Supported by tetfund TERTIARY EDUCATION TRUST FUND

TRENDS IN MACHINE LEARNING ON AUTOMATIC DETECTION OF HATE SPEECH ON SOCIAL MEDIA PLATFORMS: A SYSTEMATIC REVIEW

Hyellamada Simon[1], Benson Yusuf Baha[2], Etemi Joshua Garba[3]

[1]Department of Computer Science, Federal Polytechnic Mubi, Adamawa State, Nigeria.
[2]Department of Information Technology, Modibbo Adama University Yola, Adamawa State, Nigeria.
[3]Department of Computer Science, Modibbo Adama University Yola, Adamawa State, Nigeria.
Corresponding author's e-mail address: hyellasimon@gmail.com

Received: December 13, 2021 Accepted: February 20, 2022

ABSTRACT Social media provides a user-friendly platform for interested persons or groups to express opinions and discuss freely their topics of interest which enhances the propagation of online hate speech, which is considered a serious issue in the web community because cyber hate speech has the potential to cause harm to individuals and society at large. The main objective of this paper is to study current literatures on the detection of online hate speech to determine the trends in online hate speech detection tasks. Various databases (Elsevier, IEEE Xplore, ACM digital library, Springer, and Google Scholar) were searched to obtain the materials used for this review. The method adopted for this review is the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). Various databases (ScienceDirect, IEEE Xplore, ACM digital library, Springer, and Google Scholar) were searched. A total of 31,714 publications from 2015 to 2020 were studied, a total of 31,673 papers were excluded based on exclusion criteria, and 41 papers were included based on inclusion criteria. The results show that the Support Vector Machine learning algorithm was the most commonly used algorithm for online hateful text classification, though, deep learning algorithms and hybrid deep learning approaches are gaining grounds recently. This paper concludes that machine learning and deep learning approaches have proven effective in the classification of hateful text on social media. However, there is a need for the development of hybrid cross-platform models for hate speech detection and blocking.

KEYWORDS Hate speech, Machine learning, Deep learning, Social media, Hate speech detection, and Text classification.

Introduction

Recently, people have become more engaged with the widespread social network activities. The micro-blogging applications have opened up the chance for people around the globe to freely express and share thoughts in a real-time manner, which has encouraged the propagation of hateful content. Such expression gave researchers the ability to investigate the online social emotions in different events. According to Agarwal (2015), various social media platforms on the Internet such as Twitter, Facebook, YouTube, Blogs, and discussion forums are being misused by extremist groups to spread different beliefs and ideologies, also promoting radicalization, recruiting members, and building online virtual communities. Online Hate speech has been an active area of research that attracted the interest of many researchers (such as, Burnap & Williams, 2015; Gamback & Sikdar, 2017; Ruwandika & Weerasinghe, 2018; Park & Fung, 2018; Sharma, Kshitiz, & Shailendra, 2018; Aulia & Budi, 2019) to use diverse techniques of Machine Learning (ML) to propose several Artificial Intelligence (AI) models to automatically detect online hate speech, bullying, aggression, abusive,

and misogynous comments especially in social media. The term "hate speech" was defined by Schmidt and Wiegand (2017) as a broad umbrella term for various kinds of insulting user-created content. Other serious issues affecting most Internet users are cyberbullying and cyber aggression (Sahay et al., 2018). Popular social media platforms such as Twitter and Facebook are the most vulnerable to such attacks (Sahay et al., 2018).

However, Facebook and Twitter vowed to remove hate speech content within 24 hours when reported (Kottasova, 2016 in Zenuni et al., 2017). The European Union (EU) has launched a "code of conduct" that valid hate speech contents will be removed while the right to freedom of expression will be preserved (Commision, 2016 in Zenuni, et al., 2017). The CEO of Facebook, Mark Zuckerberg, revealed before US Congress that the problem of hate speech can be solved in the next "5 to10 years" (De Smedt et al., 2018). The Nigeria Senate also proposed a bill on hate speech such that any person who is convicted guilty of any form of hate speech resulted in the death of another person shall die by hanging (Uzochukwu & Okafor, 2019).

The problem of hate speech propagation on the internet has attracted the interest of researchers (for example, Raiyani et al., 2018; Sahay et al., 2018; Mujadia et al., 2019; Rodriguez, Argueta, & Chen, 2019) to use several Artificial Intelligence (AI) techniques such as Machine Learning (ML), Natural Language Processing (NLP), and Deep Learning (DL) to propose models that can automatically detect the presence of hate speech, cyberbullying and use of abusive languages on social media in English and English code-switch or code-mixed languages.

This paper aims to conduct a systematic review of various techniques proposed in the literature for detecting the use of hateful comments on social media from various databases, to identify challenges and suggest the way forward in this area of research. This paper is organized into 5 sections: starting with section 1 the introduction, section 2 the review of existing literature, section 3 covers the methods adopted for the study, section 4 presents our findings, section 5 discussion of findings, then section 6 is the conclusion and future research.

What is considered hate speech from several sources

Hate speech has been defined by several sources (Guermazi et al., 2007; McNamee et al., 2010; Chen, 2011; Wendling, 2015; Agarwal, 2015; Thompson, 2016; ILGA, 2016; Tarasova, 2016; Nobata et al., 2016; Del'Vigna et al., 2017; Jigsaw, 2017) though most of the definitions are similar and targeted the same points. Hate speech is defined as a language that demeans or attacks a person or group based on race, ethnic origin, religion, gender, age, disability, or sexual orientation/gender identity (Nobata et al., 2016). Similarly, Code of Conduct, between EU and companies defines hate speech as all conduct that publicly incites violence or hatred directed against a person or group of persons or a member of such a group defined by reference to religion, race, color, ancestry or national or ethnicity (Wendling, 2015). International Minorities Associations (ILGA) also stated that "Hate speech is public expressions which spread, incite, promote or justify hatred, discrimination or hostility towards a specific group. They contribute to a general climate of intolerance which in turn makes attacks more probable against those given groups" (ILGA, 2016).

