Analysis and Classification of Abusive Textual Content Detection in Online Social Media

Ovais Bashir Gashroo and Monica Mehrotra

Abstract With every passing day, the amount in which social media content is being produced is enormous. This contains a large amount of data that is abusive. Which in turn is responsible for disturbing the peace, affecting the mental health of users going through that kind of data and in some cases is responsible for causing riots and chaos leading to loss of life and property. Knowing the sensitivity of such situations, it becomes necessary to tackle the abusive textual content. This paper presents the analysis of abusive textual content detection techniques. And for these, researchers have been developing methods to automatically detect such content present on social media platforms. This study also discusses the domains of automatic detection that have been investigated utilizing various methodologies. Several machine learning techniques that have recently been implemented to detect abusive textual material have been added. This paper presents a detailed summary of the existing literature on textual abusive content detection techniques. This paper also discusses the categorization of abusive content presented in the research, as well as potential abusive communication methods on social media. Deep learning algorithms outperform previous techniques; however, there are still significant drawbacks in terms of generic datasets, feature selection methods, and class imbalance issues in conjunction with contextual word representations.

Keywords Abusive textual content · Social media · Detecting abusive content · Machine learning

1 Introduction

As we have entered the third decade of the twenty-first century, the advancements in technology are growing with every passing day. Among these advancements are social media which is one of the most popular and widely used technology. This has revolutionized the way people communicate with each other. Before its inception,

O. B. Gashroo (✉) · M. Mehrotra
Department of Computer Science, Jamia Millia Islamia, New Delhi, India
e-mail: ovais1910426@st.jmi.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

G. Rajakumar et al. (eds.), *Intelligent Communication Technologies and Virtual Mobile Networks*, Lecture Notes on Data Engineering and Communications Technologies 131,
https://doi.org/10.1007/978-981-19-1844-5_15

there were limited ways for humans to be able to communicate with other humans. These platforms are not now restricted to only being used for sharing messages between each other. They have become the largest and most commonly used source for information sharing [1] such as news, thoughts, and feelings in the form of text, images, and videos. People can use these platforms for the dissemination of all kinds of content available at their disposal. These platforms are also used to express opinions about issues, products, services, etc. This ability, when combined with the speed with which online content spreads, has elevated the value of the opinions expressed [2]. Acquiring sentiment in user opinion is crucial since it leads to a more detailed understanding of user opinion [3]. The worldwide shutdown caused by the COVID-19 pandemic resulted in a tremendous rise in social media communication. As a result, a huge volume of data has been generated, and analysis of this data will assist companies in developing better policies that will eventually help make these platforms safer for their users.

The global digital population is increasing every passing day. As per [4], the active Internet users globally were 4.66 billion as of October 2020. Out of them, almost 4.14 billion were active social media users. According to another report [5], the rate of social penetration reached almost 49% in 2020. It is expected that the number of global monthly active users of social media will reach 3.43 billion by 2023, which is around one-third of the total earth's population. Among all other social media platforms, Facebook is the most popular [5]. It became the first social media platform to surpass the 1 billion monthly active user mark. In 2020, it had almost 2.6 billion monthly active users globally, the highest among all the social media platforms. Among all the countries, India has over 680 million active Internet users. As of 2020, India is the country with a 300 million user-base of Facebook, the highest among all other countries. It is reported that on average, a normal Internet user spends almost 3 h/day on social media in India. From 326 million users of social media platforms in 2018, it is estimated that it will reach almost 450 million users by 2023 in India [6]. In the European Union (EU), 80% of people have experienced hate speech online, and 40% have felt assaulted or endangered as a result of their use of social media platforms [7]. According to research conducted by Pew Research Centre in 2021 [8], "about four-in-ten Americans (41%) have experienced some form of online harassment." Online abusive content is a very serious matter. Online extremist narratives have been linked to heinous real-world occurrences like hate crimes, mass shootings like the one in Christchurch in 2019, assaults, and bombings; and threats against prominent people [9].

The abusive textual content on social media platforms is a persistent and serious problem. The presence of abusive textual content in the user's life increases their stress and dissatisfaction. The effects of such content can be very adverse with time. The analysis and evaluation of this generated data reveal important details about the people who provided it [10]. The use of emotion recognition techniques on this data will aid in stress and mental health management. This paper is intended to assist new researchers in gaining an overall perspective of this research area by providing an overview to gain insights into this field of study. The paper is organized as follows. Firstly, in Sect. 2, the abusive textual content is defined, its overall impacts on the

users. Section 3 will discuss the latest popular approaches proposed for this task. A brief description of datasets, techniques, and results obtained is also in Sect. 3. Different ways of communication are discussed in Sect. 4. Various limitations and gaps of the approaches will be discussed in the Sect. 5 followed by Sect. 6 that will present the challenges, future scope, and conclusion.

