Multimedia Tools and Applications (2024) 83:75243–75292
https://doi.org/10.1007/s11042-024-18452-0

A survey and comparative study on negative sentiment analysis in social media data

Jayanta Paul[1] · **Ahel Das Chatterjee**[1] · **Devtanu Misra**[1] · **Sounak Majumder**[1] · **Sayak Rana**[1] · **Malay Gain**[1] · **Anish De**[1] · **Siddhartha Mallick**[1] · **Jaya Sil**[1]

Received: 12 May 2023 / Revised: 28 December 2023 / Accepted: 26 January 2024 /
Published online: 15 February 2024

Abstract
With the rapid growth of internet usage, especially on social media, forums, review platforms, and blogs, an enormous amount of data is being generated. This data often contains users' opinions, emotions, and arguments on various topics. To make informed decisions or predictions, it's crucial to analyze and organize this unstructured data effectively. Sentiment analysis of social media data has become essential, aiming to identify different forms of sentiments like hate speech, profanity, sentiment, and targeted insults. However, in the field of natural language processing (NLP), a significant challenge in sentiment analysis is the scarcity of labeled data. Researchers have traditionally used methods like lexicon-based and traditional machine learning approaches to process this unstructured social media data. Recent studies indicate that deep learning techniques have proven effective in handling this task. This study aims to provide a comprehensive overview of various classical machine learning and deep learning techniques employed in sentiment analysis. We explore different sentiment analysis categories and compare their performance using various evaluation metrics.

Keywords Sentiment analysis · Review on different sentiment types · Hate speech · Profanity · Targeted insults · Natural language processing · Lexicon based methods · Machine learning · Deep learning

1 Introduction

Sentiment Analysis (SA) in the context of social media involves gauging public opinions about entities and classifying them as positive, negative, or neutral. Over time, expressions of emotion on platforms like Facebook, Instagram, Twitter, LinkedIn, Amazon, and Flipkart have evolved, manifesting in more distinct and nuanced forms, such as hate speech, profanity,

✉ Jayanta Paul
2020csp003.jayanta@students.iiests.ac.in

✉ Siddhartha Mallick
msiddhartha1600@gmail.com

Extended author information available on the last page of the article

negative opinions, and targeted insults. Such expressions of opinions are defined as negative sentiment. This landscape has catalyzed extensive research on analyzing sentiments from such diverse data. This review paper seeks to spotlight key facets of these research endeavors, elucidating their contributions, methodologies, and algorithms. Sentiment analysis is indispensable in areas like social media monitoring, product reviews, market research, competitor analysis, and customer support for various businesses. The process typically begins with data collection in its raw form, which then undergoes preprocessing to ensure standardization. Subsequently, pertinent features are extracted, paving the way for sentiment classification through methodologies ranging from lexicon-based approaches to classical machine learning and deep learning techniques. Classical machine learning techniques, including Naive Bayes (NB), Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), Maximum likelihood, k-Nearest Neighbors (k-NN), and Conditional Random Field (CRF) have been pivotal in sentiment extraction from text. On the other hand, deep learning, a subset of machine learning, employs architectures such as Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Long short-term memory (LSTM), Gated Recurrent Unit (GRU), and Bidirectional Encoder Representations from Transformers (BERT) for sentiment classification. The dynamics of sentiment expression on social media is not merely a computational challenge; it is an intricate mosaic of cultural, sociological, and individual factors. The platforms today are utilized as repositories of human thought, influenced by global events, personal experiences, and the evolution of internet culture. For this reason, a mere algorithmic analysis is insufficient. In this survey, our approach extends beyond the methods and models. We probe deeper into the findings of various research works, seeking patterns, anomalies, and insights. Our aim is to discern not just what works and what doesn't, but why certain methods excel in one context and falter in another. We draw attention to the nuances of sentiment expression, the contextual dependencies, and the ever-evolving nature of online discourse. It's worth noting that while many papers provide results, few delve into the deeper reasons behind those results. We believe that this depth of observation and reasoning will bridge the gap, offering readers a fuller understanding of the complexities involved in social media sentiment analysis.

It's our contention that by comprehending these intricacies, the research community can better tailor solutions, anticipate challenges, and usher in the next wave of innovation in sentiment analysis. This review paper makes several contributions to the field of sentiment analysis:

- Comprehensive Overview: We conduct exhaustive experiments using classical machine learning and deep learning techniques to analyze negative sentiments of social media data and other on line domain as well. The comprehensive approach provides valuable insights of the methodologies in analyzing negative sentiments.
- Comparative Analysis: Based on the experimental results, the paper provides a thorough comparative analysis of these techniques for assessing their performance using various metrics such as accuracy (A), precision (P), recall (R), F-measure (F1), and ROC.
- Systematic Categorization: Sentiment manifestations, including hate speech, profanity, negative opinions, and targeted insults, are systematically categorized and analyzed, providing a nuanced understanding of negative sentiment expressions in diverse contexts.
- Methodological Strengths and Limitations: Our paper discusses the strengths of each methodology and identifies the limitations as the inherent weaknesses of the algorithms. This nuanced approach facilitates the readers of the paper to understand applicability of the methods in analyzing negative sentiments.