Facebook: "define hate speech as a direct attack on people based on what we call protected characteristics, such as race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability" (Facebook Community Standards, n.d.).

Twitter: "You may not promote violence against or directly attack or threaten other people based on race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease" (Twitter, n.d.).

YouTube: "Hate speech refers to content that promotes violence or hatred against individuals or groups based on certain attributes: age, caste, disability, ethnicity, gender identity, and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin, and veteran status" (YouTube, n.d.)

Based on the definitions of hate by different sources it is clear to say that hate speech is any content which incites violence or hatred or attack that is targeted at a person or group or a member of a group based on age, race, color, religion, nationality, sexual orientation, disability or illness, gender, ethnicity, etc.

Various approaches to automatic hate speech detection

Applications of text mining on social media cannot be over emphasized. This task mainly depends on text mining approaches such as NLP, ML, and DL algorithms. Thus, these algorithms are used to

Table 2.1: Summary of hateful related concepts and the distinction from hate speech

Concept	Definition of the concept	Concept distinction from hate speech
Hate	Expression of hostility without any reason for it (Tarasova, 2016).	Hate speech is hate which focused on stereotypes, and not general.
Cyberbullying	An aggressive and intentional act carried out by a person or group, using electronic means, repeatedly and over time, against a person (victim) who cannot easily defend themselves (Chen, 2011).	Hate speech is more general and does not necessarily focus on a specific person.
Abusive language	The term abusive language was used to refer to the hurtful language, including hate speech, derogatory language, and profanity (Nobata et al., 2016).	Hate speech is a kind of abusive language.
Profanity	Offensive or obscene word or expression (Del'Vigna et al., 2017).	Hate speech can use profanity, but not inevitably.
Toxic language or Comment	Toxic language is a rude, disrespectful or irrational comment that is likely to make a person leave a discussion (Jigsaw, 2017).	Not all toxic comments contain hate speech. However, some hate speech can make people discuss more.
Extremism	The ideology associated with extremists or hate groups, promoting violence, often aiming to segment populations and reclaiming status, where outgroups are presented as perpetrators or inferior populations (McNamee et al., 2010).	Extremist discourses frequently used hate speech. However, these discourses focused on other topics as well, such as new members recruitment, government and social media demonization of the in-group and persuasion (McNamee et al., 2010).
Radicalization	Online radicalization is a concept similar to the extremism concept which has been studied on multiple topics and domains, such as terrorism, anti-black communities, or nationalism (Agarwal, 2015).	Radical discursions, such as extremism used hate speech. However in radical discursions, topics like war, religion and negative emotions are common (Agarwal, 2015). While hate

		speech can be more indirect and grounded in stereotypes.
Discrimination	A process by which a difference is identified and then use such as the basis of unfair treatment (Thompson, 2016).	Hate speech is a form of discrimination, via verbal means.
Flaming	These are hostile, profane and intimidating comments that can interrupt participation in a community (Guermazi et al., 2007).	Hate speech can occur in any context while flaming aimed at a participant in the specific context of a discussion.

ML approaches are effective in the area of text classification for detection tasks. The approaches are categorized into *supervised, semi-supervised and unsupervised* approaches:

Supervised learning: This approach depends on the domain of application since it relies on a large volume of texts that are labelled manually. The labeling task is time and effort consuming though more efficient for domain-dependent events. Most of the approaches used for the task of hate speech detection are supervised methods. Molina-Gonzalez et al. (2019) proposed an ensemble learning to detect the aggressiveness on Mexican Spanish tweets using several supervised classifiers. Their result shows that a combination of the methods increased the Macro F1-score in all the classifiers.

Semi-supervised learning: The semi-supervised learning algorithms are trained using both labelled and unlabelled data. Using both labelled with unlabelled data can effectively enhance the performance of the classification algorithm as in (Hua et al., 2013). They argued that unsupervised learning has limited ability to handle small scale events.

Unsupervised learning: This is a domain-independent approach that is capable of handling diverse content while maintaining scalability. This approach does not rely on humans to label a large volume training dataset, but it dynamically extracts domain-related key terms. Gitari et al. (2015) employed a bootstrapping approach to building their lexicon from hate verbs and then expanded it iteratively. They achieved the best result by incorporating more features.

Natural language processing

The NLP is an analysis of linguistic data, mostly in the form of textual data such as documents or publications, using computational techniques. NLP generally is used to build a representation of the text and adds structure to the unstructured natural language, by using insights from linguistics. Sahay et al. (2018) proposed a robust methodology for extracting text, user, and network-based attributes. They also studied the properties of bullies and aggressors, and the features distinguished them from other users.

Deep learning

The DL models have shown a promising future for text mining tasks. It entirely depends on the Artificial Neural Networks framework but with extra depth. Deep learning tries to mimic the operations in layers of neurons and attempt to learn in a real sense by identifying patterns in the given text. Although, DL approaches are not mostly better than the traditional supervised approaches. Moreover, the performance of DL is subject to the choice of the right algorithm and many hidden layers as well as the feature representation technique. Pitsilis et al. (2018) have broken the barrier of language dependency in the word embedding approach by using the Recurrent Neural Network model with word frequency vectorization to implement the features instead of the word embedding. Their results outperformed the current

state of art deep learning approaches for the detection of hate speech.

Features Representation for detection of hate speech

There are general features considered in performing automatic detection of hate speech. The features of the corpus must be specified to enable the classification algorithms to perform the task. Thus:

Word embedding and Word2Vec: The development of word embedding eased the data sparsity problem by bringing up an extra semantic feature and generating distributed a representation that introduces dependence between words. Word2Vec is one of the techniques to construct word embedding. Word2vec has attracted a lot of interest by researchers in the text mining field and because it is compatible with both supervised and unsupervised machine learning models (Lilleberg, Zhu, & Zhang, 2015).

