big data and cognitive computing

MDPI

Article

Intelligent Multi-Lingual Cyber-Hate Detection in Online Social Networks: Taxonomy, Approaches, Datasets, and Open Challenges

Donia Gamal [1,*], Marco Alfonse [1,2], Salud María Jiménez-Zafra [3] and Mostafa Aref [1]

[1] Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo 11566, Egypt; marco_alfonse@cis.asu.edu.eg (M.A.); mostafa.aref@cis.asu.edu.eg (M.A.)
[2] Laboratoie Interdisciplinaire de l'Université Française d'Égypte (UFEID LAB), Université Française d'Égypte, Cairo 11566, Egypt
[3] Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071 Jaén, Spain; sjzafra@ujaen.es
* Correspondence: donia.gamaleldin@cis.asu.edu.eg

Abstract: Sentiment Analysis, also known as opinion mining, is the area of Natural Language Processing that aims to extract human perceptions, thoughts, and beliefs from unstructured textual content. It has become a useful, attractive, and challenging research area concerning the emergence and rise of social media and the mass volume of individuals' reviews, comments, and feedback. One of the major problems, apparent and evident in social media, is the toxic online textual content. People from diverse cultural backgrounds and beliefs access Internet sites, concealing and disguising their identity under a cloud of anonymity. Due to users' freedom and anonymity, as well as a lack of regulation governed by social media, cyber toxicity and bullying speech are major issues that need an automated system to be detected and prevented. There is diverse research in different languages and approaches in this area, but the lack of a comprehensive study to investigate them from all aspects is tangible. In this manuscript, a comprehensive multi-lingual and systematic review of cyber-hate sentiment analysis is presented. It states the definition, properties, and taxonomy of cyberbullying and how often each type occurs. In addition, it presents the most recent popular cyberbullying benchmark datasets in different languages, showing their number of classes (Binary/Multiple), discussing the applied algorithms, and how they were evaluated. It also provides the challenges, solutions, as well as future directions.

Keywords: cyber-hate; cyberbullying; sentiment analysis; online social networks; machine learning

Citation: Gamal, D.; Alfonse, M.; Jiménez-Zafra, S.M.; Aref, M. Intelligent Multi-Lingual Cyber-Hate Detection in Online Social Networks: Taxonomy, Approaches, Datasets, and Open Challenges. *Big Data Cogn. Comput.* **2023**, *7*, 58. https://doi.org/10.3390/bdcc7020058

Academic Editors: Maria Chiara Caschera, Patrizia Grifoni and Fernando Ferri

Received: 17 February 2023
Revised: 16 March 2023
Accepted: 22 March 2023
Published: 24 March 2023

1. Introduction

Sentiment Analysis (SA) is the area of Natural Language Processing (NLP) that focuses on analyzing and studying individuals' sentiments, appraisals, evaluations, emotions, and attitudes writing in texts [1]. The utilization of social media platforms, for example, Twitter, Facebook, and Instagram, have immensely increased the quantity of online social interactions and communications by connecting billions of people who prefer the exchange of opinions. The penetration of social media into the life of internet users is increasing. According to the most recent data, there will be 5.85 billion social media users globally in 2027, a 1.26 percent rise over the previous year [2,3], as shown in Figure 1.

Moreover, social media platforms offer visibility to ideas and thoughts that would somehow be neglected and unspoken by traditional media [4]. The textual content of interactions and communications that signify upsetting, disturbing, and negative phenomena such as online cyber-hate, harassment, cyberbullying, stalking, and cyber threats is increasing [5]. Therefore, this has strongly led to an expansion of attacks against certain users based on different categorizations such as religion, ethnicity, social status, age, etc. Individuals frequently struggle and battle to deal with the results and consequences of

such offenses. By employing NLP, several attempts have been put in action to deal with the issue of online cyber-hate and cyberbullying speech detection. This is because the computational analysis of language could be utilized to rapidly identify and distinguish offenses to facilitate and ease the process of dealing with and removing harsh messages [6].

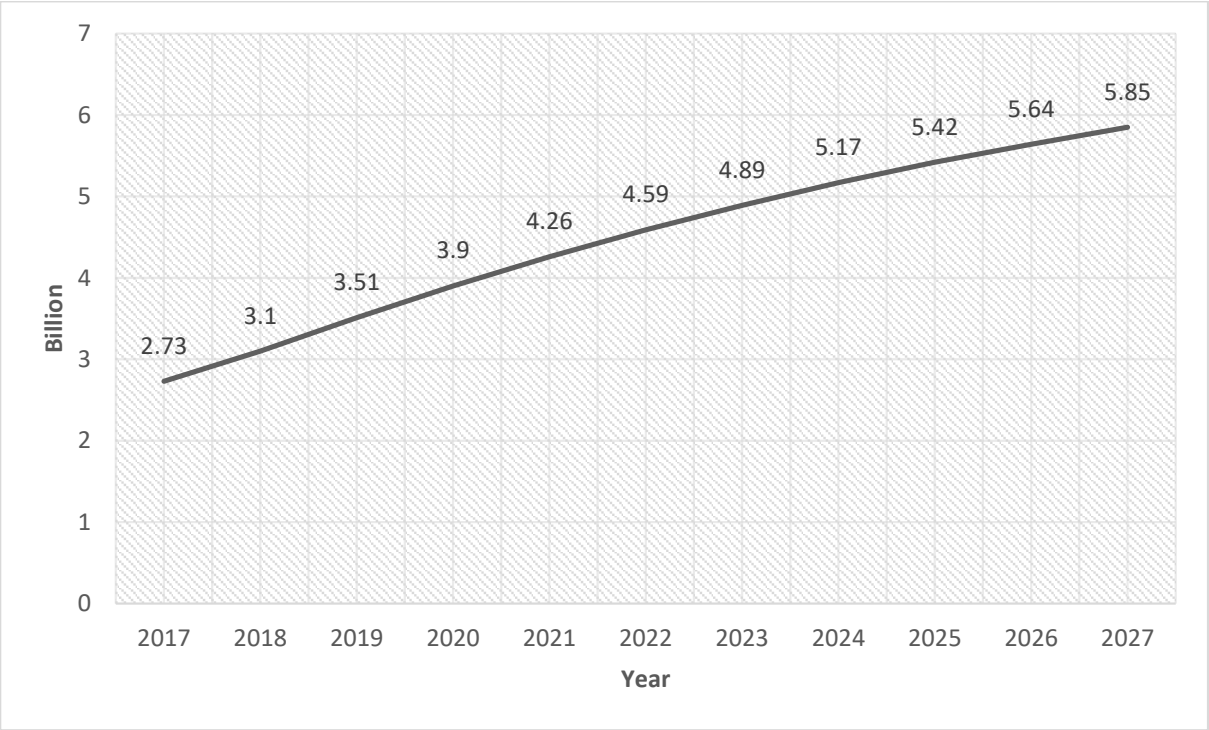

Figure 1. Number of Global Social Media Users Within 2017–2027.

The instance of hate speech and brutal communication shown over the Internet is named cyber-hate [7]. Cyber-hate, also known as cyberbullying, is defined as any utilization of electronic communications technologies to spread supremacist, racist, religious, extremist, or terrorist messages. It can target not only individuals but also entire communities [8]. Cyberbullying occurs when someone utilizes the internet to hurt or disturb a child or young person. It can occur on a social media platform, a game, an app, or any other online or electronic service or platform. Examples include posts, comments, texts, messages, chats, livestreams, memes, photos, videos, and emails. Some examples of how the internet can be used to violate someone's self-confidence is:

- Sending derogatory messages about them.
- Sharing humiliating images or videos of them.
- Spreading slanderous web rumors against them.
- Making fake accounts in their name.
- Making the public believe they are someone else.

Various types of violations are committed for many reasons in the online cyber realm through cyber innovation and technology. This insecure environment of online social networks requires consideration to prevent the harm and damage brought by these crimes to society. Several researchers are working in multiple directions to achieve the best results for automated cyberbullying detection using machine-learning techniques. In this manuscript, a taxonomy of multiple techniques being utilized in cyber-hate detection and prediction through different languages will be presented.

The rest of the manuscript is organized as follows: Section 2 provides a summary of the main properties of cyber-hate speech, and a detailed taxonomy of cyber-hate speech is given. The available datasets of cyber-hate in different languages are discussed in Section 3.

Section 4 previews the main approaches of cyber-hate classification. Section 5 introduces the comparative study of binary and multiple classifications over different datasets and various languages. Open Challenges in cyber-hate detection are discussed in Section 6. Finally, Section 7 concludes this manuscript and presents future work.

2. Cyber-Hate Speech Properties

2.1. Definition

Cyber-hate is the act of threatening, intimidating, harassing, irritating, or bullying any individual or group (for example, non-white people) through communication technology such as social media [9]. For textual content to be considered bullying, the intent of harm, such as physical, emotional, social, etc., should be obvious, as shown in Figure 2.The following situations are examples of cyber-hate speech:

- Posting threats such as physical harm, brutality, or violence.
- Any discussion intended to offend an individual's feelings, including routinely inappropriately teasing, prodding, or making somebody the brunt of pranks, tricks, or practical jokes.
- Any textual content meant to destroy the social standing or reputation of any individual on online social networks or offline communities.
- Circulating inappropriate, humiliating, or embarrassing images or videos on social networks.
- Persistent, grievous, or egregious utilization of abusive, annoying, insulting, hostile, or offensive language.

Figure 2. Examples of Online Social Media Cyberbullying.

2.2. Taxonomy

There is much more to cyber-hate than meets the eye. For instance, many people once believed that cyber-hate only consisted of physical bullying and name-calling. However, there are ten types of cyber-hate, which range from excluding and gossiping about people to making fun of their race or religion, as shown in Figure 3.

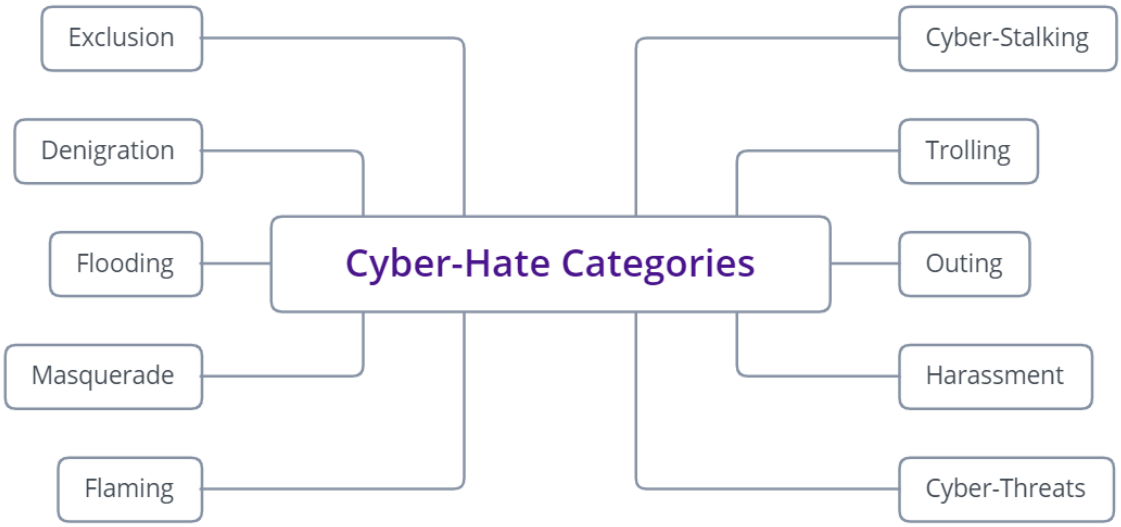

Figure 3. Cyber-hate Categories.