2 Defining Abusive Textual Content

The word Abuse according to the Washington state department of social and health services [11], "covers many different ways someone may harm a vulnerable adult." The state department has categorized abuse into seven different types based on different ways it can be taken advantage of to harm anyone. One among all those types of abuse was identified that can be inflicted using social media and that is "Mental mistreatment or emotional abuse." Other types of abuse require the physical presence of both the abuser and the victim in some way or another way. They define mental mistreatment or emotional abuse as "deliberately causing mental or emotional pain. Examples include intimidation, coercion, ridiculing, harassment, treating an adult like a child, isolating an adult from family, friends, or regular activity, use of silence to control behavior, and yelling or swearing which results in mental distress."

We are living in a world where causing someone mental and emotional pain does not require physical presence. As discussed in the section above, social media has become an alternate platform for communication [12]. Almost all the examples given in the definition of abuse above can be put into practice with the use of these social media platforms. The kind of effect this type of content has on humans is very dangerous in terms of mental and emotional health. Physical pain usually fades after a while, but emotional pain can linger for a long time and has a lot of serious consequences on one's mental health [13].

Abusive content is a broad term, and it covers a lot of different types of content. So, the next subsection makes a clear distinction among the different types of online abuse that are present on social media platforms.

2.1 Classification of Online Abusive Content

Online abusive content can be categorized based on the presence of abusive language, aggression, cyberbullying, insults, personal attacks, provocation, racism, sexism, or toxicity in it (Fig. 1). Based on the classification, it can be said that any social media content containing any kind of language or expression that fall into the category of these types will be abusive. Researchers have been very concerned about this problem of abusive content dissemination and have been devising methods to control its inception, to stop its further spreading and most importantly to come up with methods that can detect abuse present in disguise.

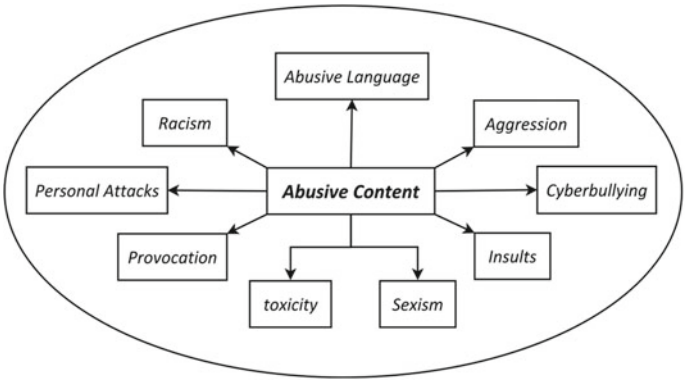

Fig. 1 Classification of abusive content

The abusive content has been classified into 9 types. The presence of any such type of textual content in online social media will be termed abusive content. Table 1 depicts the classification of abusive content and the difference among these types can be understood by the examples from social media platforms for each category. In the context of online discussions, the above highlighted types are defined as shown in Table 2. For each of the type, there are 3 examples given in Table 1. The definitions make every type distinct as every kind is expressed differently in online social media. Researchers have been developing methods to tackle mainly these types of abuse present on social media.

As per our knowledge, there is no definition of abusive content in this field of research to date. We define abusive content as:

The presence of an individual or a combination of abusive language, aggressive, bullying, insulting, sexist, toxic, provocative, personal attacking, and racist remarks in any type of social media content, that has the potential of causing mental and psychological harm to users.

3 Approaches for Detecting Textual Abusive Content

This section provides details about the different approaches employed by researchers to combat abusive content on social media sites. Table 4 gives a detailed description of the methods used for detecting abusive content on social media. The table also contains information about the datasets used, the platforms from which datasets have been obtained, and the ways of data classification. It is important to highlight that the trend in the field of abusive content detection is going toward multi-class classification rather than binary classification. The researchers have used datasets from multiple sources. In the overall review, it was found that Twitter data has been used more frequently to classify text. Also, gaming platform data, online blogs, magazines,