Dictionaries and Lexicons: This feature usually employed unsupervised machine learning. Gitari et al. (2015) used a lexicon as a primary feature to aggregate opinions and giving rates to the subjective words. Wiegand et al. (2018) also proposed a model for the detection of profane words by taking advantage of corpora and lexical resources. They used general-purpose lexical resources and several features to build their lexicon.

Bag-of-words (BOW) and N-grams: This is a word co-occurrence feature. A process of vectorization is performed on tokenized words in the corpus by assigning weight for each word according to its frequency in the tweet and its frequency in between different tweets. The vectorization process is performed using some statistical models, such as Term Frequency-Inverse Document Frequency (TF-IDF) weight, then a list of words is presented as vectors of weights (that is, BOW) (George & Joseph, 2014). n-gram is a representation of sequences of n adjacent words. Waseem and Hovy (2016) analysed the impact of using many features in combination with character N-gram for hate speech detection. They disclosed that using character n-gram representation is a great option for hate speech detection. BOW has a limitation because it needs to be accompanied by other features to improve performance, but it is computationally expensive (Tsai, 2012).

Latent Dirichlet Allocation (LDA): This is a probabilistic topic modelling method. It is generally used for an estimation of the latent topics in a data set and these latent topics will be used as features instead of words. The LDA is suitable for unsupervised and semi-supervised machine learning systems. Xiang et al. (2012) postulate that BOW did not work well for the detection of abusive text on Twitter. They used highly expressive topical features and other lexicon features by using the LDA model as an alternative to supervised methods.

Materials and methods

This study adopted the PRISMA method of literature review by Moher et al. (2015) to investigate the trends in machine learning for automatic detection of hate speech on social media platforms. The PRISMA technique takes care of literature search, selection, and summarization.

Materials deployed

In reviewing previous studies, a search was performed in reputable scientific electronic databases using keywords such as hate speech, hate speech detection, machine learning, social media, and text classification to download research papers published from 2015 to 2020. Similarly, the databases searched were ScienceDirect, IEEE Xplore, ACM digital library, Springer, and Google Scholar.

Criteria for inclusion and exclusion

The articles screening was based on inclusion criteria with consideration on some factors such as studies published from 2015 to 2020, studies presented in the English language, articles that have concentrated on the detection of hateful related text (English and English code-mixed text) on social media. Similarly, the exclusion criteria were also used to eliminate publications that were published earlier than the year 2015, non-English papers, and articles that have not concentrated on the detection of hateful related text (English and English code-mixed text).

Articles screening Using PRISMA method

Based on the keywords supplied for articles search from the electronic databases, a total of 31,714 publications were investigated from various electronic databases, such as ScienceDirect, IEEE Xplore, ACM digital library, Springer, and Google Scholar. The papers were screened based on inclusion and exclusion criteria. Figure 3.1 shows the PRISMA flow diagram for the selection of the articles used in this research.

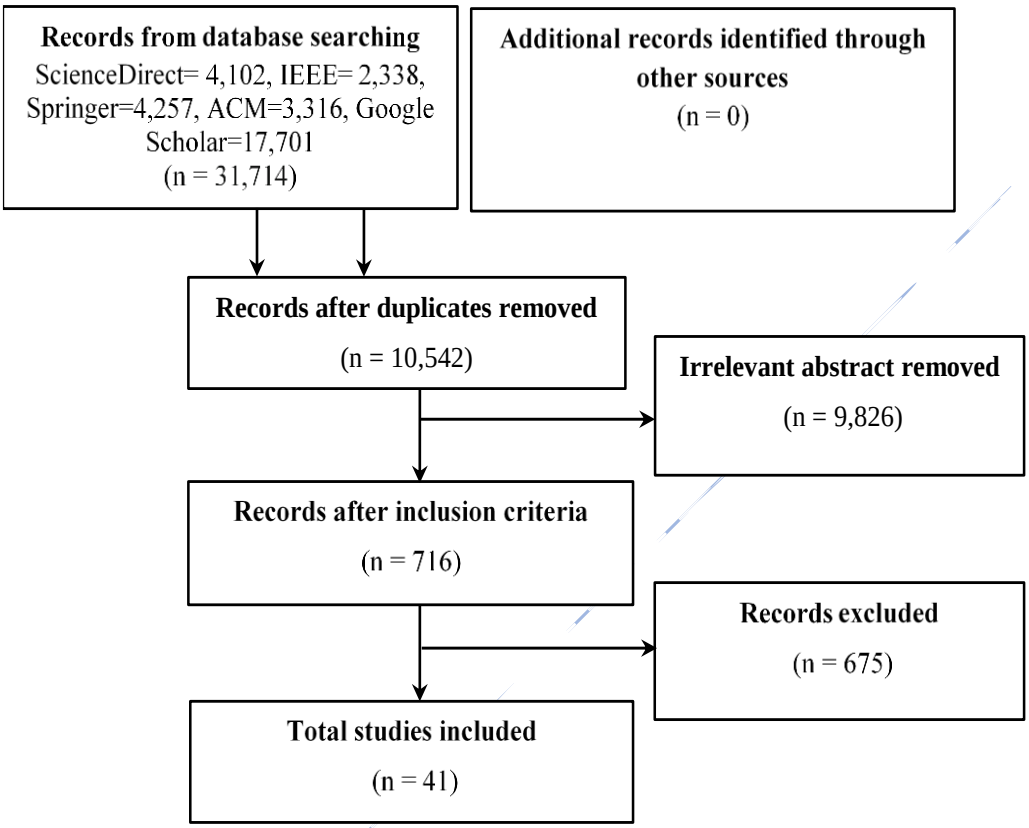

Figure 3.1: PRISMA method for the articles screening (Moher et al., 2015).

