A survey and comparative sentiment analysis on Social Media Data and Its Application

Name[1*], Name[1*], Name[1*], Name[1*], Name[1*], Name[1*], Name[1*], Name[1*] and Jaya Sil[1*]

[1,2,3,4,5,6,7,8,9*]Dept. of CST, IIEST, Shibpur, P.O. - Botanic Garden, Howrah, 711103, West Bengal, India .

*Corresponding author(s). E-mail(s): email; email; email; email; email; email; email; email; js@cs.iiests.ac.in;

Abstract

Sentiments or opinions from social media provide the most up-to-date and inclusive information, due to the proliferation of social media and the low barrier for posting the message. Despite the growing importance of sentiment analysis, this area lacks a concise and systematic arrangement of prior efforts. It is essential to: (1) analyze its progress over the years, (2) provide an overview of the main advances achieved so far, and (3) outline remaining limitations. Several essential aspects, therefore, are addressed within the scope of this survey. Sentiment analysis is the computational examination of end user's opinion, attitudes and emotions towards a particular topic or product. Sentiment analysis classifes the message according to their polarity whether it is positive, negative, or neutral. Besides, this paper highlights the research challenges associated with predicting election results and open issues related to sentiment analysis. Further, this paper also suggests some future directions in respective election prediction using social media content.

Keywords: keyword1, Keyword2, Keyword3, Keyword4

1 Introduction

The management of sentiments, opinions, and subjective text is referred to as Sentiment analysis. Sentiment analysis, which examines several tweets and

reviews, provides understanding information on public opinions. It is a reliable method for forecasting a variety of important events, including the success of movies at the box office and general elections. The opinions can be classified as being negative, positive or neutral. The purpose of sentiment analysis is to automatically determine the expressive direction of user reviews . The demand of sentiment analysis is raised due to increase requirement of analyzing and structuring hidden information which comes from the social media in the form of unstructured data. 1) Features of Sentiment Analysis: Sentiment Analysis is a field that assesses sentiments in text as either positive or negative, through the use of techniques such as machine learning and lexicon-based methods. Machine learning techniques use algorithms such as Nave Bayes, Support Vector Machine (SVM), and Maximum Entropy to extract sentiments from text. Lexicon-based techniques rely on decision trees like k-Nearest Neighbors (k-NN) and Conditional Random Field (CRF).

Deep Learning is a subfield of machine learning that uses neural networks inspired by the human brain to model complex tasks. It has been applied in various fields including Natural Language Processing (NLP) and computer vision. The success of deep learning is due to improved hardware processing and advancements in machine learning algorithms.

The combination of Sentiment Analysis and Deep Learning has been successful in performing sentiment classification by using models like Recursive Neural Network (RNN), Dynamic Convolutional Neural Network (DyCNN), and Convolutional Neural Network (ConvNets). The recent works in this field include the use of models like Paragraph Vector, Long Short-Term Memory (LSTM), and ConvNets on characters.

2 Hate Speech

Hate speech detection is a complex method, even for the human mind, more so for computers. According to the UN, hate speech refers to "offensive discourse targeting a group or an individual based on inherent characteristics (such as race, religion or gender) and that may threaten social peace." Some elements like racism, violence, misogyny, and Islamophobia are highly related to hate speech. We have analyzed both classical machine learning algorithms and deep learning algorithms.

Hate speech detection is one type of text classification problem. There are different types of classifiers available. Here the main challenge is choosing the optimal classifier. But for that, we need to have complete understanding of each existing hate speech classifier. Machine learning is generally classified into classical method, Ensemble approach and deep learning method.

Comparison between classical machine learning and deep-learning models revealed that deep-learning models outperformed popular classifiers (NB, LR, RF, SVM) in most studies. An early work by Badjatiya et al. [21] has compared HS detection with different ML models (LR , SVM, GBDT) showed deep

learning models like LSTM or CNN performed 13-20% better than classical ML algorithms.

2.1 Deep Learning Records

We reviewed around 80 documents and analyzed each one of them on the basis of the architecture of the model. BERT, LSTM and CNN were the most popularly used algorithms, and also gave the best results.

Most architectures involved two steps: i) word embedding layer employing models like TfidfVectorizer, Word2vec; ii) deep learning layer.

The best algorithms for analysis varied from author to author. For example, Jahan, M.S. [1] found that CNN performed better than LSTM, while Badjatiya et al. [2] found that LSTM performed better than CNN. Many studies performed by different authors found that the combining two or more deep learning models usually give better performance than a single deep learning model [3] [4] [5] [6] [7]. For example, CNN+LSTM and CNN+GRU both performed better when compared to LSTM and CNN applied individually [3]. Zhou et al. [4] suggested a combination of three CNN models with different parameters can significantly improve the performance of the model.

BERT has been the most popular model for hate speech classification in the past 5 years. Several works explored BERT's performance in hate speech detection. For example, [8] [9] [10] [11] and almost all authors concluded that BERT is a better model. BERT achieves exceptional performance in multilingual datasets, as in [11] [12] [13] [14] [15].

In their work, Velankar et. al. [16] presented L3Cube-MahaHate. The dataset was extracted from Twitter, and then was labelled manually. The dataset consists of over 25000 texts and they labelled their dataset into four major classes i.e hate, offensive, profane, and not. They explored mono-lingual and multi-lingual variants of BERT like MahaBERT, IndicBERT, mBERT, and xlm-RoBERTa. They concluded with their results that mono-lingual models performed better than their multi-lingual counterparts. xlm-RoBERTa gave an accuracy of 0.894, while MahaBERT [17], a model trained on Marathi monolingual datasets, gave an accuracy of 0.909. Besides, some other language-specific BERT models developed over time for monolingual tasks outperformed multilingual model mBERT. For example, AraBERT (Arabic) [18],RuBERT (Russian) [19], AlBERTo (Italian) [12], BERTje (Dutch) [20], FinBERT (Finnish) [21], CamemBERT(French) [22], Flaubert (French [23]), BERT-CRF (Portuguese) [24], BERTje (Dutch) [20] and BERTtweet (A pre-trained language model for English Tweets) [25].

There have been very few reseaarches involving hate speech detection in a mixture of languages. A Hindi-English code mix dataset is released in the paper [26]. In the paper [27], the authors proposed text classification using Hinglish text written in Roman script. The data was collected randomly from the news and Facebook comments. They proposed various combinations of feature identification methods using TF-IDF representation and concluded that

Radial Basis Function Neural Network as the best combination to classify in the Hinglish text.

In the paper [28], the authors proposed DL (deep Learning) methods for detection of hate speech from Hindi-English code mix data on benchmark dataset. They experimented DL models using domain-specific embeddings and received results with accuracy 82.62%, precision 83.34 and F-score 80.85% with CNN model.

In this paper [29], Mathur et. al. created self made Hindi-English code mix dataset with annotation and applied ML models as baseline model. Finally, they proposed Multi-Channel Transfer Learning based model (MIMCT) and concluded that the proposed model outperforms state-of-art methods.

In another of their papers [30], Mathur et. al. introduce a novel tweet dataset, titled Hindi English Offensive Tweet (HEOT). They annotated the tweets manually into three categories: non offensive, abusive and hate-speech. Further, they used a CNN model and reported accuracy 83.90%, precision 80.20%, recall 69.98% and F1-score 71.45%.

Another major subset of hate speech is racism and sexism. The authors in [31] used a pre-trained word embedding [32] and applied max/mean pooling to form a fully connected transformation of these embeddings, the output then is fed to a 2 layers neural network followed with ReLU activation functions and a soft-max output layer. This method achieved a F1 score of 0.9241. Finally, the researchers in [33] proposed a classification model that consists of pre-trained word2vec features applied to a multi-layer CNN. The F1 score result was 78.3%.

Author, year	Platform	ML Approach	Features Representation	Algorithm	P	R	F
Karayiğit et. al. [34], 2023	Instagram (Turkish)	Supervised	-	LSTM, BiLSTM, mBERT	0.93	0.88	0.90
Badjatiya et al. [2], 2017	Twitter, 16k	Supervised	FastText, Random embedding, GloVe	CNN, LSTM, GBDT	.93	.93	.93
Rizos et al. [35],2019	Twitter, 24k	Supervised	FastText, Word2Vec, GloVe	CNN, LSTM, GRU	-	-	.69
Kamble and Joshi [28],2018	Twitter, 3.8k	Supervised	Word2Vec	LSTM, BiLSTM, CNN	.83	.78	.80
Ranasinghe et al. [9], 2019	Twitter	Supervised	FastText	LSTM, GRU, BERT	-	-	.78
Faris et al. [36],2020	Twitter, (Arabic)	Supervised	Word2Vec	Aravec CNN + LSTM	.65	.79	.71
Al-Hassan and Al Dossari [3] 2021, Springer	Twitter, 11k (Arabic)	Supervised	Keras word embedding	LSTM, GURU, CNN + GRU,CNN + LSTM	.72	.75	.73
Duwairi et al. [37] 2021, Springer	Twitter, 9k, 2k (Arabic)	Supervised	SG, CNN, CBOW	CNN, CNN + LSTM, BiLSTM + CNN	.74	-	-