The categories that comprise the taxonomy of the term cyber-hate are presented and defined below:

i. Exclusion is defined as ignoring or neglecting the victim in a conversation [10]. Cyber-Exclusion is an intentional and deliberate action to make it clear to people that they do not belong to the group and that their involvement is not needed. On social networking sites, individuals can defriend or block others, which implies their inability to view their profiles, write comments, and so forth.

ii. Denigration involves the practice of demeaning, gossiping, dissing, or disrespecting another individual on social networks [11]. Writing rude, vulgar, mean, hurtful, or untrue messages or rumors about someone to another person or posting them in a public community or chat room falls under denigration. The purpose is to hurt the victim in the eyes of his or her community, as the insults are seen not only by the victim but also others.

iii. Flooding is the posting of a countless number of online social networking messages so the victim cannot post a message [12]. It consists of the bully or harasser repeatedly writing the same comment, posting nonsense comments, or holding down the enter key to not allow the victim to contribute to the chat or conversation.

iv. Masquerade is defined as the process of impersonating another person to send messages that seem to be originated by that person and cause damage or harm [13]. One of the ways to do this is by, for example, hacking into a victim's e-mail account and instantly sending these messages. Moreover, friends sharing passwords can also regularly accomplish this type of access; however, the sophisticated hacker may discover other ways, for example, by systematically testing probable passwords. This strategy is inherently hard and difficult to be recognized or detected.

v. Flaming, blazing, or bashing involves at least two users attacking and assaulting each other on a personal level. In this category of cyber-hate, flaming refers to a conversation full of hostile, unfriendly, irate, angry, and insulting communications and interactions that are regularly unkind personal attacks [14]. Flaming can occur in a diversity of environments, such as online social networking and discussion boards, group chat rooms, e-mails, and Twitter. Anger is frequently expressed by utilizing capital letters, such as 'U R AN IDIOT & I HATE U!'. Many flaming texts are vicious, horrible, and cruel and are without fact or reason.

vi. Cyberstalking is the utilization of social networks to stalk, hassle, or harass any individual, group, or organization [15]. It might contain false incriminations, accusations, criticism, defamation, maligning, slander, and libel. Cases of cyberstalking can often begin as seemingly harmless interactions. Sometimes, particularly at the

beginning, a few strange or maybe distasteful messages may even amuse. Nevertheless, if they turn out to be systematic, it becomes irritating, annoying, and even frightening.

vii. Trolling, also called baiting, attempts to provoke a fight by intentionally writing comments that disagree with other posts in the topic or thread [16]. The poster plans to excite emotions and rouse an argument, while the comments themselves inevitably turn personal, vulgar, enthusiastic, or emotional.

viii. Outing is similar to denigration but requires the bully and the victim to have a close personal relationship, either on social networks or in-person. It includes writing and sharing personal, private, embarrassing, or humiliating information publicly [17]. This information can incorporate stories heard or received from the victim or any personal information such as personal numbers, passwords, or addresses.

ix. Harassment using social networks is equivalent to harassment utilizing more conventional and traditional means [18]. Harassment refers to threatening actions dependent on an individual's age, gender, race, sexual orientation, and so forth.

x. Cyber threats include sending short messages that involve threats of harm, are scary, intimidating, are very aggressive, or incorporate extortion [19]. The dividing line where harassment becomes cyberstalking is obscured; however, one indicator may be when the victim starts to fear for his or her well-being or safety, then the act has to be considered cyberstalking.

According to a more detailed survey from 2021 by PEW Research Center (https://www.pewresearch.org/) (access on 16 February 2023), over 40% of Americans under the age of 30 have experienced online bullying [20]. The most common types of cyberbullying are, as represented in Figure 4, denigration and harassment.

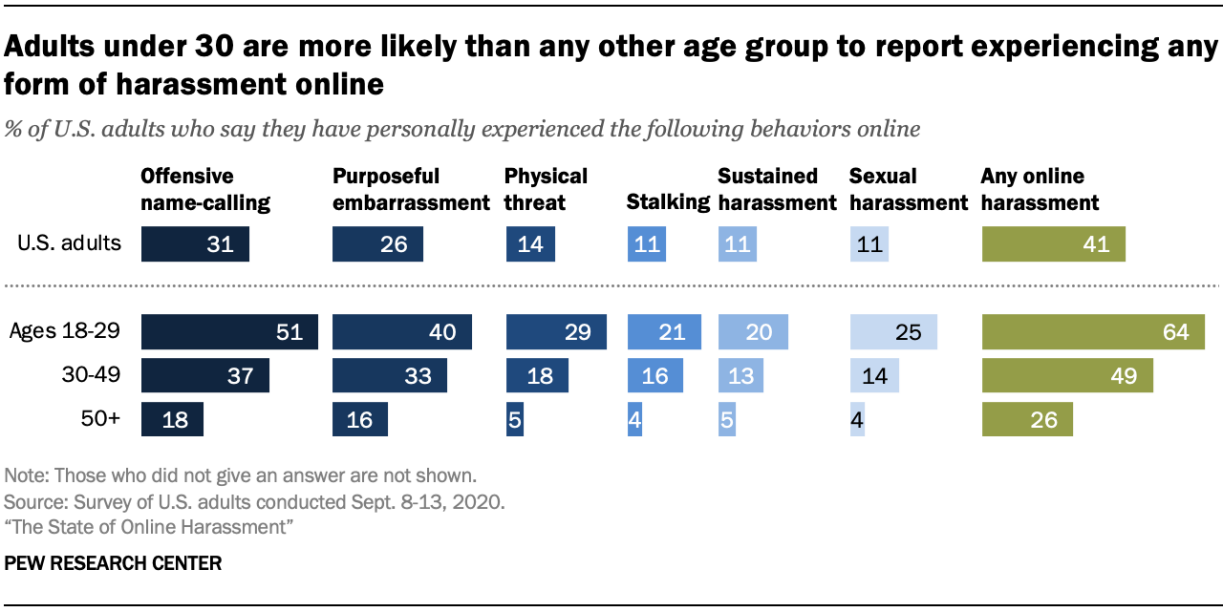

Figure 4. Experiences with certain types of online abuse vary by age, gender, race, or ethnicity in the U.S. in 2020.

3. Cyber-Hate Speech Datasets

This section presents a compilation of the datasets generated over the last five years for cyber-hate speech detection using different characteristics such as the number of classes, language, size, and availability of the datasets, as shown in Table 1. It covers various types of social network textual content such as Formspring (which contains a teen-oriented Q&A forum), Twitter, Instagram, Facebook (which are considered large microblogging platforms), WhatsApp (which is an application for instant messaging that can run on multiple platform

devices) and Wikipedia talk pages (that could be described as a collaborative knowledge repository). Each dataset states a different topic of cyber-hate speech. Twitter datasets comprise examples of offensive, racist, and sexist tweets. Facebook datasets also contain racism and sexism statuses. Instagram and YouTube datasets include examples of personal attacks. However, Formspring datasets are not explicitly about a single topic.

Mangaonkar et al. [21] proposed a binary dataset that contains two subset datasets. The first subset dataset was a balanced dataset with 170 bullying tweets and 170 non-bullying tweets. The second sample was unbalanced, with 177 bullying tweets and 1163 non-bullying tweets. The purpose of developing a balanced and imbalanced dataset is to test the performance of the ML algorithms on various dataset types. These tweets were then manually categorized as "bullying" or "nonbullying" for validation purposes.

Van Hee et al. [22] collected binary cyberbullying data from the social networking site Ask.fm (https://ask.fm/, access on 16 February 2023). They created and implemented a novel method for cyberbullying annotation that describes the existence and intensity of cyberbullying and the role of the author of the post, for example, a harasser, victim, a bystander or not.

Waseem and Hovy [23] gathered a dataset of tweets over a period of two months. They downloaded 136,052 tweets and annotated 16,914 of them, 3383 of which were sexist content sent by 613 users, 1972 of which were racist content sent by 9 users, and 11,559 of which were neither sexist nor racist and were sent by 614 users. Because hate speech is a genuine but restricted phenomenon, they did not balance the data in order to present the most realistic dataset feasible.

Zhao et al. [24] proposed a Twitter dataset that is made up of tweets retrieved from the public Twitter API stream. At least one of the following keywords appears in each tweet: bully, bullied, bullying. Retweets are eliminated by filtering tweets that contain the acronym "RT". Finally, 1762 tweets are randomly selected and manually tagged from the entire twitter collection. It is important to note that labeling is based on bullying traces. Bullying traces are defined as a reaction to a bullying encounter, which includes but vastly outnumbers instances of cyberbullying.

Singh et al. [25] used the Twitter corpus from the Content Analysis for the WEB 2.0 (CAW 2.0) dataset [26]. This corpus comprises around 900,000 postings from 27,135 users (one XML file for each user) from December 2008 to January 2009. They picked this corpus not just because it has been widely used in prior literature but also because it contains information for both textual content and social networks. They chose 800 files at random and kept the comments written with @, which represents direct paths between two people. This yielded a data set of roughly 13,000 messages. Then, they asked three students to categories each message as cyberbullying twice. They designated each post as 'yes' or 'no' based on whether it was believed to entail cyberbullying. This resulted in a data collection with 2150 user pairings and 4865 messages between them.

Al-garadi et al. [27] collected their data via Twitter during January and February of 2015. They have 2.5 million geo-tagged tweets in their data set. To avoid any privacy violations, they extract only publicly available content via the Twitter API and in accordance with Twitter's privacy policy in their study. Their dataset had an uneven class distribution, with just 599 tweets labeled as cyberbullying and 10,007 tweets classified as non-cyberbullying. Such an uneven class distribution can make it difficult for the model to appropriately categories the instances. Learning algorithms that lack class imbalance are prone to be overwhelmed by the major class while ignoring the minor. In real-world applications like fraud detection, instruction detection, and medical diagnosis, data sets frequently contain imbalanced data in which the normal class is the majority, and the abnormal class is the minority. Several solutions to these issues have been offered, including a combination of oversampling the minority (abnormal) class and under-sampling the majority (normal) class.

Hosseinmardi et al. [28] gathered data by using the Instagram API and a snowball sampling technique. They found 41 K Instagram user ids starting from a random seed

node. Of these Instagram IDs, 25 K (61%) belonged to users who had public profiles, while the remaining users had private ones.

Zhang et al. [29] gathered data from the social networking site Formspring.me. Almost 3000 messages were collected and labeled by Amazon Mechanical Turk, a web service in which three workers each voted on whether or not a document contained bullying content. As a result, each message receives an equal number of votes from the workers. At least two workers labeled approximately 6.6% of the messages as bullying posts. The authors parsed the original dataset's messages into sentences and relabeled the messages that contained at least one vote. This resulted in 23,243 sentences, with 1623 (or roughly 7%) labeled as bullying messages.

Wulczyn et al. [30] used English Wikipedia to generate a corpus of over 100 k high-quality human-labeled comments. The collected data of debate comments from English Wikipedia discussion pages are generated by computing differences throughout the whole revision history and extracting the new content for each revision. About 10 annotators using Crowdflower (https://www.crowdflower.com/, access on 16 February 2023) annotated the collected corpus into two classes, either attacking or not attacking.

Batoul et al. [31] gathered a massive amount of data. As a result, the decision was made to scrape data from both Facebook and Twitter. This decision was influenced by the fact that those two social media platforms are the most popular among Arabs, particularly Arab youth. After removing all duplicates, this dataset contains 35,273 unique Arabic tweets that were manually labeled as bullying or not bullying.

Davidson et al. [32] gathered tweets containing hate speech keywords using a crowd-sourced hate speech lexicon. They used crowdsourcing to categorize a sample of these tweets into three groups: those containing hate speech, those containing only offensive language, and those containing neither. This dataset resulted in a total of 33,458 tweets.

Sprugnoli et al. [33] presented and distributed an Italian WhatsApp dataset developed through role-playing by three classes of children aged 12–13. The publicly available data has been labeled based on user role and insult type. The Whatsapp chat corpus consists of 14,600 tokens separated into 10 chats. Two annotators used the Celct Annotation Tool (CAT) web-based application [34] to annotate all of the chats.

Founta et al. [35] have published a large-scale crowdsourced abusive tweet dataset of 60 K tweets. An enhanced strategy is applied to effectively annotate the tweets via crowdsourcing. Through such systematic methods, the authors determined that the most appropriate label set in identifying abusive behaviors on Twitter is None, Spam, Abusive, and Hateful, resulting in 11% as 'Abusive,' 7.5% as 'Hateful,' 22.5% as 'Spam,' and 59% as 'None.' They prepare this dataset for a binary classification task by concatenating 'None'/'Spam' and 'Abusive'/'Hateful'.