Results

Selected papers for the study

From figure 3.1, a total number of 21,172 papers were discarded due to repetition, 9,826 papers were also discarded due to irrelevant abstract. Similarly, a total of 675 papers were eliminated based on exclusion criteria, leaving a total of 41 full-text articles which were included for this study. In the selected studies, several models were proposed for the detection of hate speech and other related online anti-social behaviors such as cyberbullying, aggression, misogyny, inappropriate and abusive languages in social media based on ML, NLP, and DL techniques as summarized in Table 4.1.

Table 4.1: Summary of reviewed models for automatic detection of other hateful related contents on social media

No.	Author and Year	Language	Social media Platform	Classification focus	Features Representation	Algorithms	Results Efficiency
1	Burnap & Williams (2015)	English	Twitter	Hateful and aggressive comments	BoW, n-gram (n=1-5)	probabilistic, ruled-based and spatial-based classifiers	98%
2	Waseem & Hovy (2016)	English	Twitter	Hate speech	Extra-linguistic features and character n-grams (n=1-4)	Character and word n-grams	64.58 %.
3	Davidson et al. (2017)	English	Twitter	Hate speech and offensive language	Part-of-Speech (POS) tags, bigram, unigram, trigram, and TF-IDF	Naive Bayes (NB), Logistic Regression (LR), Random Forests (RF), Decision Trees (DT), and linear SVMs	90%
4	Gamback & Sikdar (2017)	English	Twitter	Hate speech	Character 4-grams, word2vec		78.3%.
5	Malmasi & Zampieri (2017)	English	social media	Hate speech, profanity, and other anti-social behavior	Character n-grams, word n-grams, and word skip-grams	Linear SVM	78%
6	De Smedt et al. (2018)	English	Twitter	Hate speech	Character trigrams, keyword extraction	Linear SVM and DT	80%
7	Martins et al. (2018)	English	social media	Hate speech	bi-grams and tri-grams	SVM, Naive Bayes and Random Fores	80.56%

8	Ahluwalia et al. (2018a)	English	Twitter	Hate speech and aggression towards women	Word unigrams and bigrams	ML	55%
9	Ruwandika & Weerasinghe (2018)	English	social media	Hate speech	Bag-of-Words features (BoW), TF-IDF, and Bag-of-Features (BoF)	SVM, NB Classifier, LR Classifier, DT Classifier and K-Means Clustering	71.9%
10	Watanabe et al. (2018)	English	Twitter	Hate speech	Unigrams and patterns	J48graft, SVM, and Random Forest	78.4%
11	Salminen et al. (2018)	English	online news media	Hate speech	TF-IDF	LR, RF, DT, Adaboost, and Linear SVM	79%
12	Pitsilis et al. (2018)	English	social media	Hate speech	User-related information	RNN	92.95%
13	Gaydhani et al. (2018)	English	Twitter	Hateful and offensive language	n-grams, TF-IDF	LR, NB, and Linear SVM	95.6%
14	Qian et al. (2018)	English	Twitter	Hate speech	bi-LSTM + attention, n-grams	Intra.+ Reinforced Inter. Rep.	77.4%
15	Sahay et al. (2018)	English	Twitter	Classification of cyberbullying and aggression.	Count vectors, TF-IDF, n-gram of up to five levels.	LR, SVM, RF, and Gradient Boost (GB).	77- 90%
16	Abdullah et al. (2018)	English	Twitter	Classification of positive, negative and neutral sentiment.	TF-IDF.	LG, NB, RF, Decision Tree (DT), SVM.	79%
17	Raiyani et al. (2018)	Hindi and English	Facebook and Twitter	Automatic detection of aggression	TF-IDF and n-gram (1-3)	LR, SVC, Multinomial NB, Bernoulli NB, Ridge Classifier, and AdaBoost Classifier.	57.64% - 59.67% for detection in Hindi, and 55.34% - 63.71% for detection in English languages.
18	Sharma et al. (2018)	English	Twitter	Classification of	Lexical Syntactic Features, TF-	LR, SVM, RF, and GB.	75-90%

				cyberbullying and aggression	IDF, words count, count of second-person pronoun in a sentence, Character n-gram (1-5) word/Document Vectors.		
19	Ahluwalia et al. (2018b)	English	Twitter	Automatic detection of misogynous and non-misogynous content	Sentiment scores, a bag of words, and lexical features	LR, SVM, RF, GB, and Stochastic Gradient Descent (SGD).	78.51%
20	Van Hee et al. (2018)	English and Dutch	Twitter	Automatic detection of cyberbullying	Word n-grams, Character n-grams, Term lists, Subjectivity lexicons, and Topic models.	linear SVM	64% for detection in English, and 61% for detection in Dutch languages.
21	Park & Fung (2018)	English	Twitter	Automatic detection of abusive language (sexist/racist)	Character and word features.	LR, SVM, FastText, CharCNN, WordCNN, and HybridCNN	82.70%
22	Yenala et al. (2018)	English	Search engines and messenger	Automatic detection of inappropriate language	Query words	Coolutional bi-LSTM, Convolution Neural Networks (CNN) and bi-LSTM	78.90%
23	Aroyehun & Gelbukh (2018)	English	Facebook and other Social media	Automatic detection of cyber aggression	Word n-grams and character n-grams	CNN, LSTM, bi-LSTM, CNN-LSTM, LSTM-CNN, CNN-bi-LSTM, and bi-LSTM-CNN.	64.25% for detection on Facebook, and 59.20% for other social media
24	Agrawal & Awekar (2018)	English	Twitter, Formspring	Automatic detection of cyberbullying	Character n-gram, word unigram, GloVe	SVM, LR, NB, RF, and LSTM, CNN,	94% for Twitter, 78.5% for