Ali et al. [38], 2022	Twitter (20k) (Urdu)	Supervised	–	FastText, FastText + BiGRU, BERT, DistilBERT, XLMRoberta	0.76	0.65	0.67
Kovács et al. [39], 2021	Hindi-English	Supervised	–	RoBERTa, CNN + LSTM, FastText	-	-	0.85
Chopra et. al. [40], 2020	Twitter (Hindi-English)	Supervised	Keras Tokenizer	FT + CNN + BiLSTM + Attn + PV + DW + Debias	-	-	0.73
Gupta V, Sehra V, Vardhan YR, et al. [41], 2021	Twitter, 3k (Hindi English)	Supervised	–	GRU + attn.	0.87	0.87	0.87
Warner et. al. [42], 2012	-	Supervised	English	Yahoo! and the American Jewish Congress (AJC) (1,000 paragraphs)	0.59	0.68	0.63
Silva et. al. [43]	Twitter (27k), English	Supervised	–	-	-	-	-
Davidson et. al. [44]	Twitter (25k), English	Supervised	–	SVM	0.91	0.90	0.90

[45]	News website (10K), Brasilian Portuguese	Supervised	-	-	-	0.70	
[46]	Twitter (975 tweets), English	-	TFIDF Vectorizer	0.81	0.77	-	

2.2 Classical Machine Learning Records

This approach is also known as the shallow method. This method depends on automatically or manually coded dataset that is used for training purposes. This labeled dataset is used to train the learning algorithms to produce a model which can be used for detecting and classifying text as hate speech or non-hate. These classical Machine Learning algorithms include support vector machines (SVM), Naive Bayes (NB), Logistic Regress (LR), Decision Trees (DT), K-Nearest neighbor (KNN), etc [47]

Table 2: Comparison of Classical Machine Learning Methods

Author	Classifier	Novelty	Features Extraction	P	R	F
B. Vidgen and T. Yasseri [48]	NB, RF, LG, DT, SVM, DL	Improvement on is lamophobia detection	Word Embedding	0.778	0.773	0.776
Aljarah, M. Habib, N. Hijazi, H. Faris, R. Qaddoura, B. Hammo, et al.[49]	SVM,NB,DT,RF	Addresses Code-switch	TF-IDF	–	–	–
K. Nugroho, E. Noersasongko,et al.[50]	RF	Improved RF for HS detection	Count Vectors	0.711	0.722	0.713
R. Martins, M. Gomes, J. J. Almeida,et al. [46]	SVM,NB,RF	Emotional Analysis	N-gram	0.768	0.736	–
H. Watanabe, M. Bouazizi,et al. [51]	SVM, J48graft	Combination 3 different different datasets which give a wider coverage	Unigrams	0.68	0.65	0.66

P. Burnap and M. L. Williams [52]	n-Gram word	Identifying cyber Hate	BoW	0.96	0.97	0.96
Abozinadah et al. [53] 2015	Naïve Bayes	Abusive Hate Speech Detection	Bag of words, N-gram	0.85	0.85	0.85
Magdy et al.[54]	SVM	Terrorism(Pro-ISIS and Anti-ISIS)	Temporal patterns, Hashtags	0.87	0.87	0.87
Kaati et al. [55]	AdaBoost	Terrorism (Support or Oppose Jihadism)	Data dependent features and data independent features	-	-	-
Alshehri et al. [56]	SVM	Adult, Regular user	Lexicon, N-grams, bag-of means	0.92	0.65	0.76
Abozinadah [57]	SVM	Abusive HS Detection	Lexicon, bag of words (BOW), N-gram	0.96	0.96	0.96

Mubarak et al. [58]	Just performed extrinsic evaluation	Abusive, Ofensive	unigram and bigram, Log Odds Ratio (LOR), Seed Words lists None	0.98	0.45	0.60
Jaki and De Smedt [59]	K-means, single-layer averaged Perceptron	Radicalization (Muslim, Terrorist, Islamo fascistoid)	Skip grams and Character trigrams	0.84	0.83	0.84
Alakrot et al. [60]	SVM	Offensive, In-offensive	N-gram	0.88	0.80	0.82
Özel et al. [61]	M-Naïve Bayes	Hate	BOW	-	-	0.79
Alfina et al. [62]	Random Forest	Hate (Hate,Non-hate)	BOW and n-gram	-	-	0.93
Haidar et al. [63]	SVM	Cyber- bullying	Feature Vector	0.93	0.94	0.92
Abdelfatah et al. [64]	K-means clustering	Violent (Violent, Non-violent)	morphological features Vector Space Model	0.56	0.60	0.58
Fernandez and Alani [65]	SVM	Radicalization	Semantic Context	0.85	0.84	0.85
Wiegand et al. [66]	SVM	Abusive	word embedding	0.82	0.80	0.81

Another important type of hate speech includes comments about race and sex(racism and sexism respectively). [67] explores some papers that attempted to classify racism and sexism. Researchers in [68] tried a very simple approach by applying word uni-gram features to a Naive Bayes classifier. The model provides a large number of false-positive predictions because the unigram features do not map the relation between words, so every tweet with certain keywords is classified as racist irrespective of the context. Waseem and Hovy [69] compared many combinations of features and applied them to a simple logistic regression classifier, using a combination of unigrams, bigrams, trigrams, and quad-grams, in addition to the gender of the tweet writer provided the highest F1 score of 73.93%. Researchers in [70] used a combination of 1-3 word n-gram, 1-7 char ngram, in addition to sexism related lexicons. The tweets with those features were classified using the SVM classifier achieving an F1 score of 89% using one of the sexism datasets. Researchers in [71] used a non-common feature combination by combining a pre-processed 512-word embedding, TF-IDF, and 300-dimensional BoWV. The features were applied to a logistic regression classifier achieving an F1 score of 70.04%. Authors in paper [70] worked on the three datasets whose details are reported in [72] and [73]. On dataset [72], the use of char ngram as features and SVM as the classifier achieved the best results with accuracies of 78.77% and 75.44% respectively. On dataset [73], bag of words and sequences of words feature with SVM as the classifier provided the best accuracy with 89.32%.

3 Sentiment

Social media sentiment analysis involves collecting and analyzing opinions and emotions expressed on social media platforms about a specific brand, service, or product. It provides valuable insights for businesses looking to manage their image and brand reputation by gathering data on customers and competitors. This process is also known as opinion mining and is an essential part of any social media monitoring plan.

Sentiment analysis, a subtask of text classification, is used to identify subjective information and sentiments from different texts. It involves recognizing the emotions or intent behind a piece of text or speech. Common use cases include tracking customer feedback, improving customer service, and monitoring the impact of product or service changes on customer sentiment over time.

Sentiment can be categorized into positive, negative, and neutral classes. This area falls under the larger field of natural language processing and has been a popular topic in NLP since its inception. Various machine learning algorithms have been used in sentiment analysis, with the most recent and popular ones being CNN, LSTM, and Transformer models. Every data scientist should have a good understanding of sentiment analysis as it has completely transformed the way businesses operate, from opinion polls to creative marketing strategies.

Transformers are superior for text classification because of their ability to handle large amounts of sequential data effectively. They have a unique architecture that utilizes self-attention mechanisms, allowing them to capture long-term dependencies in text data. This results in better representations of the input text and improved performance in text classification tasks. Transformers have been shown to outperform traditional recurrent neural network (RNN) models in text classification and other NLP tasks, and have become the state-of-the-art models for NLP.

3.1 Deep Learning Records

In this study, we conducted a comprehensive review of several documents focusing on the architecture and features of the models used. Our analysis covered both deep learning and traditional machine learning algorithms. Among the popular algorithms employed, BERT, LSTM, and CNN were extensively used.

The analyzed architectures generally followed a two-step process. Firstly, a word embedding layer was applied, utilizing models such as TfidfVectorizer and Word2vec. The second step was the application of a deep learning layer, which was the core of the architecture. The use of these algorithms allowed for the creation of complex models that effectively analyzed and interpreted large amounts of text data.

It is worth noting that the success of deep learning models in NLP tasks is largely attributed to the ability of these models to capture complex relationships and patterns in text data. The use of word embedding and deep learning layers in the architectures reviewed further strengthen the ability of these models to handle and analyze large amounts of text data.

Table 3: Sentiment Analysis comparison based on methods and context

Author/ Year	Title	Method / Tools	Applica- tion / Result	Context	A	P	R
Yuliyanti, Djatna and Sukoco. (2017) [74]	Sentiment Mining of Community Development Program Evaluation Based on Social Media	Lexicon-based and machine learning	Success level of the community development program	Twitter	0.82	-	-
Martin-Domingo, Martin, and Mandsberg. (2019) [75]	Social media as a resource for sentiment analysis of Airport Service Quality	Machine learning	Analyse airport service quality	Twitter account	0.78	0.99	0.89
Mansour. (2018) [76]	Social Media Analysis of User's Responses to terrorism using sentiment analysis and text mining	Lexicon-based	Most user view ISIS as a threat and fear	Twitter	-	-	-
Saragih and Girsang. (2017) [77]	Sentiment Analysis of Customer Engagement on Social Media in Transport Online	Lexicon-based	Evaluate the business performance of online transport.	Face book and Twitter comments	-	-	-
Hassan, Hussain, Husain, Sadiq, Lee. (2017) [78]	Sentiment Analysis of Social Networking Sites (SNS) Data using Machine Learning Approach for the Measurement of Depression	Machine learning	Find the depression level of a person	Twitter and newsgroup	0.91	0.83	0.79