De Gibert et al. [36] demonstrated the first dataset of textual hate speech annotated at the sentence level. Sentence-level annotation enables dealing with the smallest unit containing hate speech while decreasing noise generated by other clean sentences. About 10,568 sentences were collected from Storm-front and categorized as hate speech or not, as well as two other auxiliary types.

Nurrahmi and Nurjanah [37] gathered information from Twitter. Because the data was unlabeled, the authors created a web-based labeling tool to categorize tweets as cyberbullying or non-cyberbullying. The tool used provided them with 301 cyberbullying tweets and 399 non-cyberbullying tweets.

Albadi et al. [38] investigated the issue of religious hate speech in the Arabic Twittersphere and developed classifiers to detect it automatically. They gathered 6000 Arabic tweets referring to various religious groups and labeled them using crowdsourced workers. They provided a detailed analysis of the labeled dataset, identifying the primary targets of religious hatred on Arabic Twitter. Following the preprocessing of the dataset, they used various feature selection methods to create different lexicons comprised of terms found in tweets discussing religion, as well as scores reflecting their strength in distinguishing sentiment polarity (hate or not hate).

Bosco et al. [39] released a Twitter dataset for the HaSpeeDe (Hate Speech Detection) shared task at Evalita 2018, the Italian evaluation campaign for NLP and speech processing tools. This dataset contains a total of 4000 tweets, with each tweet having an annotation that falls into one of two categories: "hateful post" or "not".

Corazza et al. [40] provided a set of 5009 German tweets manually annotated at the message level with the labels "offense" (abusive language, insults, and profane statements) and "other". More specifically, 1688 messages are tagged as "offense", while 3321 messages are as "other".

Mulki et al. [41] presented the first publicly available Levantine Hate Speech and Abusive (L-HSAB) Twitter dataset, intending to serve as a reference dataset for the automatic identification of online Levantine toxic content. The L-HSAB is a political dataset because the majority of tweets were gathered from the timelines of politicians, social/political activists, and TV anchors. The dataset included 5846 tweets divided into three categories: normal, abusive, and hateful.

Ptaszynski et al. [42] provided the first dataset for the Polish language that included annotations of potentially dangerous and toxic words. The dataset was designed to investigate negative Internet phenomena such as cyberbullying and hate speech, which have recently grown in popularity on the Polish Internet as well as globally. The dataset was obtained automatically from Polish Twitter accounts and annotated by lay volunteers under the supervision of a cyberbullying and hate-speech expert with a total number of 11,041 tweets.

Ibrohim and Budi [43] offered an Indonesian multi-label hate speech and abuse dataset with over 11,292 tweets based on a diverse collection of 126 keywords. These tweets include 6187 non-hate speech tweets and 5105 hate speech tweets and are annotated by 3 annotators.

Basile et al. [44] investigated the detection of hate speech from a multilingual perspective on Twitter. They focused on two specific targets, immigrants and women, in Spanish and English. They made a dataset containing English (13,000) and Spanish (6600) tweets tagged concerning the prevalence of hostile content and its target.

Banerjee et al. [45] investigated the identification of cyberbullying on Twitter in the English language. The dataset used on Twitter consists of 69,874 tweets. A group of human annotators manually labeled the selected tweets as either "0" non-cyberbullying or "1" cyberbullying.

Lu et al. [46] presented a new Chinese Weibo (https://us.weibo.com/index, access on 16 February 2023) comment dataset designed exclusively for cyberbullying detection. They collected a dataset of 17 K comments from more than 20 celebrities with bad reputations or who have been involved in violent incidents. Three members who are familiar with Weibo and have a good understanding of the bloggers manually annotated all data.

Moon et al. [47] offered 9.4 K manually labeled entertainment news comments that were collected from a popular Korean online news portal for recognizing toxic speech. About 32 annotators labeled the comments manually.

Romim et al. [48] created a large dataset of 30,000 comments, 10,000 of which are hate speech, in the Bengali Language. All user comments on YouTube and Facebook were annotated three times by 50 annotators, with the majority vote serving as the final annotation.

Karim et al. [49] offered an 8 K dataset of hateful posts gathered from various sources such as Facebook, news articles, blogs, and so on in the Bengali language. A total of 8087 posts were annotated by three annotators (a linguist, a native Bengali speaker, and an NLP researcher) into political, personal, geopolitical, and religious.

Luu et al. [50] presented the ViHSD, a human-annotated dataset for automatically detecting hate speech on social networks. This dataset contains over 30,000 comments, each of which has one of three labels: CLEAN, OFFENSIVE, or HATE. The ViHSD contains 33,400 comments.

Sadiq et al. [51] presented Data Turks' Cyber-Trolls dataset for text classification purposes. To assist or prevent trolls, this dataset is used to classify tweets. There are two categories: cyber-aggressive (CA) and non-aggressive (NCA). The dataset contains 20,001 items, of which 7822 are cyber-aggressive, and 12,179 are not.

Beyhan et al. [52] compiled a hate speech dataset extracted from tweets in Turkish. The Istanbul Convention dataset is made up of tweets sent out following Turkey's departure from the Istanbul Convention. The Refugees dataset was produced by collecting tweets regarding immigrants and filtering them based on regularly used immigration keywords.

Ollagnier et al. [53] presented the CyberAgressionAdo-V1 dataset, which contains aggressive multiparty discussions in French obtained through a high-school role-playing game with 1210 messages. This dataset is based on scenarios that mimic cyber aggression situations that may occur among teenagers, such as ethnic origin, religion, or skin color. The collected conversations have been annotated in several layers, including participant roles, the presence of hate speech, the type of verbal abuse in the message, and whether utterances use different humor figurative devices such as sarcasm or irony.

ALBayari and Abdallah [54] introduced the first Instagram Arabic corpus (multi-class sub-categorization) concentrating on cyberbullying. The dataset is primarily intended for detecting offensive language in the text. They ended up with 200,000 comments, with three human annotators annotating 46,898 of them manually. They used SPSS (Kapa statistics) to evaluate the labeling agreements between the three annotators in order to use the dataset as a benchmark. The final score was 0.869, with a p-value of 103, indicating a near-perfect agreement among the annotators.

Patil et al. [55] investigated hate speech detection in Marathi, an Indian regional language. They presented the L3Cube-MahaHate Corpus, the largest publicly available Marathi hate speech dataset. The dataset was gathered from Twitter and labeled with four fine-grained labels: Hate, Offensive, Profane, and None. The dataset contains over 25,000 samples that have been manually labeled with the classes.

Kumar and Sachdeva [56] developed two datasets FormSpring.me and MySpace. The Formspring.me dataset is an XML file containing 13,158 messages published by 50 different users on the Formspring.me website. The dataset is divided into two categories: "Cyberbullying Positive" and "Cyberbullying Negative". While negative messages represent messages that do not include cyberbullying, positive messages include cyberbullying. There are 892 messages in the Cyberbullying Positive class and 12,266 messages in the Cyberbullying Negative class. The Myspace dataset is made up of messages gathered from Myspace group chats. The dataset's group chats are labeled and organized into ten message groups. If a group conversation contains 100 messages, the first group contains 1–10 messages, the second group contains 2–11 messages, and the final message group contains 91–100 messages. Labeling is done once for each group of ten messages, and it is labeled whether or not those ten messages contain bullying. This dataset contains 1753 message groups divided into 10 groups, each with 357 positive (Bullying) and 1396 negative (Non-Bullying) labels.

Atoum [57] collected two datasets (Dataset-1 and Dataset-2) from Twitter (one month apart). Twitter dataset 1 consists of 6463 tweets, with 2521 cyberbullying tweets and 3942 non-cyberbullying tweets. Twitter dataset 2 consists of 3721 with 1374 cyberbullying tweets and 2347 non-cyberbullying tweets.

Nabiilah et al. [58] proposed a dataset of toxic comments that were manually collected, processed, and labeled. Data is gathered from Indonesian user comments on social media platforms such as Instagram, Twitter, and Kaskus (https://www.kaskus.co.id/, access on 16 February 2023), which have multi-label characteristics and allow for the classification of more than one class. Pornography, Hate Speech, Radicalism, and Defamation are among the 7773 records in the dataset.

Below Table 1 is a brief description of each of the datasets.

Table 1. Cyber-hate Speech Datasets.

Dataset	Category	Number of Classes	Classes	Social Network Platform	Language	Size	Availability	Year
Mangaonkar et al. [21]	Trolling and Harassment	2	Bullying Non-Bullying	Twitter	English	1340	N/A	2015
Van Hee et al. [22]	Cyber Threats and Harassment	2	Bullying Non-Bullying	Ask.fm	Dutch	85,485	N/A	2015
Waseem and Hovy [23]	Cyber Threats and Harassment	3	Racism Sexism None	Twitter	English	16 K	[59]	2016
Zhao et al. [24]	Trolling and Harassment	2	Bullying Non-Bullying	Twitter	English	1762	N/A	2016
Singh et al. [25]	Trolling and Harassment	2	Bullying Non-Bullying	Twitter	English	4865	N/A	2016
Al-garadi et al. [27]	Trolling and Harassment	2	Bullying Non-Bullying	Twitter	English	10,007	N/A	2016
Hosseinmardi et al. [28]	Flaming and Stalking and Harassment	2	Bullying Non-Bullying	Instagram	English	1954	N/A	2016
Zhang et al. [29]	Trolling and Harassment	2	Bullying Non-Bullying	Formspring	English	13 K	N/A	2016
Wulczyn et al. [30]	Denigration and Masquerade and Harassment	2	Attacking Non-Attacking	Wikipedia	English	100 K	[60]	2017
Batoul et al. [31]	Trolling and Harassment	2	Bullying Non-Bullying	Twitter	Arabic	35,273	N/A	2017
Davidson et al.	Trolling and Harassment	3	Bullying Non-Bullying Neither	Twitter	English	33,458	[61]	2017
Sprugnoli et al. [33]	Flaming and Stalking and Harassment and Trolling	10	Defense General Insult Curse or Exclusion Threat or Blackmail Encouragement to the Harassment Body Shame Discrimination-Sexism Attacking relatives Other Defamation	WhatsApp	Italian	14,600	[62]	2018

Table 1. *Cont.*

Dataset	Category	Number of Classes	Classes	Social Network Platform	Language	Size	Availability	Year
Founta et al. [35]	Cyber Threats and Harassment	7	Offensive Abusive Hateful Aggressive Cyberbullying Spam Normal	Twitter	English	100 K	[63]	2018
De Gibert et al. [36]	Trolling and Harassment	2	Hateful Non-Hateful	Stormfront	English	10,568	[64]	2018
Nurrahmi and Nurjanah [37]	Trolling and Harassment	2	Bullying Non-Bullying	Twitter	Indonesian	700	N/A	2018
Albadi et al. [38]	Trolling and Harassment	2	Hateful Non-Hateful	Twitter	Arabic	6 K	[65]	2018
Bosco et al. [39]	Trolling and Harassment	2	Bullying Non-Bullying	Twitter	Italian	4 K		2018
Corazza et al. [40]	Trolling and Harassment	2	Bullying Non-Bullying	Twitter	German	5009		2018
Mulki et al. [41]	Trolling and Harassment	3	Normal Abusive Hate	Twitter	Arabic	6 K	[66]	2019
Ptaszynski et al. [42]	Trolling and Harassment	3	Non-harmful Cyberbullying Hate-speech and other harmful contents	Twitter	Polish	11,041	[67]	2019
Ibrohim and Budi [43]	Trolling and Harassment	2	Hateful Non-Hateful	Twitter	Indonesian	11,292	[68]	2019
Basile et al. [44]	Trolling and Harassment	2	Hateful Non-Hateful	Twitter	English Spanish	13,000 6600	[69]	2019
Banerjee et al. [45]	Trolling and Harassment	2	Bullying Non-Bullying	Twitter	English	69,874	N/A	2019
Lu et al. [46]	Cyber Threats and Harassment	3	Sexism Racism Neither	Sina Weibo	Chinese	16,914	[70]	2020
Moon et al. [47]	Trolling and Harassment	3	Hateful Offensive None	Online News Platform	Korean	9.4 K	[71]	2020
Romim et al. [48]	Trolling and Harassment	2	Hateful Non-Hateful	Facebook and YouTube	Bengali	30 K	[72]	2021
Karim et al. [49]	Trolling and Harassment	2	Hateful Non-Hateful	Facebook, YouTube comments, and newspapers	Bengali	8087	[73]	2021
Luu et al. [50]	Trolling and Harassment	3	Hate Offensive Clean	Facebook and YouTube	Vietnamese	33,400	[74]	2021

