemEval series [83,113]. On the other hand, researchers have paid close attention to SemEval 2019 Task 6 [18]—OffensEval: Identifying and Categorizing Offensive Language on Social Media. The Offensive Language Identification Dataset (OLID) [101], which contains over 14,000 English tweets, is the most

recent (similar tasks as in HASOC). The HASOC 2020 track dataset [19] includes 3708 English tweet samples, however, it is considered a substantial and competitive dataset. Nearby random (reported F1-score of 0.52) had the best performance, whereas fine-grained multilabel categorization had the worst performance (0.26 F1-measure). ElSherief et al. [131] found 25,278 hate instigators and 22,857 target users in their study. Note that datasets that were not built for Twitter (e.g., [132,133]) were not evaluated, and datasets that were not built for English were discarded (e.g., [134]). According to our reasoning, such datasets are not appropriate due to the variety of beneficial qualities. The list of reviewed datasets is shown in Appendix B.

4. Discussion

The analysis of collected papers revealed that defining hate speech is still an issue, as previously discussed [7]. According to Culpeper [135], incitement, face, and impoliteness are all related to hate speech, but they are distinct; the author stated that mapping between the concepts was complicated due to the Oxford English Corpus containing multiple genres during the period of 2000–2006, with varying frequencies of hate and impoliteness. To comprehend white supremacist heteropatriarchy, intersectionality between race, sex, place, and gender has been found to represent a framework of hate speech-related notions on social media [13]. Racist and homophobic tweets were shown to be more hateful than offensive sexist ones [25]. Wasim and Hovy [28] described offensive tweets using 11 criteria, including racial slurs, attacks on minorities, supporting hate speech, misrepresenting the truth, and defending xenophobia or sexism. As a result, hate speech has many meanings, including offensive, abusive, harassing, aggressive, cyberbullying, destructive, and hate speech [136]. However, offensive comments, such as hate speech, are among the most researched NLP topics [129]. According to MacAvaney [10], there is no standard hate speech definition. Nevertheless, according to Schmidt and Wiegand [11], hate speech is a broad phrase for various disparaging content produced by humans.

Hate speech detection is a complex process, partly due to the availability of the datasets and the task of developing a machine learning model with rigid performance. Table 1 divides the essential themes of the papers under consideration into seven categories. As shown in Figure 5, most publications use machine learning models to categorize information into binary classifications with a maximum of three class labels. The HASOC and CrowdFlower datasets are uneven in terms of the positive class, with more offensive tweets in CrowdFlower and more classed as 'neither' in HASOC, as stated in many research studies. Furthermore, detecting abusive language is problematic for various reasons, including word obfuscation, difficulties tracing racist and slurs, and professionally written and abusive language that crosses sentence boundaries. Several projects (such as the SemEval and HASOC tasks) require more robust, trustworthy, and massive datasets due to the wide applications of hate speech detection. However, small datasets are insufficient for generalizing conclusions or capturing features. For example, some research merely employed 200 tweets, generating good results but raising concerns about the reported performance's generality. Therefore, the size of the dataset does not guarantee diversity [137]. Appendix B shows the most relevant papers and their limitations. It was not practical to list all 138 papers; therefore, we selected 23 papers from the themes discussed in Section 3.1. We tried to present different methods that are published in different datasets.

Table 1. Themes of papers.

Category of Papers	Meaning	Papers
Small datasets	A dataset is considered small if it has less than the initial dataset of Waseem, which was 16 K tweets.	[33,41,49,55,76,81,115,115]
Binary classification	A model is considered a binary classification model if it presents a work that classifies hate speech into two or three classes.	[26,28,28,31,37,37,41,41,46,51,59,61,65,65, 67,69,69,78,80,80,87,88,90,93,94,108–110,113,113,114,114,115,115,121,123,126]
Low performance	A low-performance classifier is considered as that which reports a binary classification below 0.6.	[22,22,22,28,30,41,44,46,47,49,55,59,59,59, 78,82,121,138]
Topic-wise detection	A study that uses topic-wise categorization instead of classification.	[41,46,71,76,76,117,117,139]
Lexicon-based dictionary generalization	A low-performance classifier is considered as those which report a binary classification below 0.6.	[24,102,104,106,122,140]
Semi-supervised, clustering, rule-based	A study that uses semi-supervised learning or rule-based methods instead of classification to solve the issue of hate detection.	[7,113,125,141]
Specialized (one dimension of hate speech)	Focused on a specific category of hate speech or dimension of hate speech.	Sarcasm [116] Racism [66,70] Sexism [25,66] General sexist tweets hide a sentiment of hate or misogynistic attitude [34] Detecting profiles [53,54] UK only [40] Retweeting [111] Hate intensity [141] Multi-mixed languages [56] Multi-hierarchical classification [142] Hate speech against immigrants [46] Comments and large text [74,112,143,144]

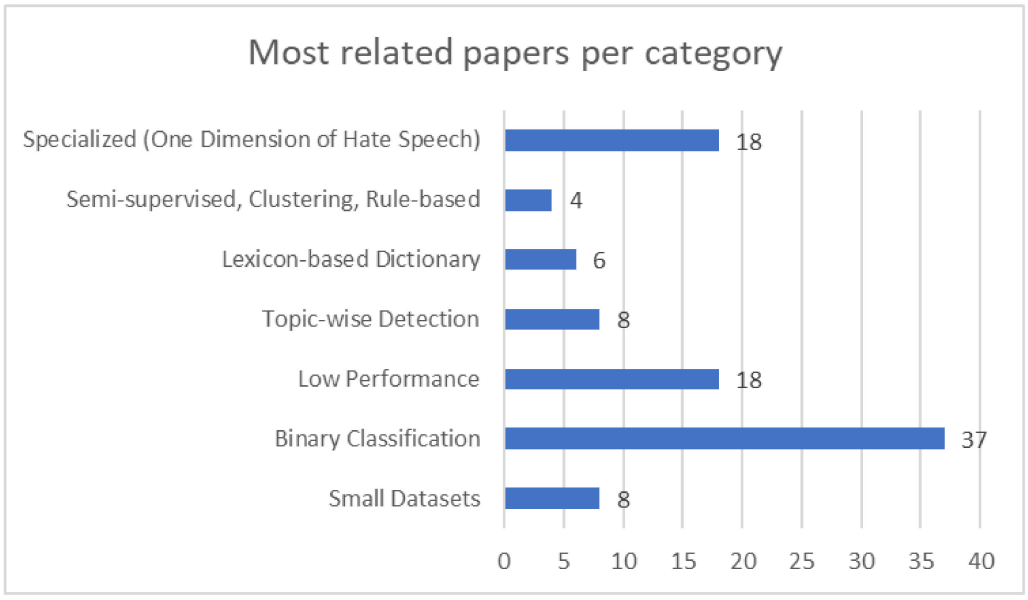

Figure 5. Literature distribution per category of issues.

4.1. Challenges of Machine Learning Models

The literature typically utilizes the following algorithms based on the described standard machine learning hate detection methods: support vector machines (SVM); Naive Bayes (NB); Logistic Regression (LR); Decision Trees (DTs); and K-Nearest Neighbor (KNN).