Joyce and Deng. (2017) [79]	Sentiment Analysis of Tweets for the 2016 US Presidential Election	Lexicon-based and machine learning	Calculate sentiment expressed and compare with polling data to see the correlation	Twitter	-	-	-
Ikoro, Harmina, Malik, and BatistaNavarro. (2018) [80]	Analyzing Sentiments Expressed on Twitter by UK Energy Company Consumers	Lexicon-based	Analyze energy provider company and the sentiment that users show	Twitter	-	-	-
Hao and Dai. (2016) [81]	Social media content and sentiment analysis on consumer security breaches	Lexicon-based	Security breaches can be detected in the early stages and prevent further destruction	Twitter	-	-	-
Shayaa, Wai, Chung, Sulaiman, Jaafar and Zakaria. (2017) [82]	Social Media Sentiment Analysis on Employment in Malaysia	Lexicon-based	Negative sentiment score on employment	Multiple channel social media	-	-	-
Isah, Trundle and Neagu. (2014) [83]	Social Media Analysis for Product Safety using Text Mining and Sentiment Analysis	Lexicon-based and machine learning	Monitor brand in order to act in even of a sudden rise in negative sentiment.	Facebook comment and Twitter	0.75	0.86	0.99

Ali, Dong, Bouguettaya, Erradi and Hadjidj. (2017) [84]	Sentiment Analysis as a Service: A social media-based sentiment analysis framework	Machine Learning	Identify the location of disease outbreaks	Twitter, Reddit, Instagram, news forum.	-	-	-
Laszlo Nemes (2020) [85]	Social media sentiment analysis based on COVID-19	RNN	Emotional manifestations in covid 19	Twitter	-	-	-
Idan, Lihi and Feigenbaum, Joan (2020) [86]	Show Me Your Friends, and I Will Tell You Whom You Vote for: Predicting Voting Behavior in Social Networks	Naive Bayes	using a novel Bayesian-network model that combines demographic, behavioral, and social features; we apply this novel approach to the 2016 U.S. Presidential election	Facebook	0.82	0.8	0.87
Sayyida et. al. (2022) [87]	Transformer-based deep learning models for the sentiment analysis of social media data	Transformer	propose a generalized SA model that can handle noisy data, OOV words, sentimental and contextual loss of reviews data	Social Media	0.9	-	0.91

Alwakid et. al. (2022) [88]	MULDASA: Multifactor Lexical Sentiment Analysis of Social-Media Content in Nonstandard Arabic Social Media	Lexicon Based	define a novel lexical sentiment analysis approach for studying Arabic language tweets (TTs) from specialized digital media platforms	Twitter	0.86	0.89	-
Qian et. al. (2022) [89]	Understanding public opinions on social media for financial sentiment analysis using AI-based techniques	Deep Neural Network	Determine the reasons for the growing acceptance of NFTs through sentiment and emotion analysis	Twitter	-	-	-
Arias et. al. [90]	Sentiment Analysis of Public Social Media as a Tool for Health-Related Topics	Deep Learning	Serve as an introduction to the methods, challenges, and implications of SA technology for disciplines different or adjacent to computer science	Twitter, Instagram, Youtube, etc	-	-	-
Zhigang Jin et. al. (2022) [91]	Social Media Sentiment Analysis Based on Dependency Graph and Co-occurrence Graph	Multi-feature hierarchical graph attention model (MH-GAT)	This article takes diverse structural information, part of speech, and position association information into consideration simultaneously	Weibo Dataset	-	-	-

Chandrasekaran et. al. (2022) [92]	Visual Sentiment Analysis Using Deep Learning Models with Social Media Data	Deep Learning	fine-tuned transfer learning models to handle the issues of image sentiment analysis.	Facebook, Instagram, Youtube, etc	0.89	0.86	0.88
Hassan et. al. (2022) [93]	Visual Sentiment Analysis from Disaster Images in Social Media	Deep Learning	propose a deep visual sentiment analyzer for disaster-related images	Twitter	0.73	0.63	0.8
Sufi, Fahim K. and Khalil, Ibrahim (2022) [94]	Automated Disaster Monitoring From Social Media Posts Using AI-Based Location Intelligence and Sentiment Analysis	AI based	present a new fully automated algorithm based on AI and NLP, for extraction of location-oriented public sentiments on global disaster situation	Twitter	-	-	-

3.2 Classical Machine Learning Records

This approach refers to detection of sentiment of a particular word segment using classical Machine Learning Algorithms. The two main methods covered in this section are classification with lexicons and standalone machine learning algorithms and ensemble learning. Various classifiers Naive Bayes, Support Vector Machine, Maximum Entropy, K-Nearest Neighbours, Logistic Regression, are used. The results suggests that deep learning algorithms give superior results than classical machine learning algorithms in general. The best accuracy obtained among the classical records is 88.34 [95].

Classification with lexicons and standalone learning algorithms

Study	Feature Set	Lexicon	Classifier	Dataset	A	P	F
Read [96] (2005)	N-Gram	Emotions	Naive Bayes and SVM	Read [96]	82.9	-	-
Go et al.[97] (2009)	N-gram and POS	–	Naive Bayes, Maximum Entropy, and SVM	Go et al. [97]	-	-	70
Davidov et al. [98] (2010)	Punctuation, n-grams, patterns, and tweet-based features	–	KNN	O'Connor et al. [99]	-	-	86
Zhang et al.[100] (2011)	N-gram, emoticons and hashtags	Ding et al.[101]	SVM	Zhang et al. [100]	85.4	68.7	74.9
Agarwal et al. [102] (2011)	POS, Lexicon, percentage of capitalized text, exclamation, capitalized text	Emoticons listed from Wikipedia, anacronym dictionary	SVM	Agarwal et al. [102]	60.5	-	60.2
Speriosu et al. [103] (2011)	N-gram, hashtags, emoticons, lexicon and Twitter follower graph	Wilson et al.[104]	Maximum Entropy	Go et al.[97] and Speriosu et al.[103]	71.2	-	-
Saif et al. [105] (2012)	N-gram, POS and semantic features	–	Naive Bayes	Go et al.[97], Speriosu et al.[103] and Shamma et al. [106]	83.9	84.2	83.9

Study	Feature Set	Lexicon	Classifier	Dataset	A	P	F
Hu et al. [107] (2013)	N-gram, POS, Data Representation of Social Relations	-	-	Go et al. [97] and Shamma et al. [106]	79.6	-	-
Saif et al. [95] (2013)	N-gram, capitalized text, POS, lexicons	Mohammand and Yang[108], Wilson et al.[104], Hu and Liu[109], and other lexicons constructed from hashtags	SVM	Nakov et al [110].	88.34	-	-

Ensemble Learning

Study	Feature Set	Base Learner	Ensemble Mehods	Dataset	A	P	F
Lin and Kolcz [111] (2012)	Feature Hashing	Logistic Regression classifier	Majority vote	Private dataset [111].	-	-	-
Rodrı gues et al. [112] (2013)	N-gram, lexicon, POS, tweet-based features and Senti-Wordnet	CRF, SVM and heuristic method	Majority vote, upper bound, ensemble vote	Nakov et al [110].	-	-	-
Clark et al. [113] (2013)	N-gram, lexicon and polarity strength	Naive Bayes	Weighted voting scheme	Nakov et al [110].	-	-	-
Hassan et al. [114] (2013)	A combination of unigrams and bigrams of simple words, part-of-speach and semantic features derived from WordNet [115] and SentiWordNet 3.0 [116]	RBF Neural Network, Random Tree, REP Tree, Naive Bayes, Bayes Net, Logistic Regression and SVM.	A bootstrap model by combining dataset, feature and classifier parameters	Sanders - Twitter Sentiment Corpus	-	-	-

4 Profane

Profanity and offensive language on social media platforms, such as Twitter, have been the subject of increasing concern and study in recent years. According to the Wikipedia, profanity is defined as "cursing, cussing, swearing, bad

language, foul language, obscenities, expletives or vulgarism, is a socially offensive use of language." Research has shown that profanity is relatively common on Twitter, with some studies finding that as much as 10-20% of all tweets contain profanity.

Recent studies have shown that deep learning models can achieve high accuracy in detecting profanity and offensive language on twitter as compared to classical machine learning models, with some studies reporting accuracy rates of over 90%. Furthermore, they have been applied to a wide range of use cases such as detecting hate speech, cyberbullying, sarcasm and irony detection, and even detecting mental health issues.

4.1 Deep Learning Records

Deep learning techniques have been increasingly used in recent years to analyze profanity and offensive language on Twitter. These techniques, which include neural networks such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have been shown to be highly effective in identifying and classifying profanity in tweets.

In the field of detecting offensive language, early studies utilized machine learning based classifiers. For example, Warner and Hirschberg (2012) and Burnap and Williams (2015) were pioneers in this area. Another study by Djuric et al. (2015) added representation through word embeddings, as introduced by Mikolov et al. (2013). Nobata et al. (2016) utilized a combination of pre-defined language elements and word embeddings to create a regression model. Waseem (2016) employed logistic regression along with n-grams and user-specific factors such as gender and location. Davidson et al. (2017) delved deeper into various forms of abusive language. Badjatiya et al. (2017) investigated the use of deep learning models, specifically ensemble gradient boost classifiers, for multi-class classification of sexist and racist language.

We have analyzed documents employing deep learning and basic machine learning algorithms. BERT, LSTM and CNN were the most popular algorithms used in the architectures.Furthermore, in recent years few more extensive research has been performed.