Table 1 Exemplar messages of each abusive content category

S. No.	Example	Ref.	Abusive content category
1	"Violently Raping Your Friend Just for Laughs"	[14]	Abusive language
2	"Kicking your Girlfriend in the Fanny because she won't make you a Sandwich,"	[14]	Abusive language
3	"#GolpearMujeresEsFelicidad" ("Beating Women Is Happiness")	[15]	Abusive language
4	"You should kill yourself"	[16]	Aggression
5	"No one likes you,"	[16]	Aggression
6	"You're such a slut,"	[16]	Aggression
7	"Ha, you're a bitch. I hope you get Crocs for your birthday."	[17]	Cyberbullying
8	"Hey, do the world a favor and go kill yourself."	[18]	Cyberbullying
9	"You don't have the balls to act like a man,"	[18]	Cyberbullying
10	"Please, give it a little thought…. oh wait, you don't have any!"	[19]	Insults
11	"Shut up with your non purposeful existant, you are just wasting oxygen!"	[19]	Insults
12	"Buddy you sound like a complete clown on here running your mouth"	[19]	Insults
13	"You are a woman why do you want to talk about football?"	[20]	Sexism
14	"The place of a woman in modern society is clear, it is in the kitchen"	[20]	Sexism
15	"Theresa May succeeds David Cameron. No better at cleaning than a woman."	[20]	Sexism
16	"Death to liberals"	[21]	Toxicity
17	"Russia is #1 terrorist nation"	[21]	Toxicity
18	"Youre a god damn idiot!!"	[21]	Toxicity
19	"You can rest assured I and the rest of the world are pleased your piece of shit family member is dead and rotting in the ground"	[22]	Provocation
20	"We laugh at your suffering and think it's pathetic you are upset because your family member was an insignificant worm"	[22]	Provocation
21	"Get a life you bored faggots."	[22]	Provocation
22	"You can go die"	[23]	Personal Attacks
23	"You are stupid"	[23]	Personal Attacks
24	"Shut up" "You Suck"	[23]	Personal Attacks

(continued)

Table 1 (continued)

S. No.	Example	Ref.	Abusive content category
25	"It is the foreigners that elicit hate and racism from natives"	[24]	Racism
26	"You can't help the fact that you belong to the race that has less intellect and sense in their brains than the smelly behind of a PIG!"	[24]	Racism
27	"Once again we have to put up with the filthiest scum, it doesn't even surprise me anymore!"	[24]	Racism

Table 2 Definitions of different types of abusive content

Types of abusive content	Definition
Abusive language	"Using rude and offensive words" [25]
Aggression	"Spoken or physical behavior that is threatening or involves harm to someone or something" [26]
Cyberbullying	"The activity of using the Internet to harm or frighten another person, specially by sending them unpleasant messages" [27]
Insults	"Insulting, inflammatory, or negative comments toward a person or a group of people." [28]
Sexism	"The belief that the members of one sex are less intelligent, able, skillful, etc. than the members of the other sex, especially that women are less able than men" [29]
Toxicity	"A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion" [28]
Provocation	"An action or statement that is intended to make someone angry" [30]
Personal attacks	"An intentionally offensive remark about someone's character or appearance" [31]
Racism	"Harmful or unfair things that people say, do, or think based on the belief that their own race makes them more intelligent, good, moral, etc. than people of other races" [32]

newspapers, and Wikipedia generated datasets have been also used. Table 3 lists the numerous abbreviations and their complete form used throughout this article in order to improve readability and maintain the uniformity of the many notations utilized.

Papegnies et al. [23] has classified the dataset into abusive and non-abusive messages depicting binary classification. A dataset containing users' in-game messages from a multiplayer online game has been used. First-stage naïve Bayes classifier for performing the task of detecting abusive and non-abusive messages using content-based features is used in the paper. Chen et al. [33] used 9 datasets containing data from multiple platforms like YouTube and Myspace. Support vector machine, convolutional neural network, and recurrent neural network were applied to detect abusive content. Their results reveal that the SVM classifier achieved the

Table 3 Abbreviations and their full forms

Abbreviation	Full Form	Abbreviation	Full form
CNN	Convolutional Neural Network	LR	Linear Regression
GRU	Gated Recurrent Units	TF-IDF	Term Frequency-inverse Document Frequency
BERT	Bidirectional Gated Recurrent Unit Network	BOW	Bag-of-Word
PoS tag	Part-of-speech Tagging	LSTM	Long Short-term Memory
RNN	Recurrent Neural Network	NB	Naïve Bayes
SVM	Support Vector Machine	RoBERTa	Robustly Optimized BERT Pertaining Approach
RNN	Recurrent Neural Network	LDA	Latent Dirichlet allocation
RBF	Radial Basis Function	MLP	Multilayer Perceptron

best results in terms of average recall on balanced datasets and deep learning models performed well on extremely unbalanced datasets. The latest research is incorporating a variety of datasets from multiple platforms. In [34], logistic regression has been set as baseline model and CNN-LSTM and BERT-LSTM have been implemented on a combination of 6 datasets containing more than 60 K records of twitter to classify data into 3 classes. They demonstrated that BERT has the highest accuracy among all models. Table 4 describes other techniques implemented; their results are also shown along with the feature extraction techniques/features used for various machine learning models. After comparing the performance of all these techniques, deep learning models are more effective in classifying abusive text. These approaches have outperformed other existing approaches for text-based classification [35].