Table 1. *Cont.*

Dataset	Category	Number of Classes	Classes	Social Network Platform	Language	Size	Availability	Year
Sadiq et al. [51]	Trolling and Harassment	2	Bullying Non-Bullying	Twitter	English	20,001	[75]	2021
Beyhan et al. [52]	Trolling and Harassment	2	Hateful Non-Hateful	Twitter	Turkish	2311	[76]	2022
Ollagnier et al. [53]	Flaming, Stalking, Harassment and Trolling	2	Hateful Non-Hateful	WhatsApp	French	1210	[77]	2022
ALBayari and Abdallah [54]	Flaming, Stalking, Harassment and Trolling	2	Bullying Non-Bullying	Instagram	Arabic	46,898	[78]	2022
Patil et al. [55]	Trolling and Harassment	4	Hate Offensive Profane None	Twitter	Marathi	25 K	[79]	2022
Kumar and Sachdeva [56]	Trolling and Harassment	2	Bullying Non-Bullying	Formspring MySpace	English	13,158 1753	N/A	2022
Atoum [57]	Trolling and Harassment	2	Bullying Non-Bullying	Twitter Dataset 1 Twitter Dataset 2	English	6463 3721	N/A	2023
Nabiilaha et al. [58]	Trolling, Harassment and Flaming	2	Bullying Non-Bullying	Instagram, Twitter and Kaskus	Indonesian	7773	N/A	2023

As demonstrated in Table 1, the majority of the studies and experiments were implemented on Twitter datasets. This is due to the effortless accessibility and availability of tweets that can be crawled utilizing the Twitter API. Out of all, most of the research focuses on the identification of hate speech and differentiating them from non-hate (or offensive) texts. Most of the research into cyber-hate was applied in the English language, while related work in other languages is scarce due to the lack of available datasets or the difficulty of their morphology as in the Arabic language.

4. Cyber-Hate Speech Detection Approaches

In recent years, some sentiment-based methods have been published to detect and identify abusive language [80]. These approaches are the machine-learning approach, the lexicon-based approach, and the hybrid approach, as shown in Figure 5.

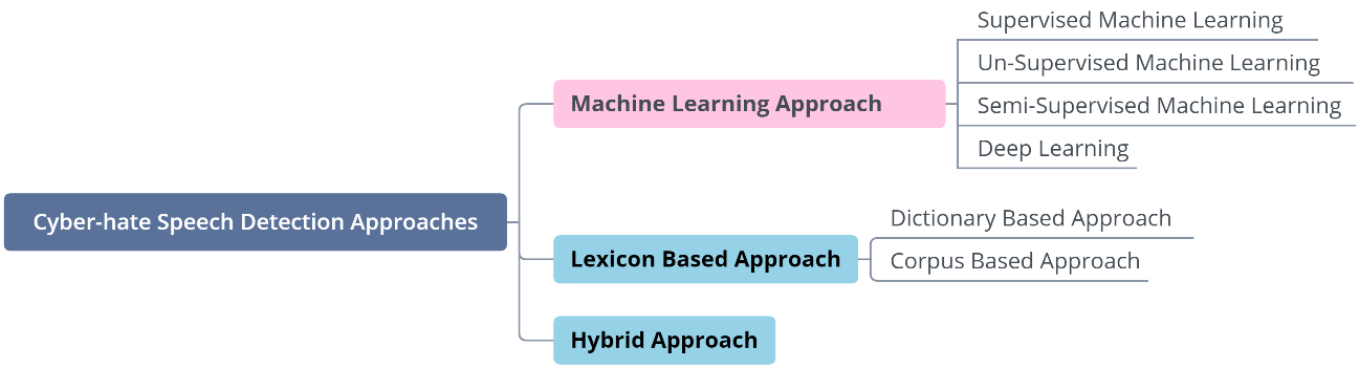

Figure 5. Cyber-hate Speech Detection Approaches.

On the one hand, the Machine Learning Approach (MLA) comprises the following methods: supervised machine learning, un-supervised machine learning, semi-supervised machine learning and deep learning. In a supervised machine learning approach, the classifier is built to learn the properties of categories or classes automatically from a set of pre-annotated training textual content. When utilizing the supervised machine learning approach, some main issues and challenges have to be considered, such as the categories to be used to classify the instances, the labeled training data, the extracted and selected features to be used to represent each unknown textual content, and the selected algorithm to be used for categorization [81]. In unsupervised machine learning, the machine attempts to find and understand the hidden structure within unlabeled data [82]. Semi-supervised learning is concerned with how the combination of labeled and unlabeled data will change the behavior of learning and designing algorithms that benefit from such a combination [83]. The Deep Learning approach, inspired by artificial neural networks, is an evolving branch of machine learning [84]. With the aid of the hierarchy of layers, it provides ways of learning data representations in a supervised and unsupervised manner, allowing multiple processing [85].

On the other hand, the Lexicon-Based Approach (LBA) comprises making a list of words that is called the dictionary, which is searched and counted in the textual content. These calculated frequencies can be utilized explicitly as features or to calculate scores for classifying textual content. A potential limitation of this approach regarding its classification efficacy is its dependency on domain-specific words presented in a dictionary; also, it needs an automatic methodology for the classification and scoring of words to reduce the amount of manpower required for the manual scoring of domain-specific words [86]. A corpus-based approach utilizes a collection of sentiment words with pre-defined polarity to recognize new sentiment words and their polarity in a large corpus [87]. A corpus-based approach provides a data-driven approach where one has access not only to sentiment labels but also to a context that one can use. A dictionary-based approach exploits the lexicographical tools such as Artha (https://sourceforge.net/projects/artha/, access on 16 February 2023), Tematres (https://www.vocabularyserver.com/, access on 16 February 2023), Wordhoard (https://wordhoard.northwestern.edu/, access on 16 February 2023), or WordNet (https://wordnet.princeton.edu/, access on 16 February 2023). In these, the key strategy methods are gathering an initial collection of sentiment words and manually orienting them; then, looking in a dictionary to enlarge this collection by finding their synonyms and antonyms [88].

Finally, the Hybrid Approach (HA) is the amalgamation of both machine learning and lexicon-based methods.

5. Cyber-Hate Detecting Techniques

Textual data mining and analysis have become an active and attractive research field. The global availability of such data makes text analytics acquire a major consideration. Hate speech detection tasks can be performed as binary or multi-class classifications based on the number of classes in these datasets.

5.1. Binary Cyber-Hate Classification

Cyber-hate detection has been approached as a binary classification task (cyberbullying -vs.- non-cyber bullying) or (Hate–Non-Hate). This section presents a summary of studies in cyber-hate binary classification techniques.

Mangaonkar et al. [21] applied different algorithms to classify tweets, then performed AND and OR parallelism. They combined the output of multiple classifiers to enhance the performance. They classified tweets using a four-node detection system and experimented with homogeneous (all computing nodes use the same classification algorithm), heterogeneous (each node uses a different algorithm), and selective (the best-performing node is chosen as the expert, and all other nodes defer to it) collaborations. Each tweet is processed by all nodes and classified as cyberbullying if more than half of the nodes in the AND

configuration flag it as bullying or if any node flags it as bullying in the OR configuration. They discovered that OR parallelism produces the highest recall values at 60%, while AND parallelism produces the highest accuracy at 70%.

Van Hee et al. [22] presented a proposed system in the Dutch language for the automatic detection of cyberbullying. The dataset enclosing cyberbullying posts was gathered from the social networking site Ask.fm. The experimental results showed that Support Vector Machines (SVM) achieved an F1-score of 55.39%.

Nandhini and Sheeba [89] proposed a system for detecting the existence of cyberbullying activity on social networks in the English language in order to help the government take action before more people become cyberbullying victims. The used dataset has a record of almost 4 K, which is gathered from social networks (Formspring.me, Myspace.com) [90]. For this purpose, they used the Naïve Bayes (NB) classifier achieving 92% on the Formspring dataset and 91% on MySpace.me dataset.

Zhao et al. [24] have presented Embedding-enhanced Bag-of-Words (EBoW), a unique representation learning method for cyberbullying detection. EBoW combines bag of words features, latent semantic characteristics, and bullying features. Bullying characteristics are generated from word embeddings, which can capture the semantic information behind words. When the final representation is learned, a linear SVM is used to detect bullying messages with a recall of 79.4%.

Singh et al. [25] employed probabilistic fusion approaches to mix social and text information as the classifier's input. The proposed methodology has been applied to the English Twitter dataset. The accuracy of the obtained results was 89%.

Al-garadi et al. [27] utilized supervised machine learning algorithms such as NB, SVM, Random Forest (RF), and K Nearest Neighbor (KNN) to detect cyberbullying on Twitter in the English language. Based on an evaluation, their model accuracy is 70.4% by NB, 50% by SVM, 62.9% by Random Forest (RF), and 56.8% by KNN.

Hosseinmardi et al. [28] investigated the problem of predicting cyberbullying in the Instagram media-based social network. They demonstrated that non-text features such as image and user metadata were important in predicting cyberbullying, with a Logistic Regression (LR) classifier achieving 72% recall and 78% precision.

Zhang et al. [29] proposed a novel Pronunciation-based Convolutional Neural Network (PCNN) to detect cyberbullying. They assessed the performance of their model using a cyberbullying dataset in English from Formspring.me. Their experiment revealed that PCNN can achieve an accuracy of 88.1%.

Wulczyn et al. [30] demonstrated a methodology in cyberbullying detection by applying LR and Multi-Layer Perceptron (MLP) to Wikipedia, resulting in an open dataset of over 100 k high-quality human-labeled comments. They evaluated their models using Area Under the Receiver Operating Characteristic (AUROC) and achieved 96.18% using LR and 96.59% using MLP.

Batoul et al. [31] proposed a system for detecting Arabic cyberbullying. They worked on an Arabic Twitter dataset that contains 35,273 unique tweets after removing all duplicates. NB and SVM obtained F-measure with 90.5% and 92.7%.

De Gibert et al. [36] conducted a thorough qualitative and quantitative analysis of their dataset, as well as several baseline experiments with various classification models, which are SVM, Convolution Neural Networks (CNN), and Long-Short Term Memory (LSTM). The experiments employ a well-balanced subset of labeled sentences. All of the HATE sentences were collected, and an equal number of NOHATE sentences were randomly sampled, a total of 2 k labeled sentences. Eighty percent of this total has been used for training, with the remaining 20% for testing. The evaluated algorithms, SVM, CNN, and LSTM, achieved 71%, 66%, and 73% accuracy, respectively.

Nurrahmi and Nurjanah [37] studied cyberbullying detection for Indonesian tweets to recognize cyberbullying text and actors on Twitter. The study of cyberbullying has successfully identified tweets that contain cyberbullying with an F1-score of 67% utilizing the SVM algorithm.

Albadi et al. [38] are the first to address the issue of identifying and recognizing speech promoting religious hate on Arabic Twitter. They implemented different classification models utilizing lexicon-based, n-gram-based, and deep-learning-based approaches. They concluded that a straightforward Recurrent Neural Networks (RNN) architecture with Gated Recurrent Units (GRU) and pre-trained word embeddings could sufficiently detect religious hate speech since it gives an AUROC of 84%.

Basile et al. [39] evaluated the SVM on a dataset of 13,000 tweets in English and 6600 tweets in Spanish [45], 60% of which were labeled as hate speech. In terms of performance, the system had an F1-score of 65% for the English tweets and 73% for the Spanish tweets.

Ibrohim and Budi [43] conducted a combination of feature, classifier, and data transformation methods between word unigram, Random Forest Decision Tree (RFDT), and Label Power-set (LP) to identify abusive language and hate speech. Their system achieved an accuracy of 77.36% for the classification of hate speech without identifying the target, categories, and level of hate speech. Moreover, their system identifies abusive language and hate speech, including identifying the target, categories, and level of hate speech with an accuracy of 66.1%.