Table 7 Records in Profane and their comparison on the basis of Accuracy(A), Precision(P), Recall(R) and F1 score(F)

Author/ Year	Title	Method / Tools	Context	A	P	R	F1
Phoey Lee Teh and Chi-Bin Cheng (2020) [117]	Profanity and Hate Speech Detection	LSTM, BLSTM and BERT	Twitter	-	-	-	-
Raktim Chatterjee, Sukanya Bhattacharya, and Soumyajeet Kabi. (2021) [118]	Profanity detection in social media text using a hybrid approach of NLP and machine learning	NLP and Machine Learning	Twitter	0.918	0.917	0.702	0.77
Pratik Ratadiya and Deepak Mishra. (2019) [119]	An attention ensemble based approach for multilabel profanity detection	Attention and Bi-GRU	Social Media	0.974	0.82	0.84	0.76
Ji Ho Park and Pascale Fung. (2017) [120]	One-step and Two-step Classification for Abusive Language Detection on Twitter	LR, SVM, FastText, CharCNN, HybridCNN, WordCNN, HybridCNN	Twitter	-	0.880	0.859	0.869
Basavraj Chinagundi, Muskaan Singh, Tirthankar Ghosal, Prashant Singh Rana and Guneet Singh Kohli. (2021) [121]	Classification of Hate, Offensive and Profane content from Tweets using an Ensemble of Deep Contextualized and Domain Specific Representations	NB, LR, KNN, SVM, DT, RF, Bagging, AdaBoost, Voting, GloVe, ERNIE 2.0, TwitterRobertaOffensive, HateBERT	Twitter	0.81	0.81	0.79	0.81
Levent Soykan, Cihan Karsak, ˙Ilknur Durgar Elkahlout and Burak Aytan. (2022) [122]	A Comparison of Machine Learning Techniques for Turkish Profanity Detection	Logistic-Regression, SGDClassifier, LinearSVC, RandomForestClassifier, LSTM, BERT, Electra, T5	Social Media	0.98	0.98	0.87	0.93
Dadvar, Maral and Eckert, Kai (2018) [123]	Cyberbullying detection in social networks using deep learning based models; a reproducibility study	Deep Learning	Social Media	-	0.99	0.99	0.97

Author/ Year	Title	Method / Tools	Context	A	P	R	F1
Ba Wazir, Abdulaziz Saleh and Karim, Hezerul Abdul and Abdullah, Mohd Haris Lye and AlDahoul, Nouar and Mansor, Sarina and Fauzi, Mohammad Faizal Ahmad and See, John and Naim, Ahmad Syazwan. (2021) [124]	Design and Implementation of Fast Spoken Foul Language Recognition with Different End-to-End Deep Neural Network Architectures	CNN, RNN, LSTM	Selection of profane language collected through the recordings of various areas and environments	-	0.9785	-	0.9765
Kim, Cheong-Ghil and Hwang, Young-Jun and Kamyod, Chayapol. (2022) [125]	A Study of Profanity Effect in Sentiment Analysis on Natural Language Processing Using ANN	LSTM	Movie Reviews	0.834	-	-	-
Kumari, Kirti and Singh, Jyoti Prakash. (2019) [126]	AI ML NIT Patna at HASOC 2019: Deep Learning Approach for Identification of Abusive Content	One-hot, GloVe, fastText Embeddings followed by CNN	Social Media	-	-	-	0.7834
Gottipati, Swapna and Tan, Annabel and Chow, David and Shan, Jing and Lim, Joel and Kiat, Wee. (2020) [127]	Leveraging Profanity for Insincere Content Detection-A Neural Network Approach	Profanity based Sincerity Classifier based on fastText	Social Media	-	-	-	0.9507
Hsu Yang and Chuan-Jie Lin (2020) [128]	TOCP: A Dataset for Chinese Profanity Processing	CNN, BiLSTM	Social Media	-	0.852	0.875	0.863

Author/ Year	Title	Method / Tools	Context	A	P	R	F1
Woo, Jiyoung and Park, Sung Hee and Kim, Huy Kang. (2022) [129]	Profane or Not: Improving Korean Profane Detection using Deep Learning	CNN	Social Media	0.904	0.9222	0.9392	0.9268
Sazzed, Salim. (2021) [130]	BengSentiLex and BengSwearLex: creating lexicons for sentiment analysis and profanity detection in low-resource Bengali language	LR, SVM, SGD, CNN, LSTM, BiLSTM	Social Media	-	-	-	-
Al-Hashedi, Mohammed and Soon, Lay-Ki and Goh, Hui-Ngo. (2019) [131]	Cyberbullying detection using deep learning and word embeddings: An empirical study	GRU, LSTM and BLSTM	Social Media	-	-	0.98	-
Nobata, Chikashi and Tetreault, Joel and Thomas, Achint and Mehdad, Yashar and Chang, Yi. (2016) [132]	Abusive language detection in online user content	NLP	Social Media	-	0.837	0.842	0.839
Bhowmick, Rajat Subhra and Ganguli, Isha and Paul, Jayanta and Sil, Jaya (2021) [133]	A multimodal deep framework for derogatory social media post identification of a recognized person	Distil-BERT, ELECTRA, XLM-RoBERTa	Twitter	0.908	-	-	-
Malik, Pranav and Aggrawal, Aditi and Vishwakarma, Dinesh K (2021) [134]	Toxic speech detection using traditional machine learning models and bert and fasttext embedding with deep neural networks	LR, SVM, DT, RF, XGBoost, CNN, MLP, LSTM	Social Media	0.82	0.83	0.82	0.81
Marwa, Tolba and Salima, Ouadfel and Souham, Meshoul (2018) [135]	Deep learning for online harassment detection in tweets	SVM, Naive Bayes, CNN, LSTM, BLSTM	Social Media	-	0.80	0.80	0.71

Many researchers have used different approaches to build profanity detection models that help in blocking, filtering, or alerting the users about the abusive language used on these platforms. Among the various methods used for profanity detection, deep learning techniques have shown promising results. The methods and tools used in the research papers include NLP, machine learning, LSTM, BLSTM, BERT, Attention and Bi-GRU, LR, SVM, FastText, CharCNN, HybridCNN, WordCNN, NB, KNN, DT, RF, Bagging, AdaBoost, GloVe, ERNIE 2.0, TwitterRoBERTaOffensive, HateBERT, Logistic-Regression, SGDClassifier, LinearSVC, RandomForestClassifier, CNN, RNN, and one-hot. The research papers analyzed different social media platforms like Twitter and movie reviews in various languages, including Turkish and Chinese. Among these, the highest F1 score was achieved by Levent Soykan et al. (2022) with 0.93 in Turkish Profanity Detection using Logistic Regression, SGD Classifier, Linear SVC, Random Forest Classifier, LSTM, BERT, Electra, and T5 techniques. The highest Recall was 0.99 achieved by Dadvar, Maral, and Eckert Kai (2018) for Cyberbullying detection in social networks using deep learning-based models. The highest Precision was 0.99 achieved by the same authors for the same task.It is evident from the studies that deep learning models can be used successfully for profanity detection. The studies that compared multiple models showed that ensembling models can provide better accuracy and F1 scores than individual models.

4.2 Classical Machine Learning Records

Modern social media is an open platform and people often misuse the freedom. A major display of profanity is cyberbullying. yberbullying is a huge phenomenon among teenagers as a victim or predator or bystander [5]. Authors in [136] have used a dataset from Twitter, which has seen maximum instances of cyberbullying. They created a dataset of around 1k data points and manually labelled them. They used SVM along with TF-IDF and obtained a F1-score of 75%. In another paper [127], authors performed profanity analysis on a dataset obtained from Quora. They achieved a F1-score of 0.591 using Logistic Regression and 0.742 using fastText. Some of the research papers which are exclusively based on Machine Learning are mentioned below.

Author/ Year	Title	Method / Tools	Context	A	P	R	F1
Haidar, Batoul and Chamoun, Maroun and Serhrouchni, Ahmed (2017) [63]	A multilingual system for cyberbullying detection: Arabic content detection using machine learning	Naive Bayes, SVM	Social Media	-	0.934	0.941	0.927
Chaudhari, Apoorva and Davda, Palak and Dand, Monil and Dholay, Surekha (2021) [137]	Profanity Detection and Removal in Videos using Machine Learning	CharCNN, Word-CNN and Hybrid-CNN	Social Media	0.824	-	-	-
Sood, Sara and Antin, Judd and Churchill, Elizabeth (2012) [138]	Profanity use in online communities	Machine Learning	Social news sites	0.913	0.636	0.412	0.457
Sood, Sara Owsley and Antin, Judd and Churchill, Elizabeth (2012) [139]	Using crowdsourcing to improve profanity detection	SVM	Social News sites	0.94	0.90	0.64	0.63
Chin, Hyojin and Kim, Jayong and Kim, Yoonjong and Shin, Jinseop and Yi, Mun Y (2018) [140]	Explicit content detection in music lyrics using machine learning	Naive Bayes, Decision Tree, SVM, MCES	Social Media	-	-	-	-

Machine learning algorithms such as Naive Bayes and SVM have proven to be effective in detecting profanity and cyberbullying in social media and online communities. However, the effectiveness of these methods largely depends on the quality and quantity of data used for training and testing. Crowdsourcing can improve the accuracy of the models by providing a more diverse set of data. Additionally, CNN-based models such as CharCNN, WordCNN, and Hybrid-CNN can also be effective for detecting profanity in videos, but their effectiveness needs to be further investigated.