One of the most difficult tasks when using machine learning is choosing the proper features to solve a problem. The terms textual features and content features were used interchangeably by the researchers. Textual features include Bag-of-Words (BoW), TF-IDF, N-grams, and so on. Part-of-speech (POS) tagging, and dependency relations are two syntactic features that have also been used. Traditional feature extraction methods such as the Bag-of-Words model and word embedding implemented with word2vec, fastText, and Glove were used. In natural language processing, BoW with TF-IDF is a traditional and simple feature extraction method, and Word embedding represents the document in a vector space model. Unlike the Bag-of-Words model, it captures the context and semantics of a word. Word embedding keeps word contexts and relationships intact, allowing it to detect similar words more accurately. Available literature has shown that word embeddings used with deep learning models outperform traditional feature extraction methods. Ahammad et al. [36] showed that long short-term memory (LSTM) and Gated recurrent unit (GRU) give better accuracy compared to others on trained embedding and Glove, respectively. Table 4 lists various feature extraction techniques/features used by researchers.

To deal with the problem of abusive textual content detection, researchers have presented numerous machine learning algorithms and their variants. A lot of research

Table 4 Comparative Analysis

Ref.	Feature extraction techniques/features	Results	Platform	Datasets used	Techniques used	Limitations	Classification of data
[23]	Content-based features (morphological, language, and context features)	Unbalanced data: (full feature set with advanced preprocessing) Precision = 68.3 Recall = 76.4 F-measure = 72.1 Balanced data: (full feature set with advanced preprocessing) Precision = 76.1 Recall = 76.9 F-measure = 76.5	SpaceOrigin (a multiplayer online game)	Database of users' in-game interactions (Containing total 40,29,343 messages)	Developed a system to classify abusive messages from an online community Developed on top of a first-stage naïve Bayes classifier	The performance is insufficient to be used directly as a fully automatic system, completely replacing human moderation	Abusive messages and non-abusive messages
[33]	n-Grams word vectors	Based on the results across 9 datasets, they showed the SVM classifier achieved the best results in terms of average recall on balanced datasets. And the average recall results by deep learning models performed well on extremely imbalanced datasets	YouTube, myspace, formspring, congregate, Slashdot, general news platform	9 datasets from various social media platforms like Twitter, YouTube, myspace, formspring, congregate, Slashdot, general news platform	SVM, CNN, and RNN	Precision, Accuracy, and F1 score not evaluated. And excessive over-sampling and under-sampling used	Abusive content detection

(continued)

Table 4 (continued)

Ref.	Feature extraction techniques/features	Results	Platform	Datasets used	Techniques used	Limitations	Classification of data
[37]	Generic linguistic features (character-, word-, sentence-, dictionary-, syntactic-, lexical and discourse-based characteristics) and embedded word features (google word embeddings) and BoW	The linguistic (stylistic) features are better than the baseline in terms of accuracy, while 92.14% accuracy for embedded lexical features captures the best distinction between all types of extremist material and non-extremist material	Generic extremist material, Jihadist Magazines, and white supremacist forum	1. Their heterogeneous corpus containing 1744 texts from different sources 2. Publicly available corpus of 2685 quotes from 2 English magazines 3. Corpus containing 2356 complete threads from an online forum	SVM	Small-sized dataset used	Classification of extremist online material (offensive and non-offensive, jihadist and non-jihadist, white supremacist, and non-white supremacist)
[38]	Pre-trained word vectors also known as word embedding or distributed representation of words	Results show that CNN outperforms all other models Accuracy = 0.762 Precision = 0.774 Recall = 0.740 F1 = 0.756 AUC = 0.762	Twitter	10 k tweets were collected using Twitter's streaming API	CNN	Labeling tweets based on keywords only	Identifying misogynistic abuse

(continued)

Table 4 (continued)