			, and Wikipedia.		embeddings, and SSWE embeddings.	bi-LSTM, and bi-LSTM with Attention	Formspring, and 95% for Wikipedia datasets.
25	Pelle et al. (2018)	English and Portuguese	Twitter and other social media	Automatic detection of offensive comments	Word and character n-grams, and Word2Vec.	Logistic Regression	90% -97%
26	Mujadia et al. (2019)	English, German and Hindi	Twitter	Automatic detection of abusive language	TF-IDF vectors, word and character level n-grams (n=1-5).	SVM, RF, Adaboost, and Voting classifiers, LSTM,	69.70% for English, 47.7% for German, and 80.32% for Hindi languages.
27	Sigurbergsson & Derczynski (2019)	English and Danish	Facebook and Reddit	Automatic detection of offensive language	Linguistic features, pre-trained word embeddings, and sentiment scores.	LR, Learned-BiLSTM, Fast-bi-LSTM, and AUX-Fast-bi-LSTM.	74% for English language, and 70% for Danish language.
28	Altin et al. (2019)	English	Twitter	Automatic detection of offensive language	Pre-trained Word embeddings	bi-LSTM	82.9%
29	Umar et al. (2019)	English	Twitter	Classification of abusive language and the user's involvement.	word embedding	User profiling algorithm and deep LSTM	89.14% for classifying abusive language, and 83.33% for detection of user's involvement.
30	Naveen & Kumar (2019)	English	Twitter	Automatic detection of offensive, hateful, and clean comments	n-grams and TF-IDF	LR, NB, and Linear SVM.	95.6%
31	Garain & Basu (2019)	English	Twitter	Hate speech and aggressive behavior	bi-LSTM	Neural Network	57.3% for hate speech detection and 76.3% for aggression detection

32	MacAvaney et al. (2019)	English	Facebook and Stormfront	Hate speech	TF-IDF and unigram	Multi-view SVM	53.68% for Facebook, and 80.33% for Stormfront
33	Faris et al. (2020)	Arabic	Twitter	Automatic detection of cyber hate	Word embeddings	Hybrid CNN and LSTM networks	71.688%
34	Sadiq et al. (2020)	English	Twitter	Automatic detection of aggression	n-grams (n=1-2)	CNN-LSTM and CNN-bi-LSTM in deep neural network	92%
35	Jain, Kumar, & Garg (2020)	English and Hindi	Twitter	Sarcasm detection	GloVe	bi-LSTM and CNN	92.71% and 89.05%
36	Kapil & Ekbal (2020)	English	Facebook and other social media	Hate speech	Word embedding and character embedding	CNN + Gated Recurrent Unit (GRU)	80.68% for detection in Facebook, and 86.52% for detection in other social media.
37	Salminen et al. (2020)	English	Multi-platform (YouTube, Twitter, Wikipedia, and Reddit).	Hate speech	Bag-of-Words, Word2Vec, TF-IDF, BERT, and their combinations	SVM, LR, NB, XGBoost, and Neural Networks	92%
38	Chopra et al. (2020)	Hindi-English code-switched	Twitter	Profanity detection	deep graph embeddings, and author profiling	node2vec + Dense, DeepWalk + Dense, CNN + bi-LSTM + Attn + n2v, and CNN + bi-LSTM + Attn + DeepWalk	73%
39	Dorris et al. (2020)		Twitter	Hate speech and offensive language detection	GloVe embeddings, and LSTM	Neural Networks	90.82% for hate speech, and 89.10% for offensive language detection
40	Modha et al. (2020)	English and code-mixed Hindi	Twitter and Facebook	Online aggression	Attention-based model, and BERT pre-trained language	Linear SVM, LR, and CNN	64% for detection on Facebook, and 58% for

							detection on Twitter
41	Sreelakshmi et al. (2020)	Hindi-English code-mixed	Facebook	Hate speech	pre-trained word embedding, and FastText	Linear SVM-Radial Basis Function	85.51%

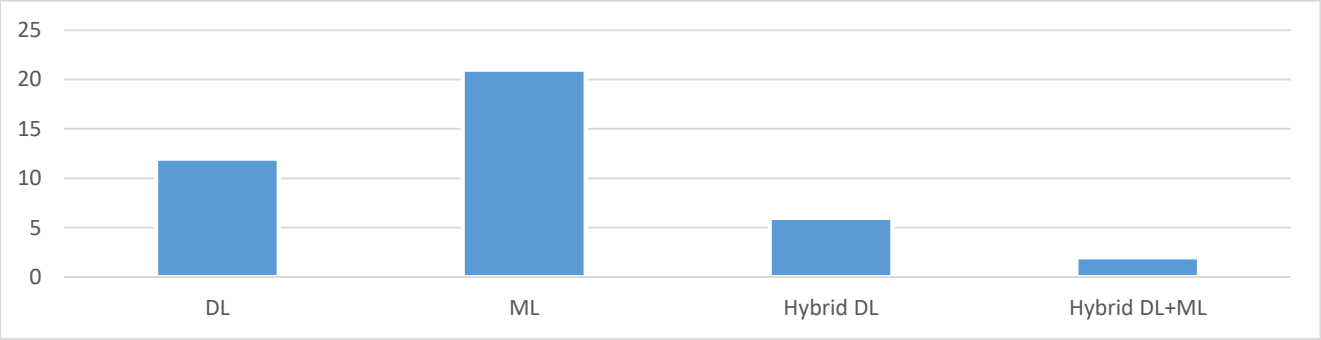

Figure 4.1: Graphical representation of reviewed publications from 2015-2020.

Discussions

Findings from Table 4.1 revealed that most of the models proposed for the detection of hate speech and other anti-social behaviors on social media have been treated as a text classification task. Most of those models were based on supervised ML algorithms (Del'Vigna et al., 2016; Davidson et al., 2017; Pelle et al., 2018; Sharma et al., 2018; Naveen & Kumar, 2019). Though, few pieces of literature incorporate supervised and unsupervised learning algorithms (Gitari et al., 2015; Ahluwalia et al., 2018; Ruwandika & Weerasinghe, 2018).