While the approaches' performance varies depending on the dataset, the LR consistently outperforms the others. Most algorithms rely on n-grams, POS tags, and feelings to detect hate speech, while some integrate lexicons as a bonus feature. However, based on the varied datasets, these techniques produce contradictory conclusions. The fact that these methods have been used for binary classification jobs is one of the major concerns. As a result, without considering the word's context and reviewing vast and robust datasets, their performance may, in most situations, be in direct opposition to deep learning models.

The most promising algorithms in deep learning methods, according to research, are CNN and LSTM. For example, Anand and Eswari [139] explored LSTM and CNN with and without GloVe embeddings. Using the same dataset as CNN, they reported an accuracy of above 97% (GloVe and CNN = 97.27). However, according to the research, the approaches have varying performance levels, and the majority of them focus on a binary hate detection problem (hate/not hate) with multilabel classifications only occasionally being explored. One could be due to the lack of a dataset, while the other could be due to the difficulties of detecting hate speech.

Nonetheless, CNN and LSTM have shown to be effective in several trials. Due to the modest size of the datasets, pre-trained BERT models produce varied outcomes. As a result, the problem with deep machine learning models is similar to past problems with classic features approaches and dataset size. According to the literature, the performance of most hybrid approaches has improved by at least 5% compared to previous detection methods.

Consequently, hybrid machine learning models are dominant in the literature and show promising results; however, without large datasets, generalization remains impossible.

4.2. Challenges of Datasets

Comparing hate speech detection methods is difficult due to the variety of features and datasets. With the imbalanced class distribution, hate speech lacks a discriminative or unique collection of features [140]. Depending on the notability and features or algorithms used, traditional approaches (such as SVM and NB) can outperform deep learning models. For example, the SVM method in [31] outperforms the Hierarchical Attention Network (HAN) [128] and CNN utilizing Glove word embedding [31]. Nugroho et al. [145] showed that RF outperforms a neural network model by approximately 10% (0.72). ElSherief et al. [131] found 25,278 hate instigators and 22,857 target users in their study. They discovered that agreeableness, openness, emotional range, conscientiousness, and extraversion all play a role in detecting the profiles of instigators. They also discovered that hate instigators target more visible users with common personality features, such as anger, sadness, and immoderation. Shallow lexical features [25], dictionaries [146], sentiment analysis [147], linguistic characteristics [148], knowledge-based features [149], and meta-information [21] were described in the literature as features connected to tweets.

Consequently, the major issue of datasets is how to prepare a large dataset without being biased during the manually performed annotation process. The second challenge is related to automated methods, which depend on keywords that might have conflicting definitions, as previously illustrated in Section 4.

4.3. Challenges of Feature Sets

Despite significant efforts to develop identifying traits for hate speech, the subject remains open due to differing definitions of hate speech and new online hate vocabulary. There was no agreement in the literature on what constitutes hate speech [150]. For example, Davidson et al. [25] differentiated between profanity, insults, and hate speech, whereas Warner and Hirschberg [39] defined abuse as based on a person's intrinsic qualities (e.g., ethnicity, sexual orientation, or gender). As a result, it is impossible to say whether the objectionable text is hateful or not [8]. Therefore, hate speech is a culturally dependent term with a specific definition [41]. Furthermore, hate speech may be interpreted in various ways depending on the level of democracy or policies enacted in each country, which may or may not correspond to international standards to some extent.

Feature detection [151,152] and identifying hate speech at the level of each class label, rather than in terms of a binary classification (hate and no hate), are discussed in this study. A small number of studies have reported the performance of hate speech and related domains in relation to class labels [140]. Furthermore, datasets are frequently uneven, with hateful class occurrences far outnumbering non-hate textual examples, making micro-F1-scores appear high due to the majority class. Many machine learning approaches work well on specialized datasets, but they are incapable of generalizing new hate speech content [82].

Consequently, the issue is finding a set of features that is sufficient to generalize to new datasets. Additionally, the feature set should be minimal, providing high-performance models. However, to the best of our knowledge, little research has discussed a generic model that applies to any domain of interest or any dataset.

4.4. Future Research Directions

Based on the discussed literature, the following are future research directions in the domain of textual hate speech detection:

1. There is a critical dearth of reporting in the literature on the optimal set of features for hate speech detection that can be applied to both classical and deep learning models. Therefore, extensive research is needed to develop features that work well with diverse datasets with multifaceted hate speech concepts. A successful model should also have features that can be applied to new datasets and previously unseen tweets. A direction could be research [45,153] in which more features are added to develop additional features.
2. Aside from the basic hate/no hate categorization for traditional and deep learning models, the literature lacks a detailed investigation of fine-grained hate speech detection at the label level. According to the studies gathered, there is still a gap in creating a model that successfully performs the multi-classification of hate speech, has acceptable performance, and can be generalized across settings. A starting point could be using the models of [81], where several classes were adopted.
3. There are no recommendations in the literature to ensure that hate speech detection methods are adequately compared across different datasets. Therefore, a new methodology for dataset comparison is needed so that datasets can be rigorously compared.

5. Conclusions

New datasets of hate speech from different regions with various topics of hate speech and offensive content are constantly being developed. However, many datasets are small in size whilst others lack reliability due to how the data were collected and annotated. Additionally, many datasets are small or sparse, lacking linguistic variety. Above all, the language of the content and region where the data were collected from social media make the comparison between various hate speech detection models difficult. One of the significant challenges in hate speech detection is the architecture of the machine learning model and the lack of consensus on hate speech definition. It was reported that creating large and varied hate or abusive datasets that minimize potential bias is laborious and requires specialized experts. The critical review provided evidence that the research community should enhance the currently developed datasets and constantly update them. Although many machine learning models are developed, feature set selection should be fine-grained at the label level, as well as in the generic hate/no hate classification by developing the best set of features for hate speech detection, which is applicable to traditional and deep learning models. The literature lacks guidelines that assure a proper comparison between the methods over different datasets.

Author Contributions: All authors significantly contributed to the scientific study and writing. F.A. contributed to the overall idea, model formulation and analysis, and the writing of the manuscript; X.M. contributed to the process of refining research ideas. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Summary of a sample-related papers (most relevant papers).

Paper	Dataset	Best Method	Results	Limitation
[40]	Islamophobic hate speech data set (109,488 tweets)	One-versus-one SVM	0.77 accuracy	The dataset was for the UK context and the word context was not considered.
[25]	25,000 tweets	SVM	0.91 F1-measure	The one-versus-rest classifier is trained for each class, where the class label is assigned to the highest probability scores across classifiers.
[30]	5593 tweets	SVM	0.97 F1-score	The sexual-orientation hate class only obtained a 0.51 F1-score.
[26]	14 K tweets	J48 graft	0.78 F1-measure	Hate speech classes: clean, offensive, hateful(three classes of hate mixed with offensive hate).
[34]	Automatic Misogyny Identification (AMI) IberEval [35] (3251 tweets), AMI EvalIta [36] (4000 tweets), and the SRW [28] (5006 tweets)	LR	accuracy AMI IberEval: 0.7605AMI EvalIta: 0.7947SRW:0.8937	General sexist tweets hide a sentiment of hate or misogynistic attitude. Sexist jokes could contribute to making sexism or misogyny not generic to hate speech.
[28]	16 K tweets	LR	73.93 F1-score	Based on three classes—racism, sexism, and none—results were due to false positives for multi-class labels with an F1-score of 0.53 as compared to a binary classification of 0.73 F1-score.
[22]	EVALITA shared task 2018 (5000 tweets)	LR	0.704 accuracy	Misogyny classification has a low F1-score of 0.37.
[44]	10 K tweets (English) and 5 K tweets in Spanish	SVM	0.38/0.37 F1-measure of evaluation dataset (Task A, Task B [45]). Detection of hate speech (Task A), and identifying whether the objective of hatred is a person or a group of people (Task B).	Low-performance, approximately random, and shallow feature sets.