Ref.	Feature extraction techniques/features	Results	Platform	Datasets used	Techniques used	Limitations	Classification of data
[39]	Bag of n-Grams, Word2Vec, Doc2Vec, fastText, Linguistic features	For multi-class, LR is the top-performing classification algorithm For multi-label, no classifier outperforms the remaining ones Word2Vec, Doc2Vec, and fastText deliver the best results for both multi-class & multi-label	Online comment sections of Newspaper portals	A dataset containing 521,954 comments from newspapers	Logistic regression (extended multinomial version) for the multi-class case Gradient Boosting for multi-class, Random Forest for both multi-class & multi-label, Binary relevance for the multi-label, classifier chains for multi-label	The dataset contains comments from articles on a single topic	Identification of Threats, Hate speech, Insult, Inappropriate language, and normal comments
[40]	FastText, Doc2Vec, word2vec, word-level embeddings	FastText features gave the highest accuracy of 85.81% using SVM-RBF, followed by word2vec with 75.11% accuracy using SVM-RBF and finally doc2vec with 64.15% accuracy using random forest	Twitter	10,000 texts from Twitter	SVM, SVM-RBF, Random Forest	The classification is into only 2 types	Binary classification of Hindi–English tweets to hate and no-hate

(continued)

Table 4 (continued)

Ref.	Feature extraction techniques/features	Results	Platform	Datasets used	Techniques used	Limitations	Classification of data
[41]	PoS tag, BoW	LDA works well considering 10 topics Log-likelihood = −895,399.1204 Perplexity = 1290.9525 Sparsity = 0.17290%	Twitter	2400 + tweets dataset	LDA, self-organizing maps (unsupervised ML), K-means clustering	Small-sized dataset used	Slurs, offensive language, abuses, violent and racist remark detection
[34]	BoW or TF-IDF with a regression-based model	BERT has the highest accuracy among all models	Twitter	Combination of 6 datasets containing 73,810 records	Logistic regression as the baseline model, CNN-LSTM, BERT-LSTM	The proposed models did not outperform the state-of-the-art models	Classification of data into abusive, hateful, and neither
[36]	n-Gram, TF-IDF, Trained word embedding, Pre-trained GloVe	In comparison to others, SVM shows the best Accuracy of 83% on TF-IDF, LSTM shows the best Accuracy of 84% on trained embedding and GRU shows the best Accuracy of 84% on GloVe	Twitter	9,787 publicly available users' tweet	Naïve Bayes, SVM, Random Forest, Logistic regression, MLP, LSTM, GRU	Less train and test data used. And imbalanced class distributions not used to test the performance of classifiers in experiments	Abusive behavior detection approach to identify hatred, violence, harassment, and extremist expressions

(continued)

Table 4 (continued)

Ref.	Feature extraction techniques/features	Results	Platform	Datasets used	Techniques used	Limitations	Classification of data
[42]	FastText embeddings	CNN performed better classification than the LSTM Accuracy = 99.4% F1 Score = 99.4% Recall = 99.7% Precision = 99.6%	Ask. fm website Formspring. me, Twitter (Olid, warner, and Waseem) Wikipedia	5 publicly available datasets AskFm corpus, Formspring dataset, Warner and Waseem dataset, Olid, and Wikipedia toxic comments dataset	Deep learning-based method (CNN/LSTM)	Determination of the best data augmentation approach is critical	Detect hate speech and cyberbullying content
[43]	Word2Vec, fastText, Glove	The framework improves the net F1 score by 7.1%, 5.6%, and 2.7% in the attack, aggressive, and toxicity detection	Wikipedia	A multi-wiki dataset containing 77,972 Wikipedia comments	CNN	High cost of computational complexity required for better results	Detection of attack, aggression, and toxicity

in this area is focused on extracting features from text. Many of the proposed works make use of text feature extraction techniques like BOW (Bag-of-Words) and dictionaries. It was discovered that these traits were unable to comprehend the context of phrases. N-gram-based approaches outperform their counterparts in terms of results and performance [14].

4 Abusive Content in Social Media

The content that is being shared over social media platforms can be divided in many ways (Fig. 2). The classification is discussed below. Possible cases of abusive content spread in OSN's (how can social media platforms be used to spread abusive content):

1. One to one (individual to individual): the attacker is an individual and the victim is also an individual. E.g.: personal text messages on social media platforms can be used to spread abusive content.
2. One to many (individual to the community): the attacker is one and the victim is a community. E.g.: social media platforms are used by individuals to post and sometimes it is exploited to target a community with the abusive content.
3. Many to one (community to individual): the attacker is a community targeting an individual. E.g.: a community posting or spreading abusive content about an individual on social media.
4. Many to many (community to community): a community targeting another community over issues and posting and spreading abusive content on social media platforms. E.g.: persons belonging to a community posting abusive content and targeting other communities through their content is also a kind through which abusive content is shared and disseminated.