Findings from Figure 4.1 revealed that the SVM algorithm being an ML algorithm was the most commonly applied classification technique in most of the reviewed literature (Davidson et al., 2017; Malmasi & Zampieri, 2017; De Smedt et al., 2018; Salminen et al., 2018; Ahluwalia et al., 2018; Van Hee et al., 2018; Umar et al., 2019; Mujadia et al., 2019; Naveen & Kumar, 2019; Modha et al., 2020; Sreelakshmi et al., 2020; Salminen et al., 2020) for detection of hateful related contents on social media due to its high level of performance and accuracy for text classification. Table 4.1 also presented existing works proposed for hate speech detection on social media (multi-platforms) based on ML algorithms (Raiyani et al., 2018; Yenala et al., 2018; Aroyehun & Gelbukh, 2018; Agrawal & Awekar, 2018; Pelle et al., 2018; Sigurbergsson & Derczynski, 2019; MacAvaney et al., 2019; Kapil & Ekbal, 2020; Salminen et al., 2020; Modha et al., 2020), the models presented good results with high performance in terms of hate speech detection and other anti-social behaviors on the internet.

Findings in Figure 4.1 also revealed that DL algorithms are recently gaining ground in the area of text classification problems by many researchers (Aroyehun & Gelbukh, 2018; Chopra et al., 2020; Kapil & Ekbal, 2020; Jain et al., 2020). However, some authors (Park & Fung, 2018; Sadiq et al., 2020; Faris et al., 2020) proposed DL hybrid models contribute on the task of online hate speech detection. The DL models and hybrid DL models achieved good results as seen in Table 4.1. Similarly, Table 4.1, shows that supervised learning algorithms with both word and character n-grams features range up to n=5, and TF-IDF has been proven to be effective in hate speech detection and classification of other online related anti-social behaviors (Pelle et al., 2018; Aroyehun & Gelbukh, 2018; MacAvaney et al., 2019; Naveen & Kumar, 2019; Mujadia et al., 2019; Sadiq et al., 2020). Though none of the papers tried higher values of n because the higher the value of n the more data is needed for model training which is also time-consuming.

Conclusion and Future Work

This paper examined the concept of hate speech in different platforms and contexts, and also study the current trends on the use of text mining in social networks for automatic detection of hateful messages which depends on ML, NLP, and DL algorithms. Based on our findings, several works of literature defined hate speech in different perspectives which defers from one context to

another. Hence, we proposed a more general and unified definition of the concept which includes any form of discrimination on the internet. The general definition will help researchers in labeling data to build models for automatic detection of hate speech on the internet and diverse social networking platforms. Based on our findings, SVM algorithms were the most commonly used algorithm for online hateful text detection which yielded good results. Recently deep learning algorithms and hybrid deep learning approaches have gained more grounds with high results performance for detection of hate speech and other anti-social behaviors.

Our findings also revealed that the tasks of hate speech detection were mostly treated as supervised learning problems, using basic features such as word embeddings which produced reasonable classification performance, specifically the Word2vec which is compatible with both supervised and unsupervised machine learning algorithms. However, the task of hate speech detection using supervised learning suffered from huge work of manual labeling of a large volume of data, and standard labeled datasets are not always available. Therefore, future work should attempt to use unsupervised or semi-supervised learning techniques to build a hybrid DL model that can automatically detect, report, and block hate speech on social networking platforms. The use of unsupervised/semi-supervised learning will tackle the problems of social media posts that contain emojis, and other kinds of abbreviations that do not conform to the English language standard. Accordingly, the unsupervised/semi-supervised learning techniques will be effective for hate speech tasks since they can handle a large volume of unlabelled data for model development and testing. This paper also suggests that since there is no fixed grammar, vocabulary, semantics, and spellings for text presentation on social media which posed a problem to hate speech detection models, creating a list of slurs in combination with other types of linguistic features may help in hate speech detection problems. Similarly, information derived from the text in addition to meta-information may be useful in detecting the presence of hate speech.

REFERENCES

Agarwal, S., 2015. Applying Social Media Intelligence for Predicting and Identifying On-line Radicalization and Civil Unrest Oriented Threats. Retrieved 6th November, 2019,from https://arXiv.org/abs/1511.06858

Agrawal, S., Awekar, A., 2018. Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms. Springer, BBC News Article. 141–153. Retrieved 20th November, 2019, from https://goo.gl/t6hQ7c.

Ahluwalia, R., Shcherbinina, E., Callow, E., 2018a. Detecting Hate Speech Against Women in English Tweets. URL: CEUR-WS.org/Vol-2263/paper032.pdf

Ahluwalia, R., Shcherbinina, E., Callow, E., Nascimento, A., De Cock, M., 2018. Detecting Misogynous Tweets. Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018). 242-248.

Aroyehun, S. T., Gelbukh, A., 2018. Aggression Detection in Social Media: Using Deep Neural Networks, Data Augmentation, and Pseudo Labeling. Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying. Santa Fe, USA, 90–97.

Abdullah K. K. A., Folorunso S. O., Solanke O. O., Sodimu S. M., 2018. A Predictive Model for Tweet Sentiment Analysis and Classification. Anale. Seria Informatica.16(2), 35-44.

Altin, L. S., Bravo, A., Saggion, H., 2019. LaSTUS/TALN at SemEval-2019 Task 6: Identification and Categorization of Offensive Language in Social Media with Attention-based Bi-LSTM model. Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019). Minneapolis, Minnesota, USA, 672–677

Aulia, N., Budi, I., 2019. Hate Speech Detection on Indonesian Long Text Documents Using Machine Learning Approach. Association for Computing Machinery (ACM). 164-169.