Table A1. *Cont.*

Paper	Dataset	Best Method	Results	Limitation
[46]	Semeval-2019 task 5: multilingual detection of hate speech against immigrants and women on Twitter(19,600 tweets—13,000 in English and 6600 in Spanish)	LIBSVM with RBF	TASK 1: hateful or not: 0.58 accuracy TASK 2: individual or generic: 0.81 accuracy TASK 3: aggressive or not: 0.80 accuracy	Focused on detection of hate speech against immigrants and women on Twitter (HatEval).
[116]	2228 sarcastic tweets	RF	0.83 accuracy	Most of the sarcastic tweets do not fall in the category of sarcasm where a positive sentiment contrasts with a negative situation. Some authors did not recognize sarcasm as hate speech.
[37]	CrowdFlower (Davidson et al.) [25] and the Forum for Information Retrieval Evaluation (FIRE) dataset. The FIRE task is a forum for Identifying Hate Speech and Offensive Content in Social Media Text (HASOC) [38]. CrowdFlower dataset (24,783, 9322). HASOC dataset (5852, 9292)	SVM with GLOVE	Accuracy HASOC Dataset:0.63 CrowdFlower Dataset:0.89	Binary classification classes are hate, offensive, and neither, not considering other types of hate speech.
[52]	MMHS150K dataset (150 K tweets)	LDA	0.704 F1-score	Despite using images in the dataset, it did not outperform textual models.
[65]	76 K tweets	MCD + LSTM	0.78 accuracy	The dataset was built into the following categories: sexual orientation, religion, nationality, gender, and ethnicity; however, the classifiers were trained on three classes: hateful, abusive, or neither.
[69]	6655 tweets	GRU + CNN	0.78 F1-score	The system identified racist and sexist tweets, but was not able to correctly identify the category 'both' since there are very few examples in this category.
[53,54]	120,000 tweets	Fuzzy ensemble	0.80 accuracy	Focused on detecting profiles rather than content.
[71]	12,311 tweets from COVID-19 dataset [72] 1105 tweets for US elections 4989 tweets from Waseem and Hovy	Multi-kernel convolution (MKC) of CNN	0.88 F1-score in US elections 0.83 in COVID-19 dataset 0.61 in Waseem and Hovy dataset	Focused on an election and COVID-19.
[140]	Davidson dataset [25] (24,783 tweets)	MCBiGRU	0.80–0.94 F1-score over different datasets. 0.94 in Davidson dataset of 24,783 tweets	One potential issue with pre-trained embeddings is out-of-vocabulary (OOV) words.
[115]	1235 tweets	CAT boost	0.94 F1-score	Binary classification. Small dataset.

Table A1. *Cont.*

Paper	Dataset	Best Method	Results	Limitation
[76]	13,240 tweets from OLID [101]	LDA	0.66 F1-score (Subtask A: offensive/not) 0.88 F1-score (Subtask B: categorization of offense types)	Sarcastic tweets make it difficult to discern the emotions (as per the author). Topic-wise, rather than the classification of hate speech content. Small dataset.
[59]	Dataset1: CrowdFlower (24,783, 9322)Dataset2: Waseem dataset [28] (16,093) Dataset 3: Davidson dataset [25] (24,783)	RETINA	Hate, offensive, neither) from Dataset 1, F1-score: 0.14, 0.67, 0.88	Sexism, racism, and neither labels had an F1-score of 0.04, 0, 0.92 in Dataset 3 as well as a low F1-score in Dataset 2.
[80]	SemEval-2019 Task 6 [95] dataset (14 K for subtask A: Offensive (OFF) and non-offensive(NOT))	MCD + LSTM	0.78 F1-score	Binary classification: offensive and non-offensive.
[114]	SemEval-2019 Task 6 [154]	GRU + CNN	Task A: classification of tweets into either offensive (OFF) or not offensive (NOT) 0.78 for supervised 0.77 for unsupervised approach	Binary classification: offensive and no offensive.
[117]	Davidson [25], Hateval [83], Waseem and Hovy [28], Waseem [27,81] Total of 121 annotated tweets out of 396 tweets	Cat Boost	F1-score ranging from 0.85 to 0.89 Best average F1-score 87.74 across all datasets	The classified hate is related to ethnic hate, racism, sexism, gender, and refugee hate. Similarly to HASOC Subtask 1 [38] and topic-relevant forum posts [155], where the topic of hate is detected rather than the type of hate speech.

Appendix B

Table A2. List of reviewed datasets.

No.	Dataset Name	Size (# of Tweets)	Categories of the Dataset	Ref
1	Waseem and Hovy	16,000	Racism, sexism, neither	[28]
2	Davidson et al.	24,783	Hate, offensive, neither	[25]
3	Waseem	6909	Racism, sexism, neither, both	[28]
4	SemEval Task 6 (OLID)	14,000 tweets	Level A: offensive, not offensive Level B: targeted insult, untargeted Level C: individual, group, other	[101]
5	SemEval Task 5 (HatEval)	19,600, 13,000 in English, 6600 in Spanish	Subtask A: hate, non-hate Subtask B: individual target, group target Subtask C: aggressive, non-aggressive	[83]
6	Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC)	5335 for the English dataset of HASOC 20207005 for HASOC 2019	Subtask A: hate and not offensive Subtask B: hate speech, offensive, and profanity	[19,38]
7	ElSherief et al.	25,278 hate instigators 22,857 targets 27,330 tweets	Archaic, class, disability, ethnicity, gender, nationality, religion, sexual orientation	[131]

Table A2. *Cont.*

No.	Dataset Name	Size (# of Tweets)	Categories of the Dataset	Ref
8	Founta et al.	80,000 (Size Doesn't Guarantee Diversity [137])	Offensive, abusive, hateful speech, aggressive, cyberbullying, spam, normal	[84]
	Ousidhoum et al.	5647 instances	Hateful, abusive or neither Directness ("direct/indirect"), hostility ("abusive/hateful/offensive/disrespectful/ fearful/normal"), target ("origin/gender/sexual orientation/religion/disability/ other"), group ("individual/woman/special needs/African descent/other") and the feeling aroused in the annotator by the tweet ("disgust/shock/anger/sadness/ fear/confusion/indifference")	[81]
9	MMHS150K	150 K tweets	Not hate, religion, sexist, racist, homophobic, other hate	[52]
10	ConaN	1288Pairs for English counter features.	Topics: crimes, culture, economics, generic, islamophobia, racism, terrorism, women	[156]
11	AbusEval	18,740	Offensive, targeted, not targeted, not offensive, explicitly abusive, implicitly abusive, not abusive	[103]
12	Amievalita	4000	misogynous, discredit, sexual harassment, stereotype, dominance, derailing	[36]
13	HateXplain	20,148	hate speech, offensive, normal the target community (i.e., the community that has been the victim of hate speech/offensive speech in the post), and the rationales, i.e., the portions of the post on which their labelling decision (as hate, offensive or normal)	[157]
14	Levantine Hate Speech and Abusive (L-HSAB)	5846	Hate, abusive, normal group or person target	[134]
15	News hate	1528 (Fox News)	Hate, not hate	[158]
16	Sexism	712	Benevolent sexism, hostile sexism, none	[158]
17	Women	3977	misogyny/not, stereotype, dominance, derailing, sexual harassment, discredit of misogyny, (active or passive) target	[35]
18	Hate	4972	Binary hate or not	[159]
19	Harassment	35,000	Harassment, not	[89]
20	Hate Topics	24,189	Topics: racism, sexism, appearance-related, intellectual, political	[159]