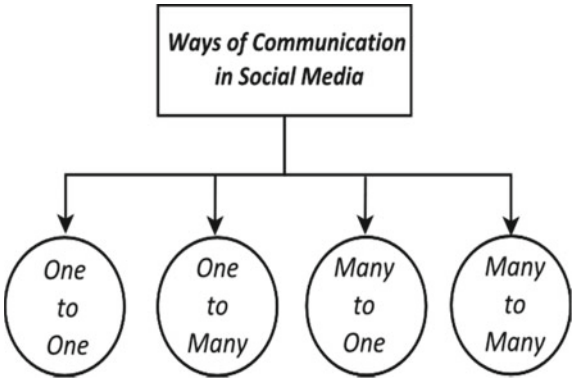

Fig. 2 Ways of communication in OSN's

5 Discussion

The approaches used to classify the textual abusive content which are discussed in Sect. 3 of this paper, when viewed critically, have their own set of limitations and gaps. The researchers will be motivated to address these issues and develop effective methods by highlighting the limitations and gaps in previous research. In terms of practical application, the performance of [23] is insufficient to be used as a fully automated system that replaces human moderation. When using machine learning techniques, performance metrics such as precision, accuracy, and F1 score provide useful information. These metrics were not evaluated in [33]. Many researchers have also used over-sampling and under-sampling throughout the study. The datasets in [33] were also subjected to these two methods. They should, however, be used with caution on datasets because excessive use of either can result in over-fitting and the loss of important information from the datasets. The authors of [42] used data augmentation and determining the best data augmentation approach is critical because if the original dataset has biases, data-enhanced from it can have biases as well. Large datasets are preferred, but small datasets are used in [37, 41], and small datasets have an impact on the machine learning model's performance, according to [44]. Because small datasets typically contain fewer details, the classification model is unable to generalize patterns learned from training data. Furthermore, because over-fitting can sometimes extend beyond training data and affect the validation set as well, it becomes much more difficult to avoid. Parametric tuning was also discovered to have produced better results in [41]. In [38], tweets are labeled solely based on keywords, resulting in the omission of tweets containing abuse and harassment in plain language that does not contain any of the authors' keywords. A dataset that only includes comments from articles on a single topic has been used in [39]. It influences the diversity of a dataset, which makes the performance of the various techniques less significant. It is also worth noting that researchers are classifying data into a variety of categories, making the problem of abusive content classification a multi-class problem. A concept for classifying abusive content was also included in Sect. 2 of this paper. Most of the models chosen for this study used multi-class classification, but [23, 40] classified the data using binary classification. The performance of the implemented models is critical in highlighting their significance. In terms of performance, the proposed models in [34] did not outperform the state-of-the-art models RoBERTa and XLM-R. It is also beneficial to consider using more train and test data to improve model performance. The researchers used less train and test data in [36], and did not use imbalanced class distributions to evaluate classifier performance in experiments. The models used in [43] have a high computational complexity cost to achieve better results. The above-discussed limitations and gaps may provide future directions for the researchers in this field.

6 Conclusion and Future Scope

This paper offered a complete overview of the topic by combining recent research articles that used cutting-edge approaches. The researchers employed machine learning approaches successfully, with Bag-of-Words (BoW) and N-grams being the most commonly used features in classification. Several recent research has adopted distributed word representations, also known as word embeddings, because of previous models' high dimensionality and sparsity. Deep learning-based architectures have lately shown promising outcomes in this discipline. Despite being introduced in 2019, BERT has become widely used. The most popular deep learning models being used are LSTM and CNN. Moreover, hybrid models, which are a combination of multiple models, such as BERT + CNN, LSTM + CNN, LSTM + GURU, and BERT + LSTM, have also been used in the research, and their performance in detecting online abuse is promising. To summarize, deeper language traits, demographic influence analyses, and precise annotation criteria are required to effectively discern between different types of abuse. Despite the abundance of work accessible, judging the usefulness and performance of various features and classifiers remains difficult because each researcher used a different dataset. For comparison evaluation, clear annotation criteria and a benchmark dataset are necessary. According to the findings, abusive content detection remains a research area of interest needing more intelligent algorithms to handle the primary issues involved and make online interaction safer for users [45].