Burnap, P., Williams, M.L., 2015. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. Policy & Internet. 7(2), 223-242.

Cao, R., Lee, R.. K., Hoang., T., 2020. DeepHate: Hate Speech Detection via Multi-Faceted Text Representations. WebSci '20: 12th ACM Conference on Web Science, 11-20.

Chen, Y., 2011. Detecting Offensive Language in Social Medias for Protection of Adolescent Online Safety. Ph.D. Dissertation. The Pennsylvania State University.

Chopra, S., Sawhney, R., Mathur, P., Shah, R. R., 2020. Hindi-English Hate Speech Detection: Author Profiling, Debasing, and Practical Perspectives. Association for the Advancement of Artificial Intelligence.

Del'Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., Tesconi. M., 2017. Hate me, hate me not: Hate speech detection on Facebook. Proceedings of the 1st Italian Conference on Cybersecurity. 86–95.

Davidson, T., Warmsley, D., Macy, M., Weber, I., 2017. Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM). 512-515.

De Smedt, T., De Pauw, G., Van Ostaeyen, P., 2018. Automatic Detection of Online Jihadist Hate Speech. Computational Linguistics & Psycholinguistics Technical Report Series. CTRS-007. ISSN 2033-3544.

Dorris, W., Hu, R., Vishwamitra, N., Luo. F., Costello., M., 2020. Towards Automatic Detection and Explanation of Hate Speech and Offensive Language. Proceedings of the Sixth International Workshop on Security and Privacy Analytics, 23-29.

Facebook. (n.d.). Hate Speech. URL: https://web.facebook.com/communitystandards/hate_speech?_rdc=1&_rdr

Faris, H., Aljarah, I., Habib, M., Castillo, P.A., 2020. Hate Speech Detection using Word Embedding and Deep Learning in the Arabic Language Context. Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2020), 453-460.

Guermazi, R., Hammami, M., Hamadou, A. B., 2007. Using a semi-automatic keyword dictionary for improving violent Web site filtering. Proceedings of the 3rd International IEEE Conference on Signal-Image Technologies and Internet-Based System (SITIS). IEEE, 337–344.

George, S. K., Joseph, S., 2014. Text Classification by Augmenting Bag of Words (BOW) Representation with Co-occurrence Feature. IOSR J. Comput. Eng., 16(1), 34–38.

Gamback, B., Sikdar, U. K., 2017. Using Convolutional Neural Networks to Classify Hate-Speech. Proceedings of the First Workshop on Abusive Language Online. 85–90.

Gitari, N.D., Zuping, Z., Damien, H., Long, J., 2015. A lexicon-based approach for hate speech detection. Int. J. Multimed. Ubiquitous Eng.10(4), 215–230.

Gao, L., Huang, R., 2018. Detecting Online Hate Speech Using Context Aware Models. URL: https://arXiv.org/abs/1710.07395

Gaydhani, A., Domay, V., Kendrez, S., Bhagwat, L., 2018. Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach. DOI:https://arXiv.org/abs/1809.08651

Garain, A., Basu, A., 2019. The Titans at SemEval-2019 Task 5: Detection of hate speech against immigrants and women in Twitter. Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019). Minneapolis, Minnesota, USA, 494–497.

Hua, T., Chen, F., Zhao, L., Lu, C. T., Ramakrishnan, N., 2013. STED: Semi-Supervised Targeted-Interest Event Detection in Twitter. Proceedings of the 19th ACM SIGKDD- International Conference on Knowledge Discovery and Data Mining. 1466–1469.

ILGA., 2016. Hate crime and hate speech. Retrieved 26th March, 2020 from https://www.ilga-europe.org/what-we-do/our-advocacy-work/hate-crime-hate-speech

Jain, D., Kumar, A., Garg, G. 2020. Sarcasm detection in mash-up language using soft-attention based bi-directional LSTM and feature-rich CNN. Applied Soft Computing, 106198. DOI:10.1016/j.asoc.2020.106198

JIGSAW., 2017. Perspective API. Retrieved 22nd February, 2020, from https://www.perspectiveapi.com

Kapil, P., & Ekbal, A. 2020. A deep neural network based multi-task learning approach to hate speech detection. Knowledge-Based Systems, 106458. DOI:10.1016/j.knosys.2020.106458

Lilleberg, J., Zhu, Y., Zhang, Y., 2015. Support vector machines and Word2vec for text classification with semantic features. Proc. of IEEE 14th Int. Conf. Cogn. Informatics Cogn. Comput. (ICCI*CC 2015), 136–140.

McNamee, G. L., Peterson, B. L., Pena, J., 2010. A call to educate, participate, invoke. and indict: Understanding the communication of online hate groups. Commun. Monogr. 77(2), 257–280.

Malmasi, S., Zampieri, M., 2017. Detecting Hate Speech in Social Media. DOI:https://arXiv.org/abs/1712.06427

Martins, R., Gomes, M., Almeida, J. J., Novais, P., Henriques, P., 2018. Hate speech classification in social media using emotional analysis. 7th Brazilian Conference on Intelligent Systems. IEEE. 61-66.

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

Molina-Gonzalez, M., Plaza-del-Arco, F., Martin-Valdivia, T. M., Urena-Lopez, L.A., 2019. Ensemble Learning to Detect Aggressiveness in Mexican Spanish Tweets. Proceedings of the Iberian Languages Evaluation Forum (IberLEF). 495-501.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., ... & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic reviews, 4(1), 1.

Modha, S., Majumder, P., Mandl, T., Mandalia, C., 2020. Detecting and Visualizing Hate Speech in Social Media: A Cyber Watchdog for Surveillance. Expert Systems with Applications, 113725. DOI: 10.1016/j.eswa.2020.113725

Mujadia, V., Mishra, P., Sharma, D. M., 2019. IIIT-Hyderabad at HASOC 2019: Hate Speech Detection. Retrieved 10th November, 2019 from http://ceur-ws.org/Vol-2517/T3-12.pdf

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y., 2016. Abusive language detection in online user content. Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 145–153.

Naveen, K., Kumar, C. P., 2019. An effective scheme for detecting hateful and offensive expressions on twitter. Intl. Journal for Innovative Engineering and Mgt. Research. 8(8). 204-209. ISSN 2456 – 5083

Qian, J., ElSherief, M., Belding, E. M., Wang, W. Y., 2018. Leveraging Intra-User and Inter-User Representation Learning for Automated Hate Speech Detection. DOI: https://arXiv.org/abs/1804.03124

Park, J. H., Fung, P. F., 2018. One-step and Two-step Classification for Abusive Language Detection on Twitter.

Pitsilis, G. K., Ramampiaro, H., Langseth, H., 2018. Effective Hate-speech Detection in Twitter data using Recurrent Neural Networks. Appl. Intelligence. 48(12), 4730-4742.

Pelle, R., Alcantara, C., Moreira, V. P., 2018. A Classifier Ensemble for Offensive Text Detection. WebMedia. 18, 237-243.

Raiyani, K., Goncalves, T., Quaresma, P., Nogueira, V. B., 2018. Fully Connected Neural Network with Advance Preprocessor to Identify Aggression over Facebook and Twitter. Proceedings of the First Workshop on Trolling, Aggression and Cyberbullyin. Santa Fe, USA, 28–41.

Ruwandika, N. D. T., Weerasinghe, A. R., 2018. Identification of Hate Speech in Social Media. International Conference on Advances in ICT for Emerging Regions (ICTer). 273 – 278.

Rodriguez, A., Argueta, C., Chen, Y., 2019. Automatic Detection of Hate Speech on Facebook Using Sentiment and Emotion Analysis. IEEE. 169-174.

Sadiq, S., Mehmood, A., Ullah, S., Ahmad, M., Choi, G. S., On, B.W., 2020. Aggression detection through deep neural model on Twitter. Elsevier, Future Generation Computer Systems, 114,120-129. DOI:10.1016/j.future.2020.07.050

Schmidt, A., Wiegand, M., 2017. A Survey on Hate Speech Detection using Natural Language Processing. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Valencia, Spain, 1–10.

Sharma, H. K., Kshitiz, K., Shailendra, 2018. NLP and Machine Learning Techniques for Detecting Insulting Comments on Social Networking Platforms. International Conference on Advances in Computing and Communication Engineering (ICACCE-2018). Paris, France, 22-23.

Sahay, K., Khaira, H. S., Kukreja, P., Shukla, N., 2018. Detecting Cyberbullying and Aggression in Social Commentary using NLP and Machine Learning. International Journal of Engineering Technology Science and Research (IJETSR). 5(1), 1428-1435. ISSN 2394 – 3386.

Salminen, J., Almerekhi, H., Milenkovic, M., Jung, S., An, J., Kwak, H., Jansen, B. J., 2018. Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM). 330-339.

Sigurbergsson, G. I., Derczynski, L., 2019. Offensive language and hate speech

detection For Danish. DOI:https://arXiv.org/abs/1908.04531

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

Sreelakshmi, K., Premjith, B., & Soman, K. P., 2020. Detection of Hate Speech Text in Hindi-English Code-mixed Data. Procedia Computer Science, 171, 737–744.

Twitter. (n.d.). Hateful conduct policy. URL: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Tsai, C. F., 2012. Bag-of-Words Representation in Image Annotation: A Review. ISRN Artif. Intell. 1–19.

Tarasova, N., 2016. Classification of Hate Tweets and Their Reasons using SVM. Master's Thesis. Uppsala Universitet.

Thompson, N., 2016. Anti-discriminatory Practice: Equality, Diversity and Social Justice. Palgrave Macmillan.

Umar, A., Bashir, S., Ochei, L. C., Adeyanju, I. A., 2019. Profiling Inappropriate Users' Tweets Using Deep Long Short-Term Memory (LSTM) Neural Network. I-Manager's Journal on Pattern Recognition. 5(4).

Uzochukwu, C. E., Okafor, E. G., 2019. Social Media, Hate Speech And Conflict: Interplay Of Influences. International Journal of Social Sciences and Humanities Reviews. 9(1), 144 – 158. ISSN: 2276-8645.

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., Hoste, V., 2018. Automatic detection of cyberbullying in social media text. PLoS ONE 13(10): e0203794.

Wendling, M., 2015. The year that angry won the internet. Retrieved 2nd March, 2020 from http://www.bbc.com/news/blogs-trending-35111707.

Wiegand, M., Ruppenhofer, J., Schmidt, A., Greenberg, C., 2018. Inducing a Lexicon of Abusive Words – a Feature-Based Approach. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 1, 1046–1056.

Watanabe, H. Bouazizi, M., Ohtsuki, T., 2018. Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. IEEE. 6. 13825- 13835.

Waseem, Z., Hovy, D., 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. Proc. NAACL Student Res. Work. 88–93.

Xiang, G., Fan, B., Wang, L., Hong, J., Rose, C., 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. Proc. 21st ACM Intl. Conf. Inf. Knowl. Mgt. – (CIKM).12.

Youtube. (n.d.). Hate speech policy. Retrieved 21st January, 2020, from https://support.google.com/youtube/answer/2801939?hl=en.#

Yenala, H., Jhanwar, A., Chinnakotla, M. K., & Goyal, J. (2018). Deep learning for detecting inappropriate content in text. International Journal of Data Science and Analytics. 6, 273–286.

Zenuni, X., Ajdari, J., Ismaili, F., Raufi, B., 2017. Automatic Hate Speech Detection In Online Contents Using Latent Semantic Analysis. PressAcademia Procedia. 368-371.