Banerjee et al. [45] represented an approach for the detection of cyberbullying in the English language. They applied CNN to a Twitter dataset of 69,874 tweets. Their proposed approach achieved an accuracy of 93.97%.

Corazza et al. [4] proposed a neural architecture for identifying the forms of abusive language, which shows satisfactory performance in several languages, namely English [23], Italian [37], and German [40]. Different components were employed in the system, which are Long Short-Term Memory (LSTM), GRU, and Bidirectional Long Short-Term Memory (BiLSTM). For the feature selection, they used n-gram, word embedding, social network-specific features, emotion lexica, and emoji. The results show that LSTM outperforms other used algorithms in multilingual classification with an F1-score of 78.5%, 71.8%, and 80.1% in English, German, and Italian, respectively.

Romim et al. [48] ran a baseline model (SVM) and several deep learning models, as well as extensive pre-trained Bengali word embedding such as Word2Vec, FastTest, and BengFastText, on their collected dataset (Facebook and YouTube comments). The experiment demonstrated that, while all of the deep learning models performed well, SVM achieved the best result with an 87.5% accuracy.

Karim et al. [49] proposed DeepHateExplainer, an explainable approach for detecting hate speech in the under-resourced Bengali language. Bengali texts are thoroughly preprocessed before being classified into political, personal, geopolitical, and religious hatreds using a neural ensemble method of transformer-based neural architectures (i.e., monolingual Bangla BERT-cased/uncased, and XLM-RoBERTa). Before providing human-interpretable explanations for hate speech detection, important (most and least) terms are identified using a sensitivity analysis and layer-wise relevance propagation (LRP). Evaluations against machine learning (linear and tree-based models) and neural networks (i.e., CNN, Bi-LSTM, and Conv-LSTM with word embeddings) baselines produce F1-scores of 78%, 91%, 89%, and 84%, respectively, outperforming both ML and DNN baselines.

Sadiq et al. [51] proposed a system using a combination of CNN with LSTM and CNN with BiLSTM for cyberbullying detection on English tweets of the cyber-troll dataset. Statistical results proved that their proposed model detects aggressive behavior with 92% accuracy.

Beyhan et al. [52] created a hate speech detection system (BERTurk) based on the transformer architecture to serve as a baseline for the collected dataset. The system is evaluated using 5-fold cross-validation on the Istanbul Convention dataset; the classification accuracy is 77%.

ALBayari and Abdallah [54] used the most basic classifiers (LR, SVM, RFC, and Multinomial Naive Bayes (MNB)) for cyberbullying detection using their dataset. As a result, the SVM classifier has a significantly higher F1-score value of 69% than the other classifiers, making it a preferable solution.

Kumar and Sachdeva [56] proposed a hybrid model, Bi-GRU Attention-CapsNet (Bi-GAC), that benefits from learning sequential semantic representations and spatial location information using a Bi-GRU with self-attention followed by CapsNet for cyberbullying detection in social media textual content. The proposed Bi-GAC model is evaluated for performance using the F1-score and the ROC-AUC curve as metrics. On the benchmark Formspring.me and MySpace datasets, the results outperform existing techniques. In comparison to conventional models, the F1-score for MySpace and Formspring.me datasets achieved by nearly 94% and 93%, respectively.

Atoum [57] developed and refined an efficient method for detecting cyberbullying in tweets that uses sentiment analysis and language models. Various machine learning algorithms are examined and compared across two tweet datasets. CNN classifiers with higher n-gram language models outperformed other ML classifiers such as DT, RF, NB, and SVM. The average accuracy of CNN classifiers was 93.62% and 91.03%.

Nabilah et al. [58] used a Pre-Trained Model trained for Indonesian to detect comments containing toxic sentences on social media in Indonesia. The Multilingual BERT (MBERT), IndoBERT, and Indo Roberta Small models were used in this study to perform a multi-label classification and evaluate the classification results. The BERT model with an F1 Score of 88.97% yielded the best results in this study.

5.2. Multi-Class Cyber-Hate Classification

Several studies have been conducted for multi-class cyber-hate classification. This section summarizes the studies of multi-class cyber-hate classification techniques in different languages.

Waseem and Hovy [23] investigated the impact of various features in the classification of cyberbullying. They used an LR classifier and 10-fold cross-validation to test and quantify the impact of various features on prediction performance with an F1-score of 73%.

Badjatiya et al. [91] investigate the use of deep neural network architectures for hate speech detection in the English language [23]. They proposed a combination of deep neural network model embeddings and gradient-boosted decision trees, leading to better accuracy values. Embeddings gained from deep neural network models, when joined with gradient-boosted decision trees, prompted the best accuracy values with an F1-score of 93%.

Park and Fung [92] proposed a two-step approach to abusive language classification for detecting and identifying sexist and racist languages. They first classify the language as abusive or not and then classify it into explicit types in a second step. With a public English Twitter corpus [23] that contains 20 thousand tweets of a sexist and racist nature, their approach shows a promising performance of 82.7% F1-score using Hybrid-CNN in the first step and 82.4% F1-score using LR in the second step.

Watanabe et al. [93] presented a methodology to detect hate speech on Twitter in the English language [23]. The proposed approach consequently detects hate speech signs and patterns using unigrams as feature extraction along with sentimental and semantic features to classify tweets into hateful, offensive, and clean. The proposed approach achieves an accuracy of 78.4% for the classification of tweets.

Mulki et al. [41] presented the first publicly available Levantine Hate Speech and Abusive Behavior (L-HSAB) Twitter dataset intending to serve as a benchmark dataset. NB and SVM classifiers were used in machine learning-based classification experiments on L-HSAB. The results showed that NB outperformed SVM in terms of accuracy, with 88.4% and 78.6%, respectively.

Lu et al. [46] proposed an automatic method for determining whether the text in social media contains cyberbullying. It learns char-level features to overcome spelling mistakes and intentional data obfuscation. On the Weibo dataset, the CNN model achieved Precision, F1-score, and Recall values of 79.0%, 71.6%, and 69.8%, respectively.

Moon et al. [47] presented 9.4 K manually labeled entertainment news comments collected from a popular Korean online news platform for identifying Korean toxic speech.

They used three baseline classifiers: a character-level convolutional neural network (Char-CNN), a bidirectional long short-term memory (BiLSTM), and a bidirectional encoder representation from a Transformer (BERT) model. BERT has the best performance, with an F1-score of 63.3%.

Luu et al. [44] created the ViHSD dataset, a large-scale dataset for detecting hate speech in Vietnamese social media texts. The dataset contains 33,400 human-annotated comments and achieves an F1-score of 62.69% using the BERT model.

Patil et al. [55] presented L3CubeMahaHate, a hate speech dataset with 25,000 distinct samples evenly distributed across four classes. They ran experiments on various deep learning models such as CNN, LSTM, BiLSTM, and transformer-based BERT. The BERT model outperformed other models with an accuracy of 80.3%.

Wang et al. [94] proposed a framework for Metamorphic Testing for Textual Content Moderation (MTTM) software. They conducted a pilot study on 2000 text messages from real users and summarized eleven metamorphic relations at three perturbation levels: character, word, and sentence. MTTM uses these metamorphic relations on the toxic textual content to generate test cases that are still toxic but are unlikely to be moderated. When the MTTM is tested, the results show that the MTTM achieves up to 83.9% error-finding rates.

5.3. Analysis of the Literature Review

Table 2 presents, for each of the previously described works on binary and multiclass classification, a summary of the dataset used in the experimentation and its number of classes, the language under study, the algorithms tested, and the results obtained.

This comparative study identified that binary classification is the most common task carried out in cyber-hate detection, as shown in Table 2.

Moreover, most of the research on cyber-hate speech detection focuses on English textual content, so most of the resources, assets, libraries, and tools have been implemented for English language use only.

Table 2. Cartography of Existing Research in Hate Speech Detection.

Author	Classes	Dataset	Language	Approach	Algorithm	Evaluation Metric	
Mangaonkar et al. 2015 [21]	2 Classes (Cyberbullying, Non-Cyberbullying)	Twitter	English	MLA	LR (OR parallelism)	Recall	60%
					LR (AND parallelism)	Accuracy	70%
Van Hee et al. 2015 [22]	2 Classes (Cyberbullying, Non-Cyberbullying)	Ask.fm	Dutch	MLA	SVM	F1-score	55.39%
						Recall	51.46%
						Precision	59.96%
Nandhini and Sheeba, 2015 [89]	2 Classes (Cyberbullying–Non-Cyberbullying)	Formspring	English	MLA	NB	Accuracy	92%
		MySpace.com					91%
Waseem and Hovy 2016 [23]	3 Classes (Sexism, Racism, Neither)	Twitter	English	MLA	LR	F1-score	73%
Zhao et al. 2016 [24]	2 Classes (Cyberbullying, Non-Cyberbullying)	Twitter	English	MLA	SVM	F1-score	79.4%
Singh et al. 2016 [25]	2 Classes (Cyberbullying, Non-Cyberbullying)	Twitter	English	LBA	Probabilistic Fusion approach	Accuracy	89%
Al-garadi et al. 2016 [27]	2 Classes (Cyberbullying, Non-Cyberbullying)	Twitter	English	MLA	NB	Accuracy	70.4%
					SVM		50%
					RF		62.9%
					KNN		56.8%
Hosseinmardi et al. 2016 [28]	2 Classes (Cyberbullying, Non-Cyberbullying)	Instagram	English	MLA	LR	Recall	72%
						Precision	78%

Table 2. *Cont.*

Author	Classes	Dataset	Language	Approach	Algorithm	Evaluation Metric	
Zhang et al. 2016 [29]	2 Classes (Cyberbullying, Non-Cyberbullying)	Formspring	English	MLA	PCCN	Accuracy	88.1%.
Wulczyn et al. 2017 [30]	2 Classes (Attacking, Non-Attacking)	Wikipedia	English	MLA	LR	AUROC	96.18%
					MLP		96.59%
Batoul et al. 2017 [31]	2 Classes (Cyberbullying–Non-Cyberbullying)	Twitter	Arabic	MLA	NB	Precision	90.1%
						Recall	90.9%
						F1-score	90.5%
					SVM	Precision	93.4%
						Recall	94.1%
						F1-score	92.7%
Badjatiya, Pinkesh et al. 2017 [91]	3 classes (Sexism, Racism, Neither)	Twitter	English	MLA	LSTM	F1-score	93%
Park, Ji Ho et al. 2017 [92]	3 classes (Sexism, Racism, Neither)	Twitter	English	MLA	CNN	F1-score	82.7%
					LR		82.4%
De Gibert et al. 2018 [36]	2 Classes (Hate, Non-Hate)	Stormfront	English	MLA	SVM	Accuracy	71%
					CNN		66%
					LSTM		73%
Nurrahmi and Nurjanah, 2018 [37]	2 Classes (Cyberbullying–Non-Cyberbullying)	Twitter	Indonesian	MLA	SVM	F1-score	67%
N. Albadi et al. 2018 [38]	2 Classes (Hate, Non-Hate)	Twitter	Arabic	MLA	GRU-based RNN	Precision	76%
						Recall	78%
						F1-score	77%
						AUROC	84%
Watanabe et al. 2018 [93]	3 Classes (Hateful, Offensive and Clean)	Twitter	English	MLA	J48graft	Precision	88%
						Recall	87.4%
						F1-score	87.5%
Mulki et al. 2019 [41]	3 Classes (Normal, Abusive, Hate)	Twitter	English		NB	Accuracy	88.4%
					SVM		78.6%
Ibrohim and Budi, 2019 [43]	2 Classes (Hateful, Non-Hateful)	Twitter	Indonesian		RFDT	Accuracy	77.36%
					LP		66.1%
Basile et al. 2019 [39]	2 Classes (Hate, Non-Hate)	Twitter	English		SVM	F1-score	65%
			Spanish				73%
Banerjee et al. 2019 [45]	2 Classes (Cyberbullying–Non-Cyberbullying)	Twitter	English	MLA	CNN	Accuracy	93.97%
Corazza, Michele et al. 2020 [4]	2 Classes (Hateful, Non-Hateful)	Twitter	English	MLA	LSTM	F1-score	78.5%
			German				71.8%
			Italian				80.1%
Lu et al. 2020 [46]	3 Classes (Sexism, Racism, and Neither)	Sina Weibo	Chinese	MLA	CNN	Precision	79%
						F1-score	71.6%
						Recall	69.7%
Moon et al. 2020 [47]	3 Classes (Hate, Offensive, None)	Korean Online News Platform	Korean	MLA	CharCNN	F1-score	53.5%
					BiLSTM		29.1%
					BERT		63.3%