In this paper, the approach for identifying the recent works in abusive content detection was limited to only those dealing with the textual form of data present on social media platforms in the form of tweets, comments, chats, reviews, blogs, etc. Only those research papers that used English language datasets were chosen. Work on detecting abusive content in languages other than English is also present. But due to the scarcity of datasets, the results are not that effective. Also, out of 9 different types of abusive content which were discussed in section II, research is focused only on a subset of these types. Research to classify all the different types from a corpus is the need of the hour and it will encourage the budding researchers to take up the task. So, the future work of this study will be to identify cross-language and cross-domain techniques being developed and analyze their performance with the already present state-of-the-art approaches.

References

1. Osatuyi B (2013) Information sharing on social media sites. Comput Human Behav 29. https://doi.org/10.1016/j.chb.2013.07.001
2. Pandian AP (2021) Performance evaluation and comparison using deep learning techniques in sentiment analysis. J Soft Comput Paradig 3. https://doi.org/10.36548/jscp.2021.2.006
3. Tripathi M (2021) Sentiment analysis of Nepali COVID19 tweets using NB, SVM AND LSTM. J Artif Intell Capsul Networks 3. https://doi.org/10.36548/jaicn.2021.3.001
4. Global digital population as of October 2020. https://www.statista.com/statistics/617136/digital-population-worldwide/
5. Social media usage in India—Statistics & Facts. https://www.statista.com/topics/5113/social-media-usage-in-india/
6. Social media—Statistics & Facts. https://www.statista.com/topics/1164/social-networks/
7. Gagliardone I, Gal D, Alves T, Martinez G (2015) Countering online hate speech. UNESCO Publishing
8. Vogels EA, The state of online harassment. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/. Last accessed 14 Jan 2022
9. Castaño-Pulgarín SA, Suárez-Betancur N, Vega LMT, López HMH (2021) Internet, social media and online hate speech. Syst Rev. https://doi.org/10.1016/j.avb.2021.101608
10. Kottursamy K (2021) A review on finding efficient approach to detect customer emotion analysis using deep learning analysis. J Trends Comput Sci Smart Technol 3. https://doi.org/10.36548/jtcsst.2021.2.003
11. Types and signs of abuse, https://www.dshs.wa.gov/altsa/home-and-community-services/types-and-signs-abuse
12. Carr CT, Hayes RA (2015) Social media: defining, developing, and divining. Atl J Commun 23. https://doi.org/10.1080/15456870.2015.972282
13. Gavin H (2011) Sticks and stones may break my bones: the effects of emotional abuse. J Aggress Maltreatment Trauma 20. https://doi.org/10.1080/10926771.2011.592179
14. Nobata C, Tetreault J, Thomas A, Mehdad Y, Chang Y (2016) Abusive language detection in online user content. In: 25th international world wide web conference, WWW 2016. https://doi.org/10.1145/2872427.2883062.
15. Do social media platforms really care about online abuse? https://www.forbes.com/sites/kalevleetaru/2017/01/12/do-social-media-platforms-really-care-about-online-abuse/?sh=659f775f45f1
16. Young R, Miles S, Alhabash S (2018) Attacks by anons: a content analysis of aggressive posts, victim responses, and bystander interventions on a social media site. Soc Media Soc 4. https://doi.org/10.1177/2056305118762444
17. Whittaker E, Kowalski RM (2015) Cyberbullying via social media. J Sch Violence 14. https://doi.org/10.1080/15388220.2014.949377
18. Hosseinmardi H, Mattson SA, Rafiq RI, Han R, Lv Q, Mishra S (2015) Analyzing labeled cyberbullying incidents on the instagram social network. In: Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and lecture notes in bioinformatics) (2015). https://doi.org/10.1007/978-3-319-27433-1_4
19. Detecting insults in social commentary. https://www.kaggle.com/c/detecting-insults-in-social-commentary/data
20. Chiril P, Moriceau V, Benamara F, Mari A, Origgi G, Coulomb-Gully M (2020) An annotated corpus for sexism detection in French tweets. In: LREC 2020—12th international conference on language resources and evaluation, conference proceedings
21. Obadimu A, Mead E, Hussain MN, Agarwal N (2019) Identifying toxicity within Youtube video comment. In: Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and lecture notes in bioinformatics). https://doi.org/10.1007/978-3-030-21741-9_22
22. McCosker A (2014) Trolling as provocation: YouTube's agonistic publics. Convergence 20. https://doi.org/10.1177/1354856513501413