References

1. Poletto, F.; Basile, V.; Sanguinetti, M.; Bosco, C.; Patti, V. Resources and benchmark corpora for hate speech detection: A systematic review. *Lang. Resour. Eval.* **2021**, *55*, 477–523. [CrossRef]
2. Theodosiadou, O.; Pantelidou, K.; Bastas, N.; Chatzakou, D.; Tsikrika, T.; Vrochidis, S.; Kompatsiaris, I. Change point detection in terrorism-related online content using deep learning derived indicators. *Information* **2021**, *12*, 274. [CrossRef]
3. Sánchez-Compaña, M.T.; Sánchez-Cruzado, C.; García-Ruiz, C.R. An interdisciplinary scientific and mathematic education, addressing relevant social problems such as sexist hate speech. *Information* **2020**, *11*, 543. [CrossRef]
4. Mondal, M.; Silva, L.A.; Benevenuto, F. A measurement study of hate speech in social media. In Proceedings of the HT 2017—28th ACM Conference on Hypertext and Social Media, Prague, Czech Republic, 4–7 July 2017; pp. 85–94. [CrossRef]
5. Sanoussi, M.S.A.; Xiaohua, C.; Agordzo, G.K.; Guindo, M.L.; al Omari, A.M.M.A.; Issa, B.M. Detection of Hate Speech Texts Using Machine Learning Algorithm. In Proceedings of the 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 26–29 January 2022; pp. 266–273. [CrossRef]

6. Forestiero, A. Metaheuristic algorithm for anomaly detection in Internet of Things leveraging on a neural-driven multiagent system. *Knowl. Based Syst.* **2021**, *228*, 107241. [CrossRef]
7. Ayo, F.E.; Folorunso, O.; Ibharalu, F.T.; Osinuga, I.A. Machine learning techniques for hate speech classification of twitter data: State-of-The-Art, future challenges and research directions. *Comput. Sci. Rev.* **2020**, *38*, 100311. [CrossRef]
8. Strossen, N. Freedom of speech and equality: Do we have to choose. *JL Pol'y* **2016**, *25*, 185.
9. Comito, C.; Forestiero, A.; Pizzuti, C. Word embedding based clustering to detect topics in social media. In Proceedings of the 2019 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2019, Thessaloniki, Greece, 14–17 October 2019; pp. 192–199. [CrossRef]
10. MacAvaney, S.; Yao, H.R.; Yang, E.; Russell, K.; Goharian, N.; Frieder, O. Hate speech detection: Challenges and solutions. *PLoS ONE* **2019**, *14*, e0221152. [CrossRef]
11. Chetty, N.; Alathur, S. Hate speech review in the context of online social networks. *Aggress. Violent Behav.* **2018**, *40*, 108–118. [CrossRef]
12. Paz, M.A.; Montero-Díaz, J.; Moreno-Delgado, A. Hate Speech: A Systematized Review. *SAGE Open* **2020**, *10*, 3022. [CrossRef]
13. Matamoros-Fernández, A.; Farkas, J. Racism, Hate Speech, and Social Media: A Systematic Review and Critique. *Telev. New Media* **2021**, *2*, 205–224. [CrossRef]
14. Fortuna, P.; Bonavita, I.; Nunes, S. Merging datasets for hate speech classification in Italian. *CEUR Workshop Proc.* **2018**, *2263*. [CrossRef]
15. Tranfield, D.; Denyer, D.; Smart, P. Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *Br. J. Manag.* **2003**, *14*, 207–222. [CrossRef]
16. Snyder, H. Literature review as a research methodology: An overview and guidelines. *J. Bus. Res.* **2019**, *104*, 333–339. [CrossRef]
17. Guest, G.; MacQueen, K.M.; Namey, E.E. *Applied Thematic Analysis*; Sage Publications: Newbury Park, CA, USA, 2011.
18. Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; Kumar, R. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 July 2019; pp. 75–86.
19. Mandl, T.; Modha, S.; Kumar, A.; Chakravarthi, B.R. Overview of the HASOC track at FIRE 2020: Hate speech and offensive content identification in Indo-European languages. *CEUR Workshop Proc.* **2020**, *2826*, 87–111.
20. Wadhwa, P.; Bhatia, M.P.S. Classification of Radical Messages on Twitter Using Security Associations. In *Case Studies in Secure Computing: Achievements and Trends*; Auerbach Publications: New York, NY, USA, 2014; pp. 273–294.
21. Rangel, F.; Sarracén, G.L.D.L.P.; Chulvi, B.; Fersini, E.; Rosso, P. Profiling Hate Speech Spreaders on Twitter Task at PAN 2021. In Proceedings of the CLEF 2021–Conference and Labs of the Evaluation Forum, Bucharest, Romania, 21–24 September 2021.
22. Saha, P.; Mathew, B.; Goyal, P.; Mukherjee, A. Hateminers: Detecting Hate speech against Women. *arXiv* **2018**, arXiv:1812.06700.
23. Cer, D.; Yang, Y.; Kong, S.Y.; Hua, N.; Limtiaco, N.; John, R.S.; Kurzweil, R. Universal Sentence Encoder. *arXiv* **2018**, arXiv:1803.11175.
24. De Andrade, C.M.V.; Gonçalves, M.A. Profiling Hate Speech Spreaders on Twitter: Exploiting textual analysis of tweets and combinations of multiple textual representations. *CEUR Workshop Proc.* **2021**, *2936*, 2186–2192.
25. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated hate speech detection and the problem of offensive language. In Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017, Quebec, MO, Canada, 15–18 May 2017; pp. 512–515.
26. Watanabe, H.; Bouazizi, M.; Ohtsuki, T. Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. *IEEE Access* **2018**, *6*, 13825–13835. [CrossRef]
27. Waseem, Z. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In Proceedings of the First Workshop on NLP and Computational Social Science, Austin, TX, USA, 5 November 2016.
28. Waseem, Z.; Hovy, D. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In Proceedings of the NAACL Student Research Workshop, San Diego, CA, USA, 16–17 June 2016; pp. 88–93. [CrossRef]
29. Aziz, N.A.A.; Maarof, M.A.; Zainal, A. Hate Speech and Offensive Language Detection: A New Feature Set with Filter-Embedded Combining Feature Selection. In Proceedings of the 2021 3rd International Cyber Resilience Conference CRC 2021, online, 29–31 January 2021. [CrossRef]
30. Burnap, P.; Williams, M.L. Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Sci.* **2016**, *5*, 1–15. [CrossRef]
31. Ombui, E.; Muchemi, L.; Wagacha, P. Hate Speech Detection in Code-switched Text Messages. In Proceedings of the 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies ISMSIT 2019, Ankara, Turkey, 11–13 October 2019; pp. 1–6. [CrossRef]
32. Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; Chang, Y. Abusive language detection in online user content. In Proceedings of the 25th International Conference on World Wide Web WWW 2016, Montreal, Canada, 11–15 May 2016; pp. 145–153. [CrossRef]
33. Martins, R.; Gomes, M.; Almeida, J.J.; Novais, P.; Henriques, P. Hate speech classification in social media using emotional analysis. In Proceedings of the 2018 Brazilian Conference on Intelligent Systems BRACIS 2018, Sao Paolo, Brazil, 22–25 October 2018; pp. 61–66. [CrossRef]
34. Frenda, S.; Ghanem, B.; Montes-Y-Gómez, M.; Rosso, P. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *J. Intell. Fuzzy Syst.* **2019**, *36*, 4743–4752. [CrossRef]