Table 2. *Cont.*

Author	Classes	Dataset	Language	Approach	Algorithm	Evaluation Metric	
Romim et al. 2021 [48]	2 Classes (Hateful, Non-Hateful)	Facebook and YouTube	Bengali	MLA	SVM	Accuracy	87.5%
					Word2Vec + LSTM		83.85%
					Word2Vec + Bi-LSTM		81.52%
					FastText + LSTM		84.3%
					FastText + Bi-LSTM		86.55%
					BengFastText + LSTM		81%
					BengFastText + Bi-LSTM		80.44%
Karim et al. 2021 [49]	2 Classes (Hateful, Non-Hateful)	Facebook, YouTube comments, and newspapers	Bengali	MLA	LR	F1-score	67%
					NB		64%
					SVM		66%
					KNN		66%
					RF		68%
					GBT		68%
					CNN		73%
					Bi-LSTM		75%
					Conv-LSTM		78%
					Bangla BERT		86%
					mBERT-cased		85%
					XML-RoBERTA		87%
					mBERT-uncased		86%
					Ensemble *		88%
Luu et al. 2021 [44]	3 Classes (Offensive, Hate, None)	Facebook and YouTube	Vietnamese	MLA	BERT	F1-score	62.69%
Sadiq et al. 2021 [51]	2 Classes (Cyber-aggressive, Non-Cyber-aggressive)	Twitter	English	MLA	CNN + LSTM + Bi-LSTM	Accuracy	92%
Beyhan et al. 2022 [52]	2 Classes (Hateful, Non-Hateful)	Twitter	Turkish	MLA	BERTurk	Accuracy	77%
ALBayari and Abdallah 2022 [54]	2 Classes (Cyberbullying–Non-Cyberbullying)	Instagram	Arabic	MLA	MNB	F1-score	66%
					RF		65%
					SVM		69%
					LR		66%
Patil et al. 2022 [55]	4 Classes (Hate, Offensive, Profane, None)	Twitter	Marathi	MLA	CNN	Accuracy	75.1%
					LSTM		75.1%
					BiLSTM		76.1%
					BERT		80.3%
Kumar and Sachdeva 2022 [56]	2 Classes (Cyberbullying–Non-Cyberbullying)	Formspring	English	HA	Bi-GAC	F1-score	94.03%
		MySpace					93.89%
Wang et al. 2023 [94]	3 Classes (Cyberbullying–Non-Cyberbullying, Neither)	Twitter	English	HA	MTTM	Error Finding Rates	83.9%

Table 2. *Cont.*

Author	Classes	Dataset	Language	Approach	Algorithm	Evaluation Metric	
Atoum, 2023 [57]	2 Classes (Cyberbullying–Non-Cyberbullying)	Twitter Dataset 1	English	MLA	CNN	Accuracy	93.62%
		Twitter Dataset 2					91.03%
Nabilah et al. 2023 [58]	2 Classes (Cyberbullying–Non-Cyberbullying)	Instagram and Twitter and Kaskus	Indonesian	MLA	BERT	F1-score	88.97%

As previously mentioned, the most common task carried out in cyber-hate detection is binary classification rather than multi-class classification. Cyber-hate texts are known as representatives of a "bullying" class, and all other documents belong to "non-bullying". Twitter is the most commonly studied data source compared to other social media platforms. Most researchers applied and compared many supervised machine learning algorithms in order to determine the ideal ones for cyber-hate detection problems. As for the traditional machine learning algorithms, SVM has been used to build prediction models for cyberbullying and has been found to be accurate and efficient. On the other hand, CNN was the most common deep learning algorithm used in cyber-hate classification for binary or multiple-class classification. Researchers measure the effectiveness of their proposed model to determine how successfully the model can distinguish cyberbullying texts from non-cyber bullying texts by using various evaluation measures such as F1-score, accuracy, recall, and Precision [95,96].

Subsequently, the algorithms used for binary and multi-class classification for the English and Arabic languages are analyzed according to the results obtained with them.

Figure 6 shows the accuracy of binary cyber-hate classification on different English datasets. As illustrated in it, CNN has better accuracy than SVM, NB, and CNN + LSTM + Bi-LSTM. In addition, NB gives an acceptable accuracy on different datasets between 91% and 92%.

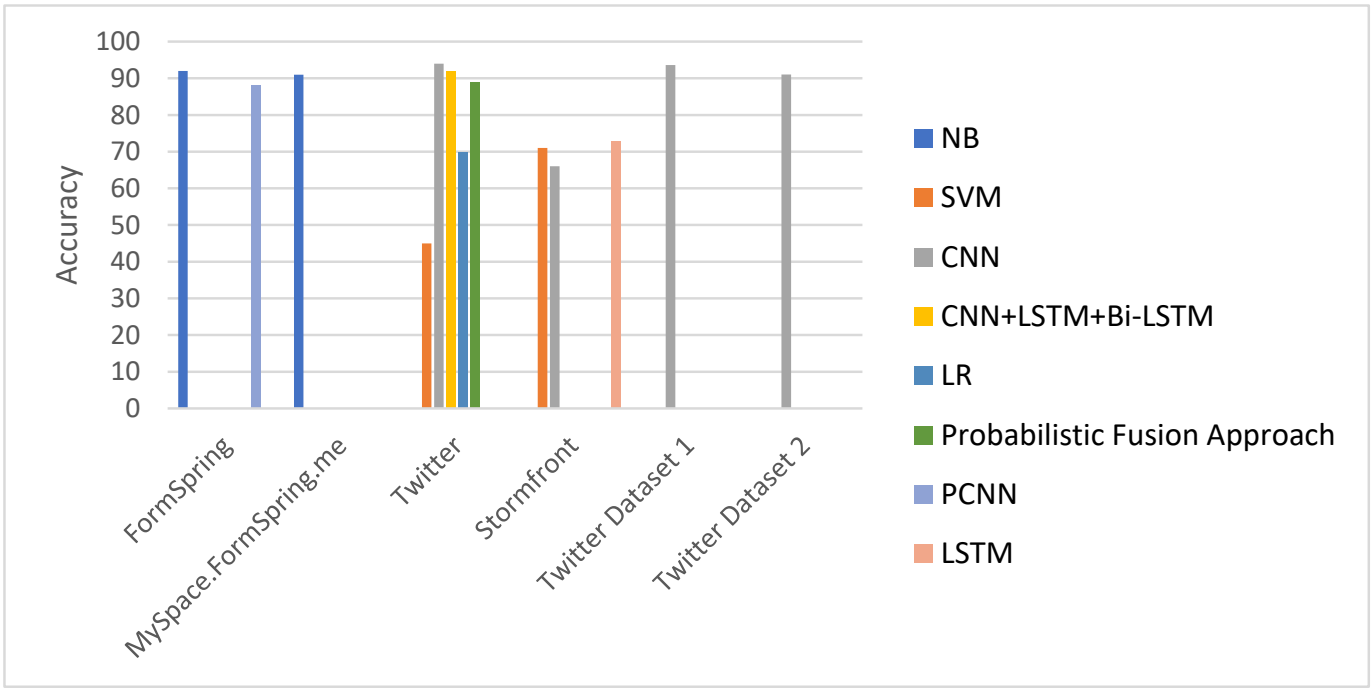

Figure 6. Accuracy of Binary Cyber-hate Classification in English.

Figure 7 shows the F1-Score of binary cyber-hate classification in Arabic on different platforms, which are Twitter and Instagram. In this language, the algorithm providing the best results is SVM on different platforms, with more than 92% in terms of F1 score. Otherwise, the combination CNN+LSTM achieves the lowest value of F1 score, which is 73% on Twitter, and RF with 65% on the Instagram platform.

Finally, Figure 8 shows the F1-Score of multiple class cyber-hate classifications on Twitter in English. It illustrates that the best performance result from different machine learning algorithms applied was LSTM, with an F1-Score of 93%.

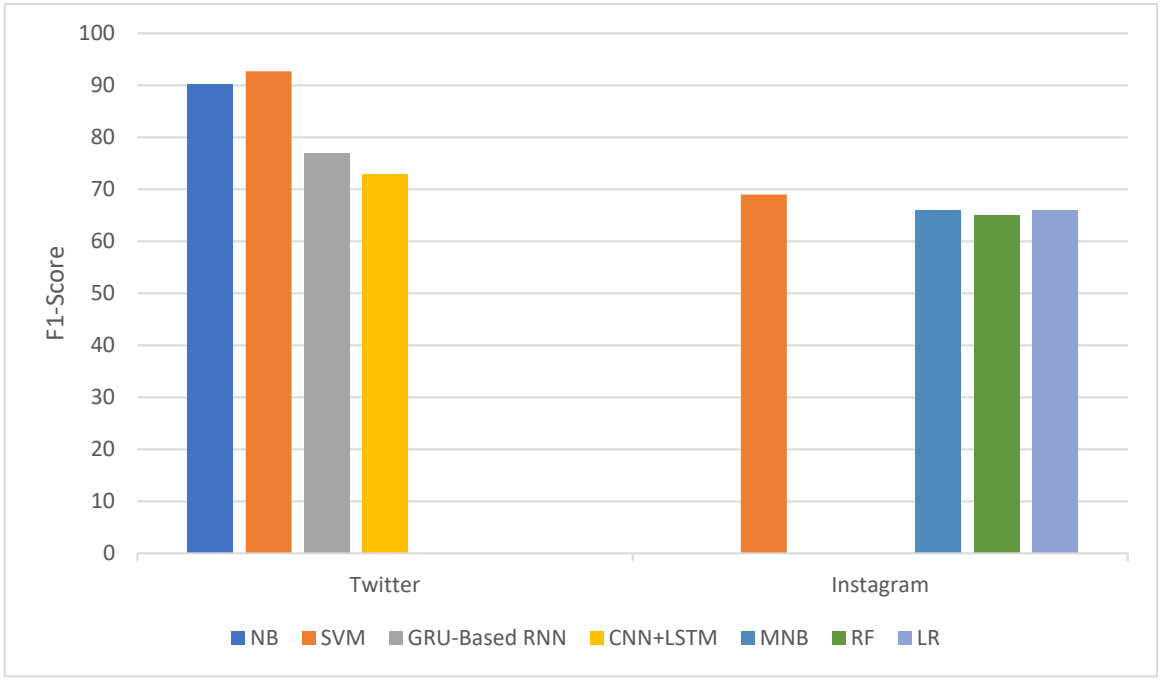

Figure 7. F1-Score of Binary Cyber-hate Classification in Arabic.