5 Targeted Insult

Targeted insult on social media refers to the act of directing abusive or offensive language or behavior towards a specific individual or group of people. It can take the form of name-calling, harassment, or even threats of violence. Statistics on targeted insult in social media can vary depending on the platform, the population being studied, and the timeframe. However, research has shown that women and marginalized groups are disproportionately affected by targeted insult on social media. Additionally, studies have found that the anonymity and lack of accountability on social media can lead to an increase in the frequency and severity of the targeted insult.

5.1 Classical Machine Learning Records

This approach is also known as the shallow method. This method depends on an automatically or manually coded dataset that is used for training purposes. This dataset is used to train the learning algorithms to produce a model which can be used for detecting and classifying text as targeted insult or non. These classical Machine Learning algorithms include support vector machines (SVM), Naive Bayes (NB), Logistic Regress (LR) etc

Author/ Year	Title	Classifier	Dataset	A	P	R	F1
R Sodhi, K Pant, R Mamidi [141]	Jibes & Delights: A Dataset of Targeted Insults and Compliments to Tackle Online Abuse	Logistic Regression, SVM	JBC Dataset	0.875	0.980	0.897	0.883
S Kurniawan, I Budi [142]	Indonesian tweets hate speech target classification using machine learning	NB, SVM	Facebook Dataset	0.77	-	-	0.84
Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini [143]	IIITT@ DravidianLangTech-EACL2021: Transfer learning for offensive language detection in Dravidian languages	Logistic Regression, SVM	Hindi-English Offensive Tweet (HEOT) dataset	-	-	-	0.96
F Alkomah, X Ma [144]	A Literature Review of Textual Hate Speech Detection Methods and Datasets	SVM	Waseem's dataset	0.704	0.76	-	0.823
Md Fahim; Swapna S. Gokhale [145]	Detecting Offensive Content on Twitter During Proud Boys Riots	Logistic Regression, Naive Bayes, SVM	Private Dataset	0.87	0.88	0.93	0.88
C Raj, A Agarwal, G Bharathy, B Narayan, M Prasad [146]	Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques	SVM	Wikipedia Attack Dataset	0.83	-	-	0.98

Author/ Year	Title	Classifier	Dataset	A	P	R	F1
Hajime Watanabe; Mondher Bouazizi [51]	Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection	SVM	Crowdflower	0.87	-	-	-
Kelly Reynolds; April Kontostathis; Lynne Edwards [147]	Using Machine Learning to Detect Cyberbullying	SVM	Private Dataset	0.81	-	-	-
Lázaro Villa, José Antonio Nugues, Pierre [148]	Identifying and categorizing offensive language in tweets using Machine Learning	SVM	Offensive Language Identification dataset (OLID)	0.89	-	-	0.66
Viviana Cotik, Natalia Debandi [149]	A study of Hate Speech in Social Media during the COVID-19 outbreak	Linear Classifier	Waseem and Hovy Dataset	-	-	-	0.75

Mostly SVM has been used to classify the text in the papers. Authors in [143] have used a dataset which has multi-lingual dataset and has both Hindi and English Language. Out of all the papers [148] has the highest accuracy of all with 0.89 and [141] has the highest precision of 0.98 and F1-score of 0.883.

5.2 Deep Learning Records

Deep learning techniques have been increasingly used in recent years to analyze profanity and offensive language on Twitter. These techniques, which include neural networks such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have been shown to be highly effective in identifying

and classifying profanity in tweets. We have analyzed documents employing deep learning and basic machine learning algorithms. BERT, LSTM and CNN were the most popular algorithms used in the architectures.

Table 10 Records in Targeted Insult detection and their comparison on the basis of Accuracy(A), Precision(P), Recall(R) and F1 score(F)

Author/ Year	Title	Method / Tools	Context	A	P	R	F1
Maral Dadvar, Kai Eckert [123]	Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study	BLSTM	Twitter	0.94	0.93	0.93	0.96
M Alotaibi, B Alotaibi, A Razaque [150]	A multichannel deep learning framework for cyberbullying detection on social media	NLP	Twitter	0.89	0.9	0.89	0.89
M Zampieri, S Malmasi, P Nakov [151]	Predicting the type and target of offensive posts in social media	Deep Learning	Social Media	-	0.79	0.89	0.80
S Sharifirad [152]	Nlp and machine learning techniques to detect online harassment on social networking platforms	BiLSTM	Twitter	0.84	0.83	0.84	0.84
K Shanmugavadivel, VE Sathishkumar [153]	Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data	NLP	Social Media	0.79	-	-	0.8

Table 11 Records in Targeted Insult detection and their comparison on the basis of Accuracy(A), Precision(P), Recall(R) and F1 score(F)

Author/ Year	Title	Method / Tools	Context	A	P	R	F1
A Omar, TM Mahmoud, T Abd-El-Hafeez, A Mahfouz [154]	Multi-label arabic text classification in online social networks	Deep Learning	Twitter	0.97	0.97	0.97	0.97
A Kalaivani, D Thenmozhi [155]	SSN-NLP-MLRG at SemEval-2020 task 12: Offensive language identification in English, Danish, Greek using BERT and machine learning approach	BERT	Social Media	-	0.8	0.81	0.77
MSA Sanoussi, C Xiaohua [156]	Detection of Hate Speech Texts Using Machine Learning Algorithm	NLP	Facebook	0.95	-	-	-
Thenmozhi D., Senthil Kumar B., Srinethe Sharavanan, Aravindan Chandraboset [157]	SSN-NLP at SemEval-2019 task 6: Offensive language identification in social media using traditional and deep machine learning approaches	NLP	Social Media	0.83	-	-	0.53
HA Nayel, HL Shashirekha [158]	DEEP at HASOC2019: A Machine Learning Framework for Hate Speech and Offensive Language Detection	Deep Learning	Social Media	-	0.91	0.90	0.90
S Alsafari, S Sadaoui, M Mouhoub [159]	Hate and offensive speech detection on Arabic social media	CNN, RNN	Social Media	-	0.81	0.84	0.82

Table 12 Records in Targeted Insult detection and their comparison on the basis of Accuracy(A), Precision(P), Recall(R) and F1 score(F)

Author/ Year	Title	Method / Tools	Context	A	P	R	F1
A Parikh, H Desai, AS Bisht [160]	DA Master at HASOC 2019: Identification of Hate Speech using Machine Learning and Deep Learning approaches for social media post	CNN, BLSTM	Twitter	-	-	-	0.64
S Thara, P Poornachandran [161]	SSN-NLP at SemEval-2019 task 6: Offensive language identification in social media using traditional and deep machine learning approaches	BiLSTM	Social Media	0.83	-	-	0.53
TL Sutejo, DP Lestari [162]	Indonesia hate speech detection using deep learning	Deep Learning	Social Media	-	-	- 0.87	
B Bharathi [163]	SSNCSE-NLP@ DravidianLangTech-EACL2021: Offensive language identification on multilingual code mixing text	BERT	Social Media	0.95	0.87	0.85	0.95
Z Zhang, D Robinson, J Tepper [164]	Detecting hate speech on twitter using a convolution-gru based deep neural network	GRU Based Deep Neural Network	Twitter	-	-	-	0.94

Mostly LSTM and NLP has been used to predict whether the text in the papers are hateful or not. Authors in [154], [164], [160], [152], [123], [150] have used a Twitter dataset. Out of all the papers [154] has the highest accuracy, precision and F1-score of all with 0.97, [163] has the second highest accuracy of 0.95 and [123] has the second highest precision 0.93.

6 Discussion

The article is on the application of social media data in detecting profanity and offensive language. The article mentions the use of both machine learning and deep learning algorithms, such as RNN, CNN, BERT, LSTM, etc. The algorithms have been applied on Twitter and other social media platforms. The article also mentions the use of hybrid approaches combining NLP and machine learning. The authors have analyzed recent research and mention that deep learning-based models have become more popular in recent years. The paper usually consists of several key components: an introduction, literature review, methodology, results, discussion, conclusion, and references. The introduction provides an overview of the topic, sets the context and defines the purpose of the research. The literature review synthesizes previous studies related to the research problem. The methodology section describes the research design, data collection and analysis procedures. The results section presents the findings of the research. The discussion interprets the results, relating them back to the research questions and the literature review. The conclusion summarizes the main findings, highlights their significance, and provides recommendations for future research. The references list all the sources used in the paper.

7 Conclusion

In conclusion, the research project on hate speech detection on Twitter has shown the need for accurate and efficient algorithms to address the growing issue of online hate speech. The results obtained from the experiments indicate that there is still room for improvement in terms of precision and recall. However, the models tested in this project have shown promising results and can serve as a starting point for further research in this area. It is imperative that we continue to develop and improve these models to effectively tackle hate speech on social media platforms and promote a safe and inclusive online environment.The future prospects of the hate speech detection project are promising, as there is scope for both enlarging the dataset and improving accuracy. By expanding the dataset, the models will be able to learn from a wider range of hate speech, leading to better generalization. Additionally, exploring advanced deep learning techniques and ensembling multiple models has the potential to significantly improve accuracy. These improvements can contribute to creating a safer and more inclusive online environment.