The paper is structured as follows: Section 2 delves into hate speech detection techniques underpinned by classical machine learning and deep learning. Section 3 explores negative opinion analysis using both classical and deep learning methodologies. Section 4 discusses profanity classification techniques, while Section 5 sheds light on targeted insults in the social media realm. A broader discussion on the scope and implications of this study is offered in Section 6. Finally, Section 7 concludes the paper.

2 Hate speech(HS) detection

Identifying hate speech is a complex task, which poses challenges even for human beings. It comprises of language that is offensive and aims to target specific individuals or groups based on inherent characteristics, such as religion, gender, or race, leading to a potential threat to social harmony. Such speech is often linked with prejudiced attitudes such as racism, violence, misogyny, and Islamophobia. We have conducted research on conventional machine learning algorithms and advanced deep learning algorithms for the detection of hate speech. The problem of identifying hate speech falls under the category of text classification, and various classifiers are available for this task. Nonetheless, the key challenge is to select the most suitable classifier, which requires a comprehensive understanding of each hate speech classifier that currently exists. Machine learning approaches are broadly categorized into classical methods, ensemble of models, and deep learning (DL) based methods. In a study conducted by Badjatiya et al. [1], different machine learning models including Logistic Regression, Support Vector Machine, and Gradient Boosting Decision Tree, were evaluated for their ability to detect hate speech. The findings of this study showed that deep learning models, such as Long Short-Term Memory (LSTM) or Convolutional Neural Networks (CNN), outperformed classical machine learning algorithms by 13-20%.

2.1 Hate speech detection using deep learning method

Based on the review of 178 research papers, it was found that BERT [2], LSTM [3], and CNN[4] were the most commonly used and best-performing algorithms in various natural language processing (NLP) tasks. The architectures of these models generally followed a three-step process: Word Embedding Layer: In this step, pre-trained word embedding models such as GloVe[5], TF-IDF Vectorizer, Word2Vec [6], etc., were used to represent words as continuous vectors. These word embeddings captured the semantic meaning of words and helped in representing text data in a numerical format that can be used as input to deep learning models. Feature Extraction: Deep learning models such as BERT, LSTM, and CNN were used for feature extraction. These models were trained on large amounts of data and were capable of capturing complex patterns and representations from the input text. BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model that captures contextual word representations, LSTM (Long Short-Term Memory) is a recurrent neural network (RNN) architecture capable of modeling sequential data, and CNN (Convolutional Neural Network) is a type of neural network architecture that uses convolutional layers to capture local patterns. Fully Connected Layer for Classification: After feature extraction, the output was passed through fully connected layers for classification. Fully connected layers are used to learn non-linear relationships between features and generate an output for the specific NLP task, such as sentiment classification, named entity recognition, or text classification. Overall, this three-step architecture involving word embedding, feature extraction using deep learning models, and fully connected layers for classification has been widely used in various research papers and has shown good performance in NLP tasks using BERT, LSTM, and CNN algorithms. Deep learning models can vary depending on various factors, including the nature of the data, the specific task, and the experimental setup. Different authors may report different findings based on their experimental setups and data analysis. For example, Jahan, M.S. [7] found that CNN performed better than LSTM, while Badjatiya et al. [1] found that LSTM performed better than CNN. This discrepancy in findings could be due to differences in the datasets, model configurations, hyperparameters, and evaluation metrics used in these studies. Moreover, many researchers have found that combining multiple deep learning models can lead to improved performance compared to using a single model. For instance, studies by Zhou et al. [8], as well as references [9–12] have shown that combining different models, such as CNN+LSTM, CNN+GRU, or multiple CNN models with different parameters, can result in better performance compared to using individual models separately. BERT, or Bidirectional Encoder Representations from Transformers, has indeed emerged as a popular model for hate speech classification in the past 5 years. Several works [13–16] have explored BERT's performance in hate speech detection and concluded that BERT is a superior model in this task. One notable work in this area is reported by Velankar et al. [17], where they presented a dataset called L3Cube-MahaHate [17], extracted from Twitter and manually labeled into four classes: hate, offensive, profane, and not. They experimented with monolingual and multilingual variants of BERT, such as MahaBERT [18], IndicBERT [19], mBERT[20], and xlm-RoBERTa[21], and found that monolingual models performed better than their multilingual counterparts [16, 22–25]. For example, the accuracy of xlm-RoBERTa was reported as 0.894, while MahaBERT, a model trained on Marathi monolingual datasets, achieved an accuracy of 0.909 [18].

In addition to the general BERT model, language-specific BERT models have also been developed for monolingual tasks and have shown superior performance compared to the multilingual model mBERT in certain cases. For example, AraBERT for Arabic [26], RuBERT for Russian [27], AlBERTo for Italian [22], BERTje for Dutch [28], FinBERT for Finnish [29], CamemBERT for French [30], Flaubert for French [31]), BERT-CRF for Portuguese [32], BERTje for Dutch [28], and BERTtweet for English Tweets [33] are some of the language-specific BERT models that have been developed over time and have demonstrated improved performance in their respective languages.

The research on hate speech detection in mixed languages, specifically Hindi-English code mix, is limited but has been gaining attention in recent years. There are several studies and datasets available in this domain. One such dataset is available in [34], which contains Hindi-English code mix text written in Roman script. The authors proposed text classification using this Hinglish text and applied deep learning (DL) methods [35]. They experimented with CNN-based DL models using domain-specific embeddings and reported accuracy of 82.62%, precision of 83.34%, and F1-score