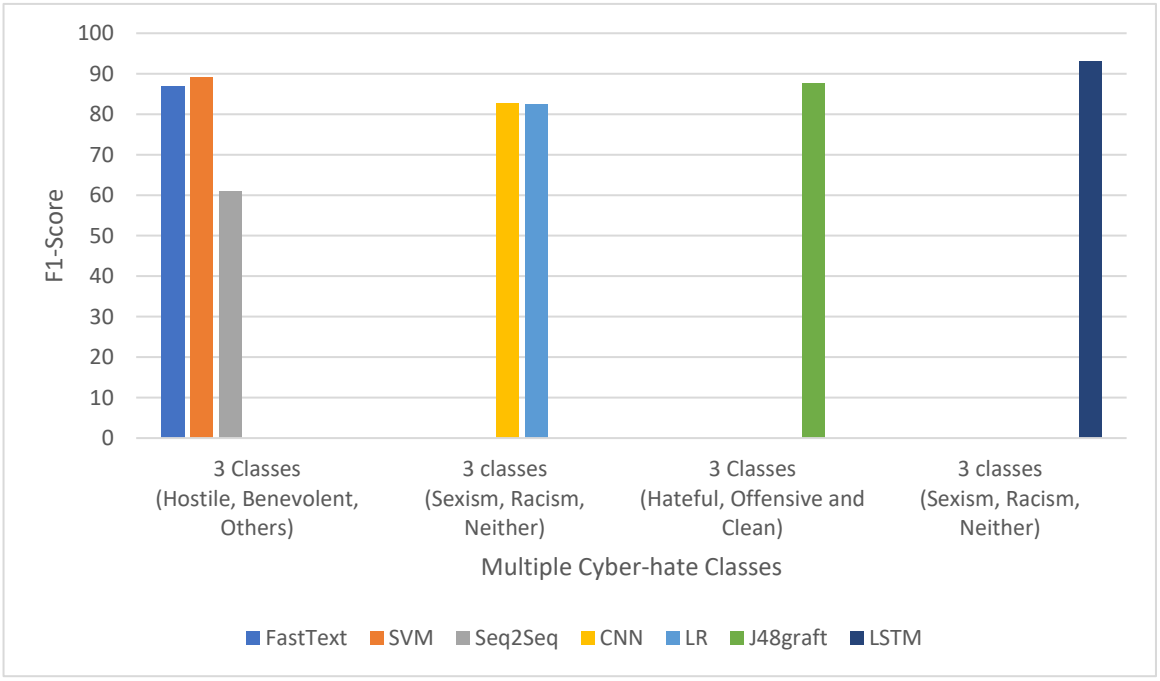

Figure 8. F1-Score of Multiple Class Cyber-Hate Classification in English.

6. Cyber-Hate Challenges

In this work, several issues were identified that affect the mainstream of the current research on cyber-hate speech detection:

- Data scarcity.
- The ambiguity of the context.
- The complexity of the Arabic language.
- Availability and accessibility to data on social networks.
- Manual Data Labelling.
- The degree of cyberbullying severity.

The field suffers from data scarcity in different languages, such as Arabic, due to the difficulty of collecting accurate cyber-hate speech data in the wild. In addition, discovering the context of a conversation is considered a challenge. The context is significant because numerous words are, in essence, ambiguous. The complexity of the Arabic language poses syntactic, semantic, and figurative ambiguity in terms of its pronunciation, vocabulary, phonetics, and morphology. This challenge could be solved by constructing an Arabic lexicon; the lexicon of offensive words may be useful in other languages to create a benchmark for the Arabic cyber-hate dataset. Current models of cyber-hate detection depend on the accessibility to accurate, relevant information from social media accounts and the experiences of potential victims. However, in actual cases, the availability of this data is influenced by consumer privacy habits and restrictions imposed by social networks. Privacy preservation is considered a challenging point. A proper solution entails the individuals' understanding of privacy preferences. Data labeling is a labor-intensive and time-consuming task, as it is necessary to select appropriate meanings of key terms that would be used during the labeling of ground truth before the process starts. The degree of cyberbullying severity is considered a challenge to be determined. Predicting various degrees of cyber-hate severity involves not only machine learning understanding but also a detailed analysis to identify and categorize the degree of cyber-hate severity from social and psychological experiences.

7. Conclusions and Future Work

In this manuscript, we have briefly reviewed the existing research on detecting cyber-hate behavior on different social media websites using various machine-learning approaches. Existing datasets of cyber-hate in different languages have been reviewed. In addition, a comparative study including binary and multiple class cyber-hate classification has been introduced, summarizing the most recent work that has been done during the last five years in different languages. Finally, the main challenges and open research issues were described in detail. Even though this research field in Arabic language is still in its early stages, existing studies confirm the importance of tackling Arabic cyberbullying detection. For future work, we aim to start with the construction of an annotated Arabic Cyber-hate dataset. Then, we will explore and apply different machine and deep learning algorithms for cyber-hate speech detection in Arabic. Future work will also include optimized real-time detection of Arabic cyberbullying.

Author Contributions: Conceptualization, D.G.; Validation, D.G., M.A. (Marco Alfonse) and S.M.J.-Z.; Formal analysis, D.G., M.A. (Marco Alfonse) and S.M.J.-Z.; Investigation, D.G.; Resources, D.G.; Writing—original draft, D.G.; Writing—review & editing, D.G., M.A. (Marco Alfonse) and S.M.J.-Z.; Visualization, D.G., M.A. (Marco Alfonse) and S.M.J.-Z.; Supervision, M.A. (Marco Alfonse), S.M.J.-Z. and M.A. (Mostafa Aref); Project administration, M.A. (Mostafa Aref). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data sharing not applicable.

Acknowledgments: This work has been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00) and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGener-

ationEU/PRTR, and Big Hug project (P20_00956, PAIDI 2020) and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government. Salud María Jiménez-Zafra has been partially supported by a grant from Fondo Social Europeo and Administración de la Junta de Andalucía (DOC_01073).

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

References

1. Alsayat, A.; Elmitwally, N. A Comprehensive Study for Arabic Sentiment Analysis (Challenges and Applications). *Egypt. Inform. J.* **2020**, *21*, 7–12. [CrossRef]
2. Available online: https://www.Statista.Com/Statistics/278414/Number-of-Worldwide-Social-Network-Users/ (accessed on 3 July 2020).
3. Available online: https://www.Oberlo.Com/Statistics/How-Many-People-Use-Social-Media (accessed on 3 July 2020).
4. Corazza, M.; Menini, S.; Cabrio, E.; Tonelli, S.; Villata, S. A Multilingual Evaluation for Online Hate Speech Detection. *ACM Trans. Internet Technol.* **2020**, *20*, 1–22. [CrossRef]
5. StopBullying.Gov. Available online: https://www.stopbullying.gov (accessed on 3 July 2020).
6. Bisht, A.; Singh, A.; Bhadauria, H.S.; Virmani, J. Detection of Hate Speech and Offensive Language in Twitter Data Using LSTM Model. *Recent Trends Image Signal Process. Comput. Vis.* **2020**, *1124*, 243–264.
7. Miro-Llinares, F.; Rodriguez-Sala, J.J. Cyber Hate Speech on Twitter: Analyzing Disruptive Events from Social Media to Build a Violent Communication and Hate Speech Taxonomy. *Des. Nat. Ecodynamics* **2016**, *11*, 406–415. [CrossRef]
8. Blaya, C. Cyberhate: A Review and Content Analysis of Intervention Strategies. *Aggress. Violent Behav.* **2018**, *45*, 163–172. [CrossRef]
9. Namdeo, P.; Pateriya, R.K.; Shrivastava, S. A Review of Cyber Bullying Detection in Social Networking. In Proceedings of the Inventive Communication and Computational Technologies, Coimbatore, India, 10–11 March 2017; pp. 162–170.
10. Hang, O.C.; Dahlan, H.M. Cyberbullying Lexicon for Social Media. In Proceedings of the Research and Innovation in Information Systems (ICRIIS), Johor Bahru, Malaysia, 2–3 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.
11. Sangwan, S.R.; Bhatia, M.P.S. Denigration Bullying Resolution Using Wolf Search Optimized Online Reputation Rumour Detection. *Procedia Comput. Sci.* **2020**, *173*, 305–314. [CrossRef]
12. Colton, D.; Hofmann, M. Sampling Techniques to Overcome Class Imbalance in a Cyberbullying Context. *Comput. Linguist. Res.* **2019**, *3*, 21–40. [CrossRef]
13. Qodir, A.; Diponegoro, A.M.; Safaria, T. Cyberbullying, Happiness, and Style of Humor among Perpetrators: Is There a Relationship? *Humanit. Soc. Sci. Rev.* **2019**, *7*, 200–206. [CrossRef]
14. Peled, Y. Cyberbullying and Its Influence on Academic, Social, and Emotional Development of Undergraduate Students. *Heliyon* **2019**, *5*, e01393. [CrossRef]
15. Dhillon, G.; Smith, K.J. Defining Objectives for Preventing Cyberstalking. *Bus. Ethics* **2019**, *157*, 137–158. [CrossRef]
16. la Vega, D.; Mojica, L.G.; Ng, V. Modeling Trolling in Social Media Conversations. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan, 7–12 May 2018; pp. 3701–3706.
17. Hassan, S.; Yacob, M.I.; Nguyen, T.; Zambri, S. Social Media Influencer and Cyberbullying: A Lesson Learned from Preliminary Findings. In Proceedings of the 9th Knowledge Management International Conference (KMICe), Miri, Sarawak, Malaysia, 25–27 July 2018; pp. 200–205.
18. Raisi, E.; Huang, B. Weakly Supervised Cyberbullying Detection Using Co-Trained Ensembles of Embedding Models. In Proceedings of the Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 28–31 August 2018; pp. 479–486.
19. Willard, N.E. *Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress*; Research Press: Champaign, IL, USA, 2007.
20. Available online: https://www.Pewresearch.Org/Internet/2021/01/13/Personal-Experiences-with-Online-Harassment/ (accessed on 3 July 2020).
21. Mangaonkar, A.; Hayrapetian, A.; Raje, R. Collaborative Detection of Cyberbullying Behavior in Twitter Data. In Proceedings of the Electro/Information technology (EIT), Dekalb, IL, USA, 21–23 May 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 611–616.
22. Van Hee, C.; Lefever, E.; Verhoeven, B.; Mennes, J.; Desmet, B.; De Pauw, G.; Daelemans, W.; Hoste, V. Detection and Fine-Grained Classification of Cyberbullying Events. In Proceedings of the Recent Advances in Natural Language Processing (RANLP), Hissar, Bulgaria, 1–8 September 2015; pp. 672–680.
23. 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, 1 June 2016; pp. 88–93.
24. Zhao, R.; Zhou, A.; Mao, K. Automatic Detection of Cyberbullying on Social Networks Based on Bullying Features. In Proceedings of the 17th International Conference on Distributed Computing and Networking, New York, NY, USA, 4 January 2016; pp. 1–6.
25. Singh, V.K.; Huang, Q.; Atrey, P.K. Cyberbullying Detection Using Probabilistic Socio-Textual Information Fusion. In Proceedings of the Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 18–21 August 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 884–887.