References

[1] Jahan, M.S., Oussalah, M.: A systematic review of hate speech automatic detection using natural language processing. arXiv preprint arXiv:2106.00742 (2021)

[2] 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, pp. 759–760 (2017)

[3] Al-Hassan, A., Al-Dossari, H.: Detection of hate speech in arabic tweets using deep learning. Multimedia Systems, 1–12 (2021)

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

[5] Kapil, P., Ekbal, A.: A deep neural network based multi-task learning approach to hate speech detection. Knowledge-Based Systems **210**, 106458 (2020)

[6] Shruthi, P., KM, A.K.: Novel approach for generating hybrid features set to effectively identify hate speech. Inteligencia Artificial **23**(66), 97–111 (2020)

[7] Kumar, A., Abirami, S., Trueman, T.E., Cambria, E.: Comment toxicity detection via a multichannel convolutional bidirectional gated recurrent unit. Neurocomputing **441**, 272–278 (2021)

[8] Nikolov, A., Radivchev, V.: Nikolov-radivchev at semeval-2019 task 6: Offensive tweet classification with bert and ensembles. In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 691–695 (2019)

[9] Ranasinghe, T., Zampieri, M., Hettiarachchi, H.: Brums at hasoc 2019: Deep learning models for multilingual hate speech and offensive language identification. In: FIRE (working Notes), pp. 199–207 (2019)

[10] Saleh Alatawi, H., Maatog Alhothali, A., Mustafa Moria, K.: Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. arXiv e-prints, 2010 (2020)

[11] Dowlagar, S., Mamidi, R.: Hasocone@ fire-hasoc2020: Using bert and multilingual bert models for hate speech detection. arXiv preprint arXiv:2101.09007 (2021)

[12] Polignano, M., Basile, P., De Gemmis, M., Semeraro, G., Basile, V.:

Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In: 6th Italian Conference on Computational Linguistics, CLiC-it 2019, vol. 2481, pp. 1–6 (2019). CEUR

[13] Alami, H., El Alaoui, S.O., Benlahbib, A., En-nahnahi, N.: Lisac fsdm-usmba team at semeval-2020 task 12: Overcoming arabert's pretrain-finetune discrepancy for arabic offensive language identification. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 2080–2085 (2020)

[14] Sai, S., Sharma, Y.: Siva@ hasoc-dravidian-codemix-fire-2020: Multilingual offensive speech detection in code-mixed and romanized text. In: FIRE (Working Notes), pp. 336–343 (2020)

[15] Wang, S., Liu, J., Ouyang, X., Sun, Y.: Galileo at semeval-2020 task 12: Multi-lingual learning for offensive language identification using pre-trained language models. arXiv preprint arXiv:2010.03542 (2020)

[16] Velankar, A., Patil, H., Gore, A., Salunke, S., Joshi, R.: L3cube-mahahate: A tweet-based marathi hate speech detection dataset and bert models. arXiv preprint arXiv:2203.13778 (2022)

[17] Joshi, R.: L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources. arXiv preprint arXiv:2202.01159 (2022)

[18] Antoun, W., Baly, F., Hajj, H.: Arabert: Transformer-based model for arabic language understanding. arXiv preprint arXiv:2003.00104 (2020)

[19] Kuratov, Y., Arkhipov, M.: Adaptation of deep bidirectional multilingual transformers for russian language. arXiv preprint arXiv:1905.07213 (2019)

[20] de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., Nissim, M.: Bertje: A dutch bert model. arXiv preprint arXiv:1912.09582 (2019)

[21] Araci, D.F., Genc, Z.: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063 (2019)

[22] Martin, L., Muller, B., Suárez, P.J.O., Dupont, Y., Romary, L., de La Clergerie, É.V., Seddah, D., Sagot, B.: Camembert: a tasty french language model. arXiv preprint arXiv:1911.03894 (2019)

[23] Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., Schwab, D.: Flaubert: Unsupervised language model pre-training for french. arXiv preprint

arXiv:1912.05372 (2019)

[24] Souza, F., Nogueira, R., Lotufo, R.: Portuguese named entity recognition using bert-crf. arXiv preprint arXiv:1909.10649 (2019)

[25] Nguyen, D.Q., Vu, T., Nguyen, A.T.: Bertweet: A pre-trained language model for english tweets. arXiv preprint arXiv:2005.10200 (2020)

[26] Kumar, R., Reganti, A.N., Bhatia, A., Maheshwari, T.: Aggression-annotated corpus of hindi-english code-mixed data. arXiv preprint arXiv:1803.09402 (2018)

[27] Ravi, K., Ravi, V.: Sentiment classification of hinglish text. In: 2016 3rd International Conference on Recent Advances in Information Technology (RAIT), pp. 641–645 (2016). IEEE

[28] Kamble, S., Joshi, A.: Hate speech detection from code-mixed hindi-english tweets using deep learning models. arXiv preprint arXiv:1811.05145 (2018)

[29] Mathur, P., Sawhney, R., Ayyar, M., Shah, R.: Did you offend me? classification of offensive tweets in hinglish language. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pp. 138–148 (2018)

[30] Mathur, P., Shah, R., Sawhney, R., Mahata, D.: Detecting offensive tweets in hindi-english code-switched language. In: Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, pp. 18–26 (2018)

[31] Kshirsagar, R., Cukuvac, T., McKeown, K., McGregor, S.: Predictive embeddings for hate speech detection on twitter. arXiv preprint arXiv:1809.10644 (2018)

[32] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)

[33] Gambäck, B., Sikdar, U.K.: Using convolutional neural networks to classify hate-speech. In: Proceedings of the First Workshop on Abusive Language Online, pp. 85–90 (2017)

[34] Karayiğit, H., Akdagli, A., Aci, Ç.İ.: Homophobic and hate speech detection using multilingual-bert model on turkish social media. Information Technology and Control **51**(2), 356–375 (2022)

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

[36] Faris, H., Aljarah, I., Habib, M., Castillo, P.A.: Hate speech detection using word embedding and deep learning in the arabic language context. In: ICPRAM, pp. 453–460 (2020)

[37] Duwairi, R., Hayajneh, A., Quwaider, M.: A deep learning framework for automatic detection of hate speech embedded in arabic tweets. Arabian Journal for Science and Engineering **46**(4), 4001–4014 (2021)

[38] Ali, R., Farooq, U., Arshad, U., Shahzad, W., Beg, M.O.: Hate speech detection on twitter using transfer learning. Computer Speech & Language **74**, 101365 (2022)

[39] Kovács, G., Alonso, P., Saini, R.: Challenges of hate speech detection in social media. SN Computer Science **2**(2), 1–15 (2021)

[40] Chopra, S., Sawhney, R., Mathur, P., Shah, R.R.: Hindi-english hate speech detection: Author profiling, debiasing, and practical perspectives. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 386–393 (2020)

[41] Gupta, V., Sehra, V., Vardhan, Y.R., *et al.*: Hindi-english code mixed hate speech detection using character level embeddings. In: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 1112–1118 (2021). IEEE

[42] Warner, W., Hirschberg, J.: Detecting hate speech on the world wide web. In: Proceedings of the Second Workshop on Language in Social Media, pp. 19–26 (2012)

[43] Silva, L., Mondal, M., Correa, D., Benevenuto, F., Weber, I.: Analyzing the targets of hate in online social media. In: Tenth International AAAI Conference on Web and Social Media (2016)

[44] Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, pp. 512–515 (2017)

[45] de Pelle, R.P., Moreira, V.P.: Offensive comments in the brazilian web: a dataset and baseline results. In: Anais do VI Brazilian Workshop on Social Network Analysis and Mining (2017). SBC

[46] Martins, R., Gomes, M., Almeida, J.J., Novais, P., Henriques, P.: Hate speech classification in social media using emotional analysis. In: 2018 7th Brazilian Conference on Intelligent Systems (BRACIS), pp. 61–66 (2018). IEEE

[47] Mullah, N.S., Zainon, W.M.N.W.: Advances in machine learning algorithms for hate speech detection in social media: a review. IEEE Access **9**, 88364–88376 (2021)

[48] Vidgen, B., Yasseri, T.: Detecting weak and strong islamophobic hate speech on social media. Journal of Information Technology & Politics **17**(1), 66–78 (2020)

[49] Aljarah, I., Habib, M., Hijazi, N., Faris, H., Qaddoura, R., Hammo, B., Abushariah, M., Alfawareh, M.: Intelligent detection of hate speech in arabic social network: A machine learning approach. Journal of Information Science **47**(4), 483–501 (2021)

[50] Nugroho, K., Noersasongko, E., Fanani, A.Z., Basuki, R.S., *et al.*: Improving random forest method to detect hatespeech and offensive word. In: 2019 International Conference on Information and Communications Technology (ICOIACT), pp. 514–518 (2019). IEEE

[51] 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 **6**, 13825–13835 (2018)

[52] Burnap, P., Williams, M.L.: Us and them: identifying cyber hate on twitter across multiple protected characteristics. EPJ Data science **5**, 1–15 (2016)

[53] Abozinadah, E.A., Mbaziira, A.V., Jones, J.: Detection of abusive accounts with arabic tweets. Int. J. Knowl. Eng.-IACSIT **1**(2), 113–119 (2015)

[54] Magdy, W., Darwish, K., Weber, I.: # failedrevolutions: Using twitter to study the antecedents of isis support. arXiv preprint arXiv:1503.02401 (2015)

[55] Kaati, L., Omer, E., Prucha, N., Shrestha, A.: Detecting multipliers of jihadism on twitter. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 954–960 (2015). IEEE

[56] Alshehri, A., El Moatez Billah Nagoudi, H.A., Abdul-Mageed, M.: Think before your click: Data and models for adult content in arabic twitter. In: TA-COS 2018: 2nd Workshop on Text Analytics for Cybersecurity and Online Safety, vol. 15 (2018)