23. Papegnies E, Labatut V, Dufour R, Linarès G (2018) Impact of content features for automatic online abuse detection. In: Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and lecture notes in bioinformatics). https://doi.org/10.1007/978-3-319-77116-8_30
24. Tulkens S, Hilte L, Lodewyckx E, Verhoeven B, Daelemans W (2016) The automated detection of racist discourse in Dutch social media. Comput. Linguist. Neth. J.
25. Cambridge: abusive. https://dictionary.cambridge.org/dictionary/english/abusive. Last accessed 14 Jan 2022
26. Cambridge: aggression. https://dictionary.cambridge.org/dictionary/english/aggression. Last accessed 14 Jan 2022
27. Cambridge: cyberbullying, https://dictionary.cambridge.org/dictionary/english/cyberbullying. Last accessed 14 Jan 2022
28. Jigsaw LLC (2020) Perspective API FAQs: what is perspective? https://support.perspectiveapi.com/s/about-the-api-faqs. Last accessed 14 Jan 2022
29. Cambridge: sexism, https://dictionary.cambridge.org/dictionary/english/sexism. Last accessed 14 Jan 2022
30. Cambridge: provocation. https://dictionary.cambridge.org/dictionary/english/provocation. Last accessed 14 Jan 2022
31. Cambridge: personal attacks. https://dictionary.cambridge.org/dictionary/english/personal. Last accessed 14 Jan 2022
32. Cambridge: racism, https://dictionary.cambridge.org/dictionary/english/racism. Last accessed 14 Jan 2022
33. Chen H, McKeever S, Delany SJ (2018) A comparison of classical versus deep learning techniques for abusive content detection on social media sites. In: Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and lecture notes in bioinformatics). https://doi.org/10.1007/978-3-030-01129-1_8
34. Vashistha N, Zubiaga A (2021) Online multilingual hate speech detection: experimenting with Hindi and English social media. Information 12 (2021). https://doi.org/10.3390/info12010005
35. Kompally P, Sethuraman SC, Walczak S, Johnson S, Cruz MV (2021) Malang: a decentralized deep learning approach for detecting abusive textual content. Appl Sci 11. https://doi.org/10.3390/app11188701
36. Ahammad T, Uddin MK, Yesmin T, Karim A, Halder S, Hasan MM (2021) Identification of abusive behavior towards religious beliefs and practices on social media platforms. Int J Adv Comput Sci Appl 12. https://doi.org/10.14569/IJACSA.2021.0120699
37. Soler-Company J, Wanner L (2019) Automatic classification and linguistic analysis of extremist online material. In: Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and lecture notes in bioinformatics). https://doi.org/10.1007/978-3-030-05716-9_49
38. Bashar MA, Nayak R, Suzor N, Weir B (2019) Misogynistic tweet detection: modelling CNN with small datasets. In: Communications in computer and information science. https://doi.org/10.1007/978-981-13-6661-1_1
39. Niemann M (2019) Abusiveness is non-binary: five shades of gray in German online newscomments. In: Proceedings—21st IEEE conference on business informatics, CBI 2019. https://doi.org/10.1109/CBI.2019.00009
40. Sreelakshmi K, Premjith B, Soman KP (2020) Detection of HATE Speech text in Hindi-English code-mixed data. Procedia Comput Sci. https://doi.org/10.1016/j.procs.2020.04.080
41. Saini Y, Bachchas V, Kumar Y, Kumar S (2020) Abusive text examination using Latent Dirichlet allocation, self organizing maps and k means clustering. In: Proceedings of the international conference on intelligent computing and control systems, ICICCS 2020. https://doi.org/10.1109/ICICCS48265.2020.9121090
42. Beddiar DR, Jahan MS, Oussalah M (2021) Data expansion using back translation and paraphrasing for hate speech detection. Online Soc Networks Media 24. https://doi.org/10.1016/j.osnem.2021.100153

43. Zhao Q, Xiao Y, Long Y (2021) Multi-task CNN for abusive language detection. In: 2021 IEEE 2nd international conference on pattern recognition and machine learning, PRML 2021, pp 286–291. https://doi.org/10.1109/PRML52754.2021.9520387
44. Althnian A, AlSaeed D, Al-Baity H, Samha A, Dris AB, Alzakari N, Abou Elwafa A, Kurdi H (2021) Impact of dataset size on classification performance: an empirical evaluation in the medical domain. Appl Sci 11. https://doi.org/10.3390/app11020796
45. Kaur S, Singh S, Kaushal S (2021) Abusive content detection in online user-generated data: a survey. Procedia CIRP. https://doi.org/10.1016/j.procs.2021.05.098