35. Fersini, E.; Rosso, P.; Anzovino, M. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. *IberEval@ SEPLN* **2018**, *2150*, 214–228.
36. Fersini, E.; Nozza, D.; Rosso, P. Overview of the evalita 2018 task on automatic misogyny identification (ami). *EVALITA Eval. NLP Speech Tools Ital.* **2018**, *12*, 59.
37. Srivastava, N.D.; Sharma, Y. Combating Online Hate: A Comparative Study on Identification of Hate Speech and Offensive Content in Social Media Text. In Proceedings of the 2020 IEEE Recent Advances in Intelligent Computational Systems RAICS 2020, Thiruvananthapuram, India, 3–5 December 2020; pp. 47–52. [CrossRef]
38. Mandl, T.; Modha, S.; Majumder, P.; Patel, D.; Dave, M.; Mandlia, C.; Patel, A. Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages. In Proceedings of the FIRE '19: Proceedings of the 11th Forum for Information Retrieval Evaluation, Kolkata, India, 12–15 December 2019; Volume 2826, pp. 87–111.
39. Warner, W.; Hirschberg, J. Detecting Hate Speech on the World Wide Web. Available online: http://dl.acm.org/citation.cfm?id=2390374.2390377 (accessed on 23 May 2022).
40. Vidgen, B.; Yasseri, T. Detecting weak and strong Islamophobic hate speech on social media. *J. Inf. Technol. Polit.* **2020**, *17*, 66–78. [CrossRef]
41. Capozzi, A.T.E.; Lai, M.; Basile, V.; Poletto, F.; Sanguinetti, M.; Bosco, C.; Patti, V.; Ruffo, G.F.; Stranisci, M.A. Computational linguistics against hate: Hate speech detection and visualization on social media in the 'Contro L'Odio' project. *CEUR Workshop Proc.* **2019**, *2481*, 1–6.
42. Sanguinetti, M.; Poletto, F.; Bosco, C.; Patti, V.; Stranisci, M. An italian twitter corpus of hate speech against immigrants. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; pp. 2798–2805.
43. Basile, V. 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, Minneapolis, MN, USA, 6–7 June 2019; pp. 54–63.
44. Vega, L.E.A.; Reyes-Magaña, J.C.; Gómez-Adorno, H.; Bel-Enguix, G. MineriaUNAM at SemEval-2019 Task 5: Detecting Hate Speech in Twitter using Multiple Features in a Combinatorial Framework. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 447–452. [CrossRef]
45. Tellez, E.S.; Moctezuma, D.; Miranda-Jimnez, S.; Graff, M. An Automated Text Categorization Framework Based on Hyperparameter Optimization. *Know. Based Syst.* **2018**, *149*, 110–123. [CrossRef]
46. Bauwelinck, N.; Jacobs, G.; Hoste, V.; Lefever, E. LT3 at SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (hatEval). In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 436–440. [CrossRef]
47. Perelló, C.; Tomás, D.; Garcia-Garcia, A.; Garcia-Rodriguez, J.; Camacho-Collados, J. UA at SemEval-2019 Task 5: Setting A Strong Linear Baseline for Hate Speech Detection. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 508–513. [CrossRef]
48. I Orts, Ò.G. Multilingual detection of hate speech against immigrants and women in Twitter at SemEval-2019 task 5: Frequency analysis interpolation for hate in speech detection. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 460–463.
49. Ribeiro, A.; Silva, N. INF-HatEval at SemEval-2019 Task 5: Convolutional Neural Networks for Hate Speech Detection Against Women and Immigrants on Twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 420–425. [CrossRef]
50. Indurthi, V.; Syed, B.; Shrivastava, M.; Chakravartula, N.; Gupta, M.; Varma, V. FERMI at SemEval-2019 Task 5: Using Sentence embeddings to Identify Hate Speech Against Immigrants and Women in Twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 70–74. [CrossRef]
51. Chakrabarty, N. *A Machine Learning Approach to Comment Toxicity Classification*; Springer: Singapore, 2020; Volume 999. [CrossRef]
52. Gomez, R.; Gibert, J.; Gomez, L.; Karatzas, D. Exploring hate speech detection in multimodal publications. In Proceedings of the 2020 IEEE Winter Conference on Applications on Computer Vision (WACV), Snowmass, CO, USA, 1–5 March 2020; pp. 1459–1467. [CrossRef]
53. Siino, M.; di Nuovo, E.; Tinnirello, I.; la Cascia, M. Detection of Hate Speech Spreaders using convolutional neural networks. *CEUR Workshop Proc.* **2021**, *2936*, 2126–2136.
54. Balouchzahi, F.; Shashirekha, H.L.; Sidorov, G. HSSD: Hate speech spreader detection using N-Grams and voting classifier. *CEUR Workshop Proc.* **2021**, *2936*, 1829–1836.
55. Winter, K.; Kern, R. Know-Center at SemEval-2019 Task 5: Multilingual Hate Speech Detection on Twitter using CNNs. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 431–435. [CrossRef]
56. Kamble, S.; Joshi, A. Hate Speech Detection from Code-mixed Hindi-English Tweets Using Deep Learning Models. *arXiv* **2018**, arXiv:1811.05145.
57. Djuric, N.; Zhou, J.; Morris, R.; Grbovic, M.; Radosavljevic, V.; Bhamidipati, N. Hate speech detection with comment embeddings. In Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, 18–22 May 2015; pp. 29–30.
58. Rozental, A.; Biton, D. Amobee at SemEval-2019 Tasks 5 and 6: Multiple Choice CNN Over Contextual Embedding. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 377–381. [CrossRef]

59. Khan, M.U.S.; Abbas, A.; Rehman, A.; Nawaz, R. HateClassify: A Service Framework for Hate Speech Identification on Social Media. *IEEE Internet Comput.* **2021**, *25*, 40–49. [CrossRef]
60. Yin, W.; Schütze, H. Attentive convolution: Equipping cnns with rnn-style attention mechanisms. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 687–702. [CrossRef]
61. Fortuna, P.; Soler-Company, J.; Wanner, L. Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 24 May 2020; pp. 6786–6794.
62. Margffoy-Tuay, E.; Pérez, J.C.; Botero, E.; Arbeláez, P. Dynamic multimodal instance segmentation guided by natural language queries. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–16 September 2018; pp. 630–645.
63. Suryawanshi, S.; Chakravarthi, B.R.; Arcan, M.; Buitelaar, P. Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, Marseille, France, 16 May 2020; Available online: https://www.aclweb.org/anthology/2020.trac-1.6 (accessed on 2 April 2022).
64. Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Fitzpatrick, C.A.; Bull, P.; Lipstein, G.; Nelli, T.; Zhu, R.; et al. The Hateful Memes Challenge: Competition Report. *Proc. Mach. Learn. Res.* **2021**, *133*, 344–360.
65. Vashistha, N.; Zubiaga, A. Online multilingual hate speech detection: Experimenting with hindi and english social media. *Information* **2021**, *12*, 5. [CrossRef]
66. Park, J.H.; Fung, P. One-step and Two-step Classification for Abusive Language Detection on Twitter. *arXiv* **2017**, arXiv:1706.01206.
67. Zimmerman, S.; Fox, C.; Kruschwitz, U. Improving hate speech detection with deep learning ensembles. In Proceedings of the 11th International Conference on Language Resources and Evaluation, Marseille, France, 11–16 May 2020; pp. 2546–2553.
68. Poursepanj, H.; Weissbock, J.; Inkpen, D. Uottawa: System description for semeval 2013 task 2 sentiment analysis in twitter. In Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, GA, USA, 14–15 June 2013; pp. 380–383.
69. Gambäck, B.; Sikdar, U.K. Using Convolutional Neural Networks to Classify Hate-Speech. In Proceedings of the first workshop on abusive language online, Vancouver, BC, Canada, 4 August 2017; 7491, pp. 85–90. [CrossRef]
70. Qian, J.; ElSherief, M.; Belding, E.; Wang, W.Y. Hierarchical CVAE for fine-grained hate speech classification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 3550–3559. [CrossRef]
71. Agarwal, S.; Chowdary, C.R. Combating hate speech using an adaptive ensemble learning model with a case study on COVID-19. *Expert Syst. Appl.* **2021**, *185*, 115632. [CrossRef]
72. Ziems, C.; He, B.; Soni, S.; Kumar, S. Racism is a Virus: Anti-Asian Hate and Counterhate in Social Media during the COVID-19 Crisis. *arXiv* **2020**, arXiv:2005.12423. Available online: https://europepmc.org/article/PPR/PPR268779 (accessed on 2 April 2022).
73. Agarwal, S.; Chowdary, C.R. A-Stacking and A-Bagging: Adaptive versions of ensemble learning algorithms for spoof fingerprint detection. *Expert Syst. Appl.* **2020**, *146*, 3160. [CrossRef]
74. Mehdad, Y.; Tetreault, J. Do Characters Abuse More Than Words? In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Los Angeles, CA, USA, 13–15 September 2016; pp. 299–303. [CrossRef]
75. Malmasi, S.; Zampieri, M. Challenges in discriminating profanity from hate speech. *J. Exp. Theor. Artif. Intell.* **2018**, *30*, 187–202. [CrossRef]
76. Doostmohammadi, E.; Sameti, H.; Saffar, A. Ghmerti at SemEval-2019 Task 6: A Deep Word- and Character-based Approach to Offensive Language Identification. *arXiv* **2019**, arXiv:2009.10792.
77. Garain, A.; Basu, A. The Titans at SemEval-2019 Task 6: Offensive Language Identification, Categorization and Target Identification. 2019, 759–762. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 759–762. [CrossRef]
78. Mishra, A.K.; Saumya, S.; Kumar, A. IIIT_DWD@HASOC 2020: Identifying offensive content in Indo-European languages. *CEUR Workshop Proc.* **2020**, *2826*, 139–144.
79. Mohtaj, S.; Woloszyn, V.; Möller, S. TUB at HASOC 2020: Character based LSTM for hate speech detection in Indo-European languages. *CEUR Workshop Proc.* **2020**, *2826*, 298–303.
80. Modha, S.; Majumder, P.; Patel, D. DA-LD-Hildesheim at SemEval-2019 Task 6: Tracking Offensive Content with Deep Learning using Shallow Representation. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 577–581. [CrossRef]
81. Ousidhoum, N.; Lin, Z.; Zhang, H.; Song, Y.; Yeung, D.Y. Multilingual and multi-aspect hate speech analysis. In Proceedings of the 9th International Joint Conference on Natural Language Processing Conference, Hong Kong, China, 3–7 November 2019; pp. 4675–4684. [CrossRef]
82. Wullach, T.; Adler, A.; Minkov, E. Towards Hate Speech Detection at Large via Deep Generative Modeling. *IEEE Internet Comput.* **2021**, *25*, 48–57. [CrossRef]
83. Yang, X.; Obadinma, S.; Zhao, H.; Zhang, Q.; Matwin, S.; Zhu, X. SemEval-2020 Task 5: Counterfactual Recognition. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona, Spain, 12–13 December 2020; pp. 322–335.

84. Founta, A.M. Large scale crowdsourcing and characterization of twitter abusive behavior. In Proceedings of the 12th International AAI Conference on Web and Social Media, ICWSM 2018, Palo Alto, CA, USA, 25–28 June 2018; pp. 491–500.
85. De Gibert, O.; Perez, N.; García-Pablos, A.; Cuadros, M. Hate Speech Dataset from a White Supremacy Forum. In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Brussels, Belgium, 31 October–4 November 2019; pp. 11–20. [CrossRef]
86. Radford, I.S.A.; Wu, J.; Child, R.; Luan, D.; Amodei, D. [GPT-2] Language Models are Unsupervised Multitask Learners. *OpenAI Blog* **2020**, *1*, 9. Available online: https://github.com/codelucas/newspaper (accessed on 3 April 2022).
87. Ziqi, Z.; Robinson, D.; Jonathan, T. Hate Speech Detection Using a Convolution-LSTM Based Deep Neural Network. *IJCCS* **2019**, *11816*, 2546–2553. [CrossRef]
88. Naseem, U.; Razzak, I.; Hameed, I.A. Deep Context-Aware Embedding for Abusive and Hate Speech detection on Twitter. *J. Chem. Inf. Model.* **2019**, *53*, 1689–1699.
89. Golbeck, J. A large human-labeled corpus for online harassment research. In Proceedings of the 2017 ACM Web Science Conference, Troy, NY, USA, 25–28 June 2017; pp. 229–233. [CrossRef]
90. Founta, A.M.; Chatzakou, D.; Kourtellis, N.; Blackburn, J.; Vakali, A.; Leontiadis, I. A unified deep learning architecture for abuse detection. In Proceedings of the 10th ACM Conference on Web Science, Boston, MA, USA, 30 June–3 July 2019; pp. 105–114. [CrossRef]
91. Chatzakou, D.; Kourtellis, N.; Blackburn, J.; de Cristofaro, E.; Stringhini, G.; Vakali, A. Mean birds: Detecting aggression and bullying on Twitter. In Proceedings of the 2017 ACM Web Science Conference, Troy, NY, USA, 25–28 June 2017; pp. 13–22. [CrossRef]
92. Rajadesingan, A.; Zafarani, R.; Liu, H. Sarcasm detection on twitter: A behavioral modeling approach. In Proceedings of the Eigth ACM International Conference on Web Search and Data Mining, Shanghai, China, 2–6 February 2015; pp. 97–106. [CrossRef]
93. Menini, S.; Moretti, G.; Corazza, M.; Cabrio, E.; Tonelli, S.; Villata, S. A System to Monitor Cyberbullying based on Message Classification and Social Network Analysis. In Proceedings of the Third Workshop on Abusive Language Online, Florence, Italy, 1 August 2019; pp. 105–110. [CrossRef]
94. Corazza, M.; Menini, S.; Cabrio, E.; Tonelli, S.; Villata, S. A Multilingual Evaluation for Online Hate Speech Detection. *ACM Trans. Internet Technol.* **2020**, *20*. [CrossRef]
95. Zhu, R. Enhance Multimodal Transformer with External Label and In-Domain Pretrain: Hateful Meme Challenge Winning Solution. *arXiv* **2020**, arXiv:2012.0829.
96. Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; Dai, J. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv* **2019**, arXiv:1908.08530.
97. Yu, F.; Tang, J.; Yin, W.; Sun, Y.; Tian, H.; Wu, H.; Wang, H. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv* **2020**, arXiv:2006.16934.
98. Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; Testuggine, D. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 2611–2624.
99. Lee, R.K.-W.; Cao, R.; Fan, Z.; Jiang, J.; Chong, W.-H. *Disentangling Hate in Online Memes*; Association for Computing Machinery: New York, NY, USA, 2021; Volume 1. [CrossRef]
100. Liu, P.; Li, W.; Zou, L. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In Proceedings of the NAACL HLT 2019—International Workshop on Semantic Evaluation, SemEval 2019, Proceedings of the 13th Workshop, Minneapolis, MN, USA, 6–7 June 2019; pp. 87–91. [CrossRef]
101. Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; Kumar, R. Predicting the type and target of offensive posts in social media. In Proceedings of the NAACL HLT 2019—2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 1415–1420. [CrossRef]
102. Caselli, T.; Basile, V.; Mitrović, J.; Granitzer, M. HateBERT: Retraining BERT for Abusive Language Detection in English. *arXiv* **2021**, arXiv:2010.12472.
103. Caselli, T.; Basile, V.; Mitrovic, J.; Kartoziya, I.; Granitzer, M. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 6193–6202.
104. Nguyen, D.Q.; Vu, T.; Nguyen, A.T. BERTweet: A pre-trained language model for English Tweets. *arXiv* **2020**, arXiv:2005.10200.
105. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
106. Conneau, A.; Baevski, A.; Collobert, R.; Mohamed, A.; Auli, M. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv* **2020**, arXiv:2006.13979.
107. Jahan, M.S.; Oussalah, M. A systematic review of Hate Speech automatic detection using Natural Language Processing. *arXiv* **2021**, arXiv:2106.00742.
108. Pitsilis, G.K.; Ramampiaro, H.; Langseth, H. Effective hate-speech detection in Twitter data using recurrent neural networks. *Appl. Intell.* **2018**, *48*, 4730–4742. [CrossRef]
109. Badjatiya, P.; Gupta, S.; Gupta, M.; Varma, V. Deep Learning for Hate Speech Detection in Tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, 3–7 April 2017; Volume 2, pp. 759–760. [CrossRef]

110. Paschalides, D.; Stephanidis, D.; Andreou, A.; Orphanou, K.; Pallis, G.; Dikaiakos, M.D.; Markatos, E. MANDOLA: A big-data processing and visualization platform for monitoring and detecting online hate speech. *ACM Trans. Internet Technol.* **2020**, *20*, 1–21. [CrossRef]
111. Masud, S.; Duuta, S.; Makkar, S.; Jain, C.; Goyal, V.; Das, A.; Chakraborty, T. Hate is the new infodemic: A topic-aware modeling of hate speech diffusion on twitter. *Proc. Int. Conf. Data Eng.* **2021**, *2021*, 504–515. [CrossRef]
112. Kumar, A.; Abirami, S.; Trueman, T.E.; Cambria, E. Comment toxicity detection via a multichannel convolutional bidirectional gated recurrent unit. *Neurocomputing* **2021**, *441*, 272–278. [CrossRef]
113. Wang, B.; Ding, H. YNU NLP at SemEval-2019 task 5: Attention and capsule ensemble for identifying hate speech. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 529–534. [CrossRef]
114. Wiedemann, G.; Ruppert, E.; Biemann, C. UHH-LT at SemEval-2019 Task 6: Supervised vs. Unsupervised Transfer Learning for Offensive Language Detection. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 782–787. [CrossRef]
115. Setyadi, N.A.; Nasrun, M.; Setianingsih, C. Text Analysis for Hate Speech Detection Using Backpropagation Neural Network. In Proceedings of the 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), Bandung, Indonesia, 5–7 December 2018; pp. 159–165. [CrossRef]
116. Bouazizi, M.; Otsuki, T. A Pattern-Based Approach for Sarcasm Detection on Twitter. *IEEE Access* **2016**, *4*, 5477–5488. [CrossRef]
117. Qureshi, K.A.; Sabih, M. Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text. *IEEE Access* **2021**, *9*, 109465–109477. [CrossRef]
118. Kshirsagar, R.; Cukuvac, T.; McKeown, K.; McGregor, S. Predictive Embeddings for Hate Speech Detection on Twitter. *arXiv* **2019**, arXiv:1809.10644.
119. Shen, D.; Shen, D.; Wang, G.; Wang, W.; Min, M.R.; Su, Q.; Zhang, Y.; Henao, R.; Carin, L. On the use of word embeddings alone to represent natural language sequences. In Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
120. Faris, H.; Aljarah, I.; Habib, M.; Castillo, P.A. Hate speech detection using word embedding and deep learning in the Arabic language context. In Proceedings of the ICPRAM 2020—9th International Conference on Pattern Recognition Applications and Methods, Valletta, Malta, 22–24 February 2020; pp. 453–460. [CrossRef]
121. Siddiqua, U.A.; Chy, A.N.; Aono, M. KDEHatEval at SemEval-2019 Task 5: A Neural Network Model for Detecting Hate Speech in Twitter. roceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 365–370. [CrossRef]
122. Miok, K.; Nguyen-Doan, D.; Škrlj, B.; Zaharie, D.; Robnik-Šikonja, M. Prediction Uncertainty Estimation for Hate Speech Classification. *Lect. Notes Comput. Sci.* **2019**, *11816*, 286–298. [CrossRef]
123. Sachdeva, J.; Chaudhary, K.K.; Madaan, H.; Meel, P. Text Based Hate-Speech Analysis. In Proceedings of the International Conference on Artificial Intelligence and Smart Systems, ICAIS, Tamilnadu, India, 25–27 March 2021; pp. 661–668. [CrossRef]
124. Sajjad, M.; Zulifqar, F.; Khan, M.U.G.; Azeem, M. Hate Speech Detection using Fusion Approach. In Proceedings of the 2019 International Conference on Applied and Engineering Mathematics, Taxila, Pakistan, 27–29 August 2019; pp. 251–255. [CrossRef]
125. Liu, H.; Alorainy, W.; Burnap, P.; Williams, M.L. Fuzzy multi-task learning for hate speech type identification. In Proceedings of the Web Conf. 2019—Proc. World Wide Web Conference, New York, UK, USA, 13–17 May 2019; pp. 3006–3012. [CrossRef]
126. Berthold, M.R. Mixed fuzzy rule formation. *Int. J. Approx. Reason.* **2003**, *32*, 67–84. [CrossRef]
127. Mulki, H.; Ali, C.B.; Haddad, H.; Babaoğlu, I. Tw-StAR at SemEval-2019 task 5: N-gram embeddings for hate speech detection in multilingual tweets. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 503–507. [CrossRef]
128. Yang, Z.; Yang, D.; Dyer1, C.; He, X.; Smola, A.; Hovy, E. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American chapter of the association for computational linguistics: Human language technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489.
129. Kocoń, J.; Figas, A.; Gruza, M.; Puchalska, D.; Kajdanowicz, T.; Kazienko, P. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Inf. Process. Manag.* **2021**, *58*, 102643. [CrossRef]
130. Wiegand, M.; Ruppenhofer, J.; Eder, E. Implicitly Abusive Language—What does it actually look like and why are we not getting there? In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 576–587. [CrossRef]
131. ElSherief, M.; Nilizadeh, S.; Nguyen, D.; Vigna, G.; Belding, E. Peer to peer hate: Hate speech instigators and their targets. In Proceedings of the 12th International AAAI Conference on Web and Social Media, ICWSM 2018, Pao Alto, CA, USA, 25–28 June 2018; pp. 52–61.
132. Guest, E.; Vidgen, B.; Mittos, A.; Sastry, N.; Tyson, G.; Margetts, H. An expert annotated dataset for the detection of online misogyny. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Kvyv, Ukraine, 21–23 April 2021; pp. 1336–1350. [CrossRef]
133. Qian, J.; Bethke, A.; Liu, Y.; Belding, E.; Wang, W.Y. A benchmark dataset for learning to intervene in online hate speech. In Proceedings of the EMNLP-IJCNLP 2019—2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2020; pp. 4755–4764. [CrossRef]

134. Mulki, H.; Haddad, H.; Ali, C.B.; Alshabani, H. L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Kvyv, Ukraine, 19–23 April 2019; pp. 1336–1350. [CrossRef]
135. Culpeper, J. Impoliteness and hate speech: Compare and contrast. *J. Pragmat.* **2021**, *179*, 4–11. [CrossRef]
136. Waseem, Z.; Davidson, T.; Warmsley, D.; Weber, I. Understanding abuse: A typology of abusive language detection subtasks. *arXiv* **2017**, arXiv:1705.09899.
137. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Association for Computing Machinery: New York, NY, USA, 2021; Volume 1. [CrossRef]
138. Plaza-del-Arco, F.M.; Molina-González, M.D.; Martin, M.; Ureña-López, L.A. SINAI at SemEval-2019 Task 5: Ensemble learning to detect hate speech against inmigrants and women in English and Spanish tweets. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 476–479. [CrossRef]
139. Mitrović, J.; Birkeneder, B.; Granitzer, M. nlpUP at SemEval-2019 Task 6: A Deep Neural Language Model for Offensive Language Detection. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 722–726. [CrossRef]
140. Zhang, Z.; Luo, L. Hate speech detection: A solved problem? The challenging case of long tail on Twitter. *Semant. Web* **2019**, *10*, 925–945. [CrossRef]
141. Dahiya, S. Would Your Tweet Invoke Hate on the Fly? Forecasting Hate Intensity of Reply Threads on Twitter. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Singapore, 14–18 August 2021; Volume 1, pp. 2732–2742. [CrossRef]
142. Kapil, P.; Ekbal, A. A deep neural network based multi-task learning approach to hate speech detection. *Knowl. Based Syst.* **2020**, *210*, 106458. [CrossRef]
143. Anand, M.; Eswari, R. Classification of abusive comments in social media using deep learning. In Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019), Erode, India, 27–29 March 2019; pp. 974–977. [CrossRef]
144. Tontodimamma, A.; Nissi, E.; Sarra, A.; Fontanella, L. Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics* **2021**, *126*, 157–179. [CrossRef]
145. Nugroho, K. Improving random forest method to detect hatespeech and offensive word. In Proceedings of the 2019 International Conference on Information and Communications Technology, Baku, Azerbaijan, 23–25 October 2019; pp. 514–518. [CrossRef]
146. Lingiardi, V.; Carone, N.; Semeraro, G.; Musto, C.; D'Amico, M.; Brena, S. Mapping Twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis. *Behav. Inf. Technol.* **2020**, *39*, 711–721. [CrossRef]
147. Shibly, F.H.A.; Sharma, U.; Naleer, H.M.M. *Classifying and Measuring Hate Speech in Twitter Using Topic Classifier of Sentiment Analysis*; Springer: Singapore, 2021; Volume 1165. [CrossRef]
148. ElSherief, M.; Kulkarni, V.; Nguyen, D.; Wang, W.Y.; Belding, E. Hate lingo: A target-based linguistic analysis of hate speech in social media. In Proceedings of the International AAAI Conference on Web and Social Media, ICWSM 2018, Pao Alto, CA, USA, 25–28 June 2018; pp. 42–51.
149. Abburi, H.; Sehgal, S.; Maheshwari, H. *Knowledge-Based Neural Framework for Sexism Detection and Classification*; IIIT: Hyderabad, India, 2021.
150. Fino, A. Defining Hate Speech. *J. Int. Crim. Justice* **2020**, *18*, 31–57. [CrossRef]
151. Ullmann, S.; Tomalin, M. Quarantining online hate speech: Technical and ethical perspectives. *Ethics Inf. Technol.* **2020**, *22*, 69–80. [CrossRef]
152. Mosca, E.; Wich, M.; Groh, G. Understanding and Interpreting the Impact of User Context in Hate Speech Detection. In Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, Online, 10 June 2021; pp. 91–102. [CrossRef]
153. Alizadeh, M.; Weber, I.; Cioffi-Revilla, C.; Fortunato, S.; Macy, M. Psychology and morality of political extremists: Evidence from Twitter language analysis of alt-right and Antifa. *EPJ Data Sci.* **2019**, *8*, 9. [CrossRef]
154. Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; Kumar, R. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv* **2019**, arXiv:1903.08983.
155. Yang, B.; Tang, H.; Hao, L.; Rose, J.R. Untangling chaos in discussion forums: A temporal analysis of topic-relevant forum posts in MOOCs. *Comput. Educ.* **2022**, *178*, 104402. [CrossRef]
156. Chung, Y.L.; Kuzmenko, E.; Tekiroglu, S.S.; Guarini, M. ConaN—Counter narratives through nichesourcing: A multilingual dataset of responses to fight online hate speech. In Proceedings of Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 June–2 August 2019; pp. 2819–2829.
157. Mathew, B.; Saha, P.; Yimam, S.M.; Biemann, C.; Goyal, P.; Mukherjee, A. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 14867–14875.
158. Gao, L.; Huang, R. Detecting Online Hate Speech Using Context Aware Models. In Proceedings of the International Conference Recent Advances in Natural Language Processing, {RANLP} 2017, Varna, Bulgaria, 2–8 September 2017; pp. 260–266. [CrossRef]
159. Ribeiro, M.H.; Calais, P.H.; Santos, Y.A.; Almeida, V.A.F.; Meira, W. Characterizing and detecting hateful users on twitter. *Twelfth Int. AAAI Conf. Web Soc. Media* **2018**, *12*, 676–679.