[57] Abozinadah, E.A.: Improved micro-blog classification for detecting abusive arabic twitter accounts. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol **6** (2016)

[58] Mubarak, H., Darwish, K., Magdy, W.: Abusive language detection on arabic social media. In: Proceedings of the First Workshop on Abusive Language Online, pp. 52–56 (2017)

[59] Jaki, S., De Smedt, T.: Right-wing german hate speech on twitter: Analysis and automatic detection. arXiv preprint arXiv:1910.07518 (2019)

[60] Alakrot, A., Murray, L., Nikolov, N.S.: Towards accurate detection of offensive language in online communication in arabic. Procedia computer science **142**, 315–320 (2018)

[61] Özel, S.A., Saraç, E., Akdemir, S., Aksu, H.: Detection of cyberbullying on social media messages in turkish. In: 2017 International Conference on Computer Science and Engineering (UBMK), pp. 366–370 (2017). IEEE

[62] Alfina, I., Mulia, R., Fanany, M.I., Ekanata, Y.: Hate speech detection in the indonesian language: A dataset and preliminary study. In: 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), pp. 233–238 (2017). IEEE

[63] Haidar, B., Chamoun, M., Serhrouchni, A.: A multilingual system for cyberbullying detection: Arabic content detection using machine learning. Advances in Science, Technology and Engineering Systems Journal **2**(6), 275–284 (2017)

[64] Abdelfatah, K.E., Terejanu, G., Alhelbawy, A.A., et al.: Unsupervised detection of violent content in arabic social media. Computer Science & Information Technology (CS & IT) **7** (2017)

[65] Fernandez, M., Alani, H.: Contextual semantics for radicalisation detection on twitter.(2018) (2018)

[66] Wiegand, M., Ruppenhofer, J., Schmidt, A., Greenberg, C.: Inducing a lexicon of abusive words–a feature-based approach. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1046–1056 (2018)

[67] Istaiteh, O., Al-Omoush, R., Tedmori, S.: Racist and sexist hate speech detection: Literature review. In: 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA), pp. 95–99 (2020). IEEE

[68] Kwok, I., Wang, Y.: Locate the hate: Detecting tweets against blacks. In: Twenty-seventh AAAI Conference on Artificial Intelligence (2013)

[69] 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, pp. 88–93 (2016)

[70] Frenda, S., Ghanem, B., Montes-y-Gómez, M., Rosso, P.: Online hate speech against women: Automatic identification of misogyny and sexism on twitter. Journal of Intelligent & Fuzzy Systems **36**(5), 4743–4752 (2019)

[71] Saha, P., Mathew, B., Goyal, P., Mukherjee, A.: Hateminers: Detecting hate speech against women. arXiv preprint arXiv:1812.06700 (2018)

[72] Fersini, E., Nozza, D., Rosso, P., *et al.*: Overview of the evalita 2018 task on automatic misogyny identification (ami). In: EVALITA Evaluation of NLP and Speech Tools for Italian Proceedings of the Final Workshop 12-13 December 2018, Naples (2018). Accademia University Press

[73] Andreas, J., Choi, E., Lazaridou, A.: Proceedings of the naacl student research workshop. In: Proceedings of the NAACL Student Research Workshop (2016)

[74] Yuliyanti, D., Sukoco: Sentiment mining of community development program evaluation based on social media (2017)

[75] Martin-Domingo, M., Mandsberg: Social media as a resource for sentiment analysis of airport service quality (2019)

[76] Mansour: Social media analysis of user's responses to terrorism using sentiment analysis and text mining (2018)

[77] Saragih, Girsang: Sentiment analysis of customer engagement on social media in transport online (2017)

[78] Hassan, H.S.L. Hussain: Sentiment analysis of social networking sites (sns) data using machine learning approach for the measurement of depression (2017)

[79] Joyce, Deng: Sentiment analysis of tweets for the 2016 us presidential election (2017)

[80] Ikoro, M. Harmina, BatistaNavarro: Analyzing sentiments expressed on twitter by uk energy company consumers (2018)

[81] Hao, Dai: Social media content and sentiment analysis on consumer security breaches (2016)

[82] Shayaa, C.S.J. Wai, Zakaria: Social media sentiment analysis on employment in malaysia (2017)

[83] Isah, T., Neagu: Social media analysis for product safety using text mining and sentiment analysis (2014)

[84] Ali, B.E. Dong, Hadjidj: Sentiment analysis as a service: A social media-based sentiment analysis framework (2017)

[85] Nemes, L., Kiss, A.: Social media sentiment analysis based on covid-19. Journal of Information and Telecommunication **5**(1), 1–15 (2021) https://doi.org/10.1080/24751839.2020.1790793. https://doi.org/10.1080/24751839.2020.1790793

[86] Idan, L., Feigenbaum, J.: Show me your friends, and i will tell you whom you vote for: Predicting voting behavior in social networks. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ASONAM '19, pp. 816–824. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3341161.3343676. https://doi.org/10.1145/3341161.3343676

[87] Tabinda Kokab, S., Asghar, S., Naz, S.: Transformer-based deep learning models for the sentiment analysis of social media data. Array **14**, 100157 (2022). https://doi.org/10.1016/j.array.2022.100157

[88] Alwakid, G., Osman, T., Haj, M.E., Alanazi, S., Humayun, M., Sama, N.U.: Muldasa: Multifactor lexical sentiment analysis of social-media content in nonstandard arabic social media. Applied Sciences **12**(8) (2022). https://doi.org/10.3390/app12083806

[89] Qian, C., Mathur, N., Zakaria, N.H., Arora, R., Gupta, V., Ali, M.: Understanding public opinions on social media for financial sentiment analysis using ai-based techniques. Information Processing Management **59**(6), 103098 (2022). https://doi.org/10.1016/j.ipm.2022.103098

[90] Arias, F., Zambrano Núñez, M., Guerra-Adames, A., Tejedor-Flores, N., Vargas-Lombardo, M.: Sentiment analysis of public social media as a tool for health-related topics. IEEE Access **10**, 74850–74872 (2022). https://doi.org/10.1109/ACCESS.2022.3187406

[91] Zhigang Jin, X.Z. Manyue Tao, Hu, Y.: Social media sentiment analysis based on dependency graph and co-occurrence graph (2022)

[92] Chandrasekaran, G., Antoanela, N., Andrei, G., Monica, C., Hemanth, J.: Visual sentiment analysis using deep learning models with social media data. Applied Sciences **12**(3) (2022). https://doi.org/10.3390/

app12031030

[93] Hassan, S.Z., Ahmad, K., Hicks, S., Halvorsen, P., Al-Fuqaha, A., Conci, N., Riegler, M.: Visual sentiment analysis from disaster images in social media. Sensors **22**(10) (2022). https://doi.org/10.3390/s22103628

[94] Sufi, F.K., Khalil, I.: Automated disaster monitoring from social media posts using ai-based location intelligence and sentiment analysis. IEEE Transactions on Computational Social Systems, 1–11 (2022). https://doi.org/10.1109/TCSS.2022.3157142

[95] Mohammad, S.M., Kiritchenko, S., Zhu, X.: Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. arXiv preprint arXiv:1308.6242 (2013)

[96] Read, J.: Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: Proceedings of the ACL Student Research Workshop, pp. 43–48 (2005)

[97] Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N project report, Stanford **1**(12), 2009 (2009)

[98] Davidov, D., Tsur, O., Rappoport, A.: Enhanced sentiment learning using twitter hashtags and smileys. In: Coling 2010: Posters, pp. 241–249 (2010)

[99] O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: Linking text sentiment to public opinion time series. In: Fourth International AAAI Conference on Weblogs and Social Media (2010)

[100] Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., Liu, B.: Combining lexicon-based and learning-based methods for twitter sentiment analysis. HP Laboratories, Technical Report HPL-2011 **89**, 1–8 (2011)

[101] Ding, X., Liu, B., Yu, P.S.: A holistic lexicon-based approach to opinion mining. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 231–240 (2008)

[102] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.J.: Sentiment analysis of twitter data. In: Proceedings of the Workshop on Language in Social Media (LSM 2011), pp. 30–38 (2011)

[103] Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J.: Twitter polarity classification with label propagation over lexical links and the follower graph. In: Proceedings of the First Workshop on Unsupervised Learning in NLP, pp. 53–63 (2011)

[104] Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp. 347–354 (2005)

[105] Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of twitter. In: The Semantic Web–ISWC 2012: 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I 11, pp. 508–524 (2012). Springer

[106] Shamma, D.A., Kennedy, L., Churchill, E.F.: Tweet the debates: understanding community annotation of uncollected sources. In: Proceedings of the First SIGMM Workshop on Social Media, pp. 3–10 (2009)

[107] Hu, X., Tang, L., Tang, J., Liu, H.: Exploiting social relations for sentiment analysis in microblogging. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 537–546 (2013)

[108] Mohammad, S., Turney, P.: Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 26–34 (2010)

[109] Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177 (2004)

[110] Nakov, P.: Semantic sentiment analysis of twitter data. arXiv preprint arXiv:1710.01492 (2017)

[111] Lin, J., Kolcz, A.: Large-scale machine learning at twitter. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp. 793–804 (2012)

[112] Rodríguez-Penagos, C., Atserias, J., Codina-Filba, J., García-Narbona, D., Grivolla, J., Lambert, P., Saurí, R.: Fbm: Combining lexicon-based ml and heuristics for social media polarities. In: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 483–489 (2013)