26. Available online: https://www.Ra.Ethz.Ch/Cdstore/Www2009/Caw2.Barcelonamedia.Org/Index.Html (accessed on 3 July 2020).
27. Al-garadi, M.A.; Varathan, K.D.; Ravana, S.D. Cybercrime Detection in Online Communications: The Experimental Case of Cyberbullying Detection in the Twitter Network. *Comput. Hum. Behav.* **2016**, *63*, 433–443. [CrossRef]
28. Hosseinmardi, H.; Rafiq, R.I.; Han, R.; Lv, Q.; Mishra, S. Prediction of Cyberbullying Incidents in a Media-Based Social Network. In Proceedings of the Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 18–21 August 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 186–192.
29. Zhang, X.; Tong, J.; Vishwamitra, N.; Whittaker, E.; Mazer, J.P.; Kowalski, R.; Hu, H.; Luo, F.; Macbeth, J.; Dillon, E. Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network. In Proceedings of the 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 18–20 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 740–745.
30. Wulczyn, E.; Thain, N.; Dixon, L. Ex Machina: Personal Attacks Seen at Scale. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3 April 2017; pp. 1391–1399.
31. Haidar, B.; Chamoun, M.; Serhrouchni, A. Multilingual Cyberbullying Detection System: Detecting Cyberbullying in Arabic Content. In Proceedings of the 1st Cyber Security in Networking Conference (CSNet), Rio de Janeiro, Brazil, 18–20 October 2017; pp. 1–8.
32. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated Hate Speech Detection and the Problem of Offensive Language. *Int. AAAI Conf. Web Soc. Media* **2017**, *11*, 512–515. [CrossRef]
33. Sprugnoli, R.; Menini, S.; Tonelli, S.; Oncini, F.; Piras, E. Creating a Whatsapp Dataset to Study Pre-Teen Cyberbullying. In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Brussels, Belgium, 31 October 2018; pp. 51–59.
34. Bartalesi Lenzi, V.; Moretti, G.; Sprugnoli, R. Cat: The Celct Annotation Tool. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 21–27 May 2012; pp. 333–338.
35. Founta, A.-M.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; Kourtellis, N. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In Proceedings of the Weblogs and Social Media (ICWSM), Palo Alto, CA, USA, 25–28 June 2018; pp. 491–500.
36. 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 2018; pp. 11–20.
37. Nurrahmi, H.; Nurjanah, D. Indonesian Twitter Cyberbullying Detection Using Text Classification and User Credibility. In Proceedings of the Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 6–7 March 2018; pp. 543–548.
38. Albadi, N.; Kurdi, M.; Mishra, S. Are They Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere. In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 28–31 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 69–76.
39. Bosco, C.; Felice, D.; Poletto, F.; Sanguinetti, M.; Maurizio, T. Overview of the Evalita 2018 Hate Speech Detection Task. *Ceur Workshop Proc.* **2018**, *2263*, 1–9.
40. Michele, C.; Menini, S.; Pinar, A.; Sprugnoli, R.; Elena, C.; Tonelli, S.; Serena, V. Inriafbk at Germeval 2018: Identifying Offensive Tweets Using Recurrent Neural Networks. In Proceedings of the Germ Eval Workshop, Vienna, Austria, 21 September 2018; pp. 80–84.
41. 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 Third Workshop on Abusive Language Online, Florence, Italy, 1–2August 2019; pp. 111–118.
42. Ptaszynski, M.; Pieciukiewicz, A.; Dybała, P. Results of the PolEval 2019 Shared Task 6: First Dataset and Open Shared Task for Automatic Cyberbullying Detection in Polish Twitter. In Proceedings of the Pol Eval 2019 Workshop, Warsaw, Poland, 31 May 2019; pp. 89–110.
43. Ibrohim, M.O.; Budi, I. Multi-Label Hate Speech and Abusive Language Detection in Indonesian Twitter. In Proceedings of the Third Workshop on Abusive Language Online, Florence, Italy, 1–2 August 2019; pp. 46–57.
44. Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Pardo, F.M.R.; Rosso, P.; Sanguinetti, M. 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.
45. Banerjee, V.; Telavane, J.; Gaikwad, P.; Vartak, P. Detection of Cyberbullying Using Deep Neural Network. In Proceedings of the 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Piscataway, NJ, USA, 15 March 2019; pp. 604–607.
46. Lu, N.; Wu, G.; Zhang, Z.; Zheng, Y.; Ren, Y.; Choo, K.R. Cyberbullying Detection in Social Media Text Based on Character-level Convolutional Neural Network with Shortcuts. *Concurr. Comput. Pract. Exp.* **2020**, *32*, 1–11. [CrossRef]
47. Moon, J.; Cho, W.I.; Lee, J. BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection. In Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media, Online, 10 July 2020; pp. 25–31.
48. Romim, N.; Ahmed, M.; Talukder, H.; Islam, S. Hate Speech Detection in the Bengali Language: A Dataset and Its Baseline Evaluation. In Proceedings of the International Joint Conference on Advances in Computational Intelligence, Singapore, 23–24 October 2021; pp. 457–468.

49. Karim, M.R.; Dey, S.K.; Islam, T.; Sarker, S.; Menon, M.H.; Hossain, K.; Hossain, M.A.; Decker, S. Deephateexplainer: Explainable Hate Speech Detection in under-Resourced Bengali Language. In Proceedings of the 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), Porto, Portugal, 6–9 October 2021; pp. 1–10.
50. Luu, S.T.; Van Nguyen, K.; Nguyen, N.L.-T. A Large-Scale Dataset for Hate Speech Detection on Vietnamese Social Media Texts. In Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Kitakyushu, Japan, 19–22 July 2021; pp. 415–426.
51. Sadiq, S.; Mehmood, A.; Ullah, S.; Ahmad, M.; Choi, G.S.; On, B.-W. Aggression Detection through Deep Neural Model on Twitter. *Futur. Gener. Comput. Syst.* **2021**, *114*, 120–129. [CrossRef]
52. Beyhan, F.; Çarık, B.; İnanç, A.; Terzioğlu, A.; Yanikoglu, B.; Yeniterzi, R. A Turkish Hate Speech Dataset and Detection System. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 4177–4185.
53. Ollagnier, A.; Cabrio, E.; Villata, S.; Blaya, C. CyberAgressionAdo-v1: A Dataset of Annotated Online Aggressions in French Collected through a Role-Playing Game. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 867–875.
54. ALBayari, R.; Abdallah, S. Instagram-Based Benchmark Dataset for Cyberbullying Detection in Arabic Text. *Data* **2022**, *7*, 83. [CrossRef]
55. Patil, H.; Velankar, A.; Joshi, R. L3cube-Mahahate: A Tweet-Based Marathi Hate Speech Detection Dataset and Bert Models. In Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022), Gyeongju, Republic of Korea, 17 October 2022; pp. 1–9.
56. Kumar, A.; Sachdeva, N. A Bi-GRU with Attention and CapsNet Hybrid Model for Cyberbullying Detection on Social Media. *World Wide Web* **2022**, *25*, 1537–1550. [CrossRef]
57. Atoum, J.O. Detecting Cyberbullying from Tweets Through Machine Learning Techniques with Sentiment Analysis. In *Advances in Information and Communication*; Arai, K., Ed.; Springer Nature: Cham, Switzerland, 2023; pp. 25–38.
58. Nabiilah, G.Z.; Prasetyo, S.Y.; Izdihar, Z.N.; Girsang, A.S. BERT Base Model for Toxic Comment Analysis on Indonesian Social Media. *Procedia Comput. Sci.* **2023**, *216*, 714–721. [CrossRef]
59. Hate Speech Twitter Annotations. Available online: https://github.com/ZeerakW/hatespeech (accessed on 9 August 2020).
60. Wikipedia Detox. Available online: https://github.com/ewulczyn/wiki-detox (accessed on 20 August 2020).
61. Used a Crowd-Sourced Hate Speech Lexicon to Collect Tweets Containing Hate Speech Keywords. We Use Crowd-Sourcing to Label a Sample of These Tweets into Three Categories: Those Containing Hate Speech, Only Offensive Language, and Those with Neither. Available online: https://arxiv.org/abs/1703.04009 (accessed on 16 February 2023).
62. WhatsApp-Dataset. Available online: https://github.com/dhfbk/WhatsApp-Dataset (accessed on 18 August 2020).
63. Hate and Abusive Speech on Twitter. Available online: https://github.com/ENCASEH2020/hatespeech-twitter (accessed on 22 August 2020).
64. Hate Speech Dataset from a White Supremacist Forum. Available online: https://github.com/Vicomtech/hate-speech-dataset (accessed on 18 November 2022).
65. Available online: https://Github.Com/Nuhaalbadi/Arabic_hatespeech (accessed on 18 November 2022).
66. L-HSAB Dataset: Context and Topics. Available online: https://github.com/Hala-Mulki/L-HSAB-First-Arabic-Levantine-HateSpeech-Dataset (accessed on 18 November 2022).
67. Dataset for Automatic Cyberbullying Detection in Polish Laguage. Available online: https://github.com/ptaszynski/cyberbullying-Polish (accessed on 15 August 2020).
68. Multi-Label Hate Speech and Abusive Language Detection in the Indonesian Twitter. Available online: https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection (accessed on 6 February 2022).
69. HatEval. Available online: http://hatespeech.di.unito.it/hateval.html (accessed on 18 November 2022).
70. BullyDataset. Available online: https://github.com/NijiaLu/BullyDataset (accessed on 6 January 2020).
71. Korean HateSpeech Dataset. Available online: https://github.com/kocohub/korean-hate-speech (accessed on 10 February 2022).
72. Available online: https://www.Kaggle.Com/Datasets/Naurosromim/Bengali-Hate-Speech-Dataset (accessed on 10 February 2022).
73. Available online: https://Github.Com/Rezacsedu/DeepHateExplainer (accessed on 10 February 2022).
74. Available online: https://Github.Com/Sonlam1102/Vihsd (accessed on 10 February 2022).
75. Available online: https://www.Kaggle.Com/Datasets/Dataturks/Dataset-for-Detection-of-Cybertrolls (accessed on 10 February 2022).
76. Available online: https://Github.Com/Verimsu/Turkish-HS-Dataset (accessed on 10 February 2022).
77. CyberAgressionAdo-v1. Available online: https://Github.Com/Aollagnier/CyberAgressionAdo-V1 (accessed on 10 February 2022).
78. Available online: https://Bit.Ly/3Md8mj3 (accessed on 10 February 2022).
79. Available online: https://Huggingface.Co/L3cube-Pune/Mahahate-Bert (accessed on 10 February 2022).
80. 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.* **2019**, *39*, 711–721. [CrossRef]
81. Alloghani, M.; Al-Jumeily, D.; Mustafina, J.; Hussain, A.; Aljaaf, A.J. *A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science*; Springer: Cham, Switzerland, 2020.
82. Alsharif, M.H.; Kelechi, A.H.; Yahya, K.; Chaudhry, S.A. Machine Learning Algorithms for Smart Data Analysis in Internet of Things Environment: Taxonomies and Research Trends. *Symmetry* **2020**, *12*, 88. [CrossRef]

83. Rout, J.K.; Dalmia, A.; Choo, K.-K.R.; Bakshi, S.; Jena, S.K. Revisiting Semi-Supervised Learning for Online Deceptive Review Detection. *IEEE Access* **2017**, *5*, 1319–1327. [CrossRef]
84. Li, Z.; Fan, Y.; Jiang, B.; Lei, T.; Liu, W. A Survey on Sentiment Analysis and Opinion Mining for Social Multimedia. *Multimed. Tools Appl.* **2019**, *78*, 6939–6967. [CrossRef]
85. Ay Karakuş, B.; Talo, M.; Hallaç, İ.R.; Aydin, G. Evaluating Deep Learning Models for Sentiment Classification. *Concurr. Comput. Pract. Exp.* **2018**, *30*, 1–14. [CrossRef]
86. Asghar, M.Z.; Khan, A.; Ahmad, S.; Qasim, M.; Khan, I.A. Lexicon-Enhanced Sentiment Analysis Framework Using Rule-Based Classification Scheme. *Peer-Rev. Open Access Sci. J. (PLoS ONE)* **2017**, *12*, e0171649. [CrossRef] [PubMed]
87. Khan, F.H.; Qamar, U.; Bashir, S. Lexicon Based Semantic Detection of Sentiments Using Expected Likelihood Estimate Smoothed Odds Ratio. *Artif. Intell. Rev.* **2017**, *48*, 113–138. [CrossRef]
88. Ahmed, M.; Chen, Q.; Li, Z. Constructing Domain-Dependent Sentiment Dictionary for Sentiment Analysis. *Neural Comput. Appl.* **2020**, *32*, 14719–14732. [CrossRef]
89. Nandhini, B.S.; Sheeba, J.I. Cyberbullying Detection and Classification Using Information Retrieval Algorithm. In Proceedings of the International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET), Tamilnadu, India, 15–16 March 2015; pp. 1–5.
90. Reynolds, K.; Kontostathis, A.; Edwards, L. Using Machine Learning to Detect Cyberbullying. In Proceedings of the 2011 10th International Conference on Machine Learning and Applications and Workshops, NW Washington, DC, USA, 18–21 December 2011; Volume 2, pp. 241–244.
91. 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, Republic and Canton of Geneva, Switzerland, 3–7 April 2017; pp. 759–760.
92. Park, J.H.; Fung, P. One-Step and Two-Step Classification for Abusive Language Detection on Twitter. In Proceedings of the First Workshop on Abusive Language Online, Vancouver, BC, Canada, 4 August 2017; pp. 41–45.
93. 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]
94. Wang, W.; Huang, J.-t.; Wu, W.; Zhang, J.; Huang, Y.; Li, S.; He, P.; Lyu, M. MTTM: Metamorphic Testing for Textual Content Moderation Software. In Proceedings of the International Conference on Software Engineering (ICSE), Lisbon, Portugal, 14–20 May 2023; pp. 1–13.
95. Roy, P.K.; Tripathy, A.K.; Das, T.K.; Gao, X.-Z. A Framework for Hate Speech Detection Using Deep Convolutional Neural Network. *IEEE Access* **2020**, *8*, 204951–204962. [CrossRef]
96. Yadav, A.; Vishwakarma, D.K. Sentiment Analysis Using Deep Learning Architectures: A Review. *Artif. Intell. Rev.* **2020**, *53*, 4335–4385. [CrossRef]