[113] Clark, S., Wicentwoski, R.: Swatcs: Combining simple classifiers with estimated accuracy. In: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 425–429 (2013)

[114] Hassan, A., Abbasi, A., Zeng, D.: Twitter sentiment analysis: A bootstrap ensemble framework. In: 2013 International Conference on Social Computing, pp. 357–364 (2013). IEEE

[115] Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)

[116] Baccianella, S., Esuli, A., Sebastiani, F., *et al.*: Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Lrec, vol. 10, pp. 2200–2204 (2010)

[117] Teh, P.L., Cheng, C.-B.: Profanity and hate speech detection. International Journal of Information and Management Sciences **31**(3), 227–246 (2020)

[118] Raktim Chatterjee, S.B., Kabi, S.: Profanity detection in social media text using a hybrid approach of nlp and machine learning (2021)

[119] Ratadiya, P., Mishra, D.: An attention ensemble based approach for multilabel profanity detection. In: 2019 International Conference on Data Mining Workshops (ICDMW), pp. 544–550 (2019). IEEE

[120] Park, J.H., Fung, P.: One-step and two-step classification for abusive language detection on twitter. arXiv preprint arXiv:1706.01206 (2017)

[121] Chinagundi, B., Singh, M., Ghosal, T., Rana, P.S., Kohli, G.S.: Classification of hate, offensive and profane content from tweets using an ensemble of deep contextualized and domain specific representations (2021)

[122] Soykan, L., Karsak, C., Elkahlout, I.D., Aytan, B.: A comparison of machine learning techniques for turkish profanity detection. In: Proceedings of the Second International Workshop on Resources and Techniques for User Information in Abusive Language Analysis, pp. 16–24 (2022)

[123] Dadvar, M., Eckert, K.: Cyberbullying detection in social networks using deep learning based models; a reproducibility study. arXiv preprint arXiv:1812.08046 (2018)

[124] Ba Wazir, A.S., Karim, H.A., Abdullah, M.H.L., AlDahoul, N., Mansor, S., Fauzi, M.F.A., See, J., Naim, A.S.: Design and implementation of fast spoken foul language recognition with different end-to-end deep neural network architectures. Sensors **21**(3), 710 (2021)

[125] Kim, C.-G., Hwang, Y.-J., Kamyod, C.: A study of profanity effect in sentiment analysis on natural language processing using ann. Journal of Web Engineering, 751–766 (2022)

[126] Kumari, K., Singh, J.P.: Ai ml nit patna at hasoc 2019: Deep learning approach for identification of abusive content. FIRE (working notes) **2517**, 328–335 (2019)

[127] Gottipati, S., Tan, A., Chow, D., Shan, J., Lim, J., Kiat, W.: Leveraging profanity for insincere content detection-a neural network approach. In: 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 0041–0047 (2020). IEEE

[128] Yang, H., Lin, C.-J.: Tocp: A dataset for chinese profanity processing. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pp. 6–12 (2020)

[129] Woo, J., Park, S.H., Kim, H.K.: Profane or not: Improving korean profane detection using deep learning. KSII Transactions on Internet and Information Systems (TIIS) **16**(1), 305–318 (2022)

[130] Sazzed, S.: Bengsentilex and bengswearlex: creating lexicons for sentiment analysis and profanity detection in low-resource bengali language. PeerJ Computer Science **7**, 681 (2021)

[131] Al-Hashedi, M., Soon, L.-K., Goh, H.-N.: Cyberbullying detection using deep learning and word embeddings: An empirical study. In: Proceedings of the 2019 2nd International Conference on Computational Intelligence and Intelligent Systems, pp. 17–21 (2019)

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

[133] Bhowmick, R.S., Ganguli, I., Paul, J., Sil, J.: A multimodal deep framework for derogatory social media post identification of a recognized person. Transactions on Asian and Low-Resource Language Information Processing **21**(1), 1–19 (2021)

[134] Malik, P., Aggrawal, A., Vishwakarma, D.K.: Toxic speech detection using traditional machine learning models and bert and fasttext embedding with deep neural networks. In: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 1254–1259 (2021). IEEE

[135] Marwa, T., Salima, O., Souham, M.: Deep learning for online harassment detection in tweets. In: 2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS), pp. 1–5 (2018). IEEE

[136] Perera, A., Fernando, P.: Accurate cyberbullying detection and prevention on social media. Procedia Computer Science **181**, 605–611

(2021)

[137] Chaudhari, A., Davda, P., Dand, M., Dholay, S.: Profanity detection and removal in videos using machine learning. In: 2021 6th International Conference on Inventive Computation Technologies (ICICT), pp. 572–576 (2021). IEEE

[138] Sood, S., Antin, J., Churchill, E.: Profanity use in online communities. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1481–1490 (2012)

[139] Sood, S.O., Antin, J., Churchill, E.: Using crowdsourcing to improve profanity detection. In: 2012 AAAI Spring Symposium Series (2012)

[140] Chin, H., Kim, J., Kim, Y., Shin, J., Yi, M.Y.: Explicit content detection in music lyrics using machine learning. In: 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 517–521 (2018). IEEE

[141] Sodhi, R., Pant, K., Mamidi, R.: Jibes & delights: A dataset of targeted insults and compliments to tackle online abuse. In: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), pp. 132–139 (2021)

[142] Kurniawan, S., Budi, I.: Indonesian tweets hate speech target classification using machine learning. In: 2020 Fifth International Conference on Informatics and Computing (ICIC), pp. 1–5 (2020). IEEE

[143] Yasaswini, K., Puranik, K., Hande, A., Priyadharshini, R., Thavareesan, S., Chakravarthi, B.R.: Iiitt@ dravidianlangtech-eacl2021: Transfer learning for offensive language detection in dravidian languages. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pp. 187–194 (2021)

[144] Alkomah, F., Ma, X.: A literature review of textual hate speech detection methods and datasets. Information **13**(6), 273 (2022)

[145] Fahim, M., Gokhale, S.S.: Detecting offensive content on twitter during proud boys riots. In: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1582–1587 (2021). IEEE

[146] Raj, C., Agarwal, A., Bharathy, G., Narayan, B., Prasad, M.: Cyberbullying detection: hybrid models based on machine learning and natural language processing techniques. Electronics **10**(22), 2810 (2021)

[147] Reynolds, K., Kontostathis, A., Edwards, L.: Using machine learning to

detect cyberbullying. In: 2011 10th International Conference on Machine Learning and Applications and Workshops, vol. 2, pp. 241–244 (2011). IEEE

[148] Viñas Redondo, B.: Identifying and categorizing offensive language in tweets using machine learning. B.S. thesis, Universitat Politècnica de Catalunya (2020)

[149] Cotik, V., Debandi, N., Luque, F.M., Miguel, P., Moro, A., Pérez, J.M., Serrati, P., Zajac, J., Zayat, D.: A study of hate speech in social media during the covid-19 outbreak (2020)

[150] Alotaibi, M., Alotaibi, B., Razaque, A.: A multichannel deep learning framework for cyberbullying detection on social media. Electronics **10**(21), 2664 (2021)

[151] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Predicting the type and target of offensive posts in social media. arXiv preprint arXiv:1902.09666 (2019)

[152] Sharifirad, S.: Nlp and machine learning techniques to detect online harassment on social networking platforms (2019)

[153] Shanmugavadivel, K., Sathishkumar, V., Raja, S., Lingaiah, T.B., Neelakandan, S., Subramanian, M.: Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data. Scientific Reports **12**(1), 21557 (2022)

[154] Omar, A., Mahmoud, T.M., Abd-El-Hafeez, T., Mahfouz, A.: Multi-label arabic text classification in online social networks. Information Systems **100**, 101785 (2021)

[155] Kalaivani, A., Thenmozhi, D.: Ssn_nlp_mlrg at semeval-2020 task 12: Offensive language identification in english, danish, greek using bert and machine learning approach. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 2161–2170 (2020)

[156] Sanoussi, M.S.A., Xiaohua, C., Agordzo, G.K., Guindo, M.L., Al Omari, A.M., Issa, B.M.: Detection of hate speech texts using machine learning algorithm. In: 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0266–0273 (2022). IEEE

[157] Thenmozhi, D., Sharavanan, S., Chandrabose, A., *et al.*: Ssn_nlp at semeval-2019 task 6: Offensive language identification in social media using traditional and deep machine learning approaches. In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 739–744 (2019)

[158] Nayel, H.A., Shashirekha, H.: Deep at hasoc2019: A machine learning framework for hate speech and offensive language detection. In: FIRE (Working Notes), pp. 336–343 (2019)

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

[160] Parikh, A., Desai, H., Bisht, A.S.: Da master at hasoc 2019: Identification of hate speech using machine learning and deep learning approaches for social media post. In: FIRE (Working Notes), pp. 315–319 (2019)

[161] Thara, S., Poornachandran, P.: Social media text analytics of malayalam–english code-mixed using deep learning. Journal of big Data **9**(1), 45 (2022)

[162] Sutejo, T.L., Lestari, D.P.: Indonesia hate speech detection using deep learning. In: 2018 International Conference on Asian Language Processing (IALP), pp. 39–43 (2018). IEEE

[163] Bharathi, B., *et al.*: Ssncse_nlp@ dravidianlangtech-eacl2021: Offensive language identification on multilingual code mixing text. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pp. 313–318 (2021)

[164] Zhang, Z., Robinson, D., Tepper, J.: Detecting hate speech on twitter using a convolution-gru based deep neural network. In: The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15, pp. 745–760 (2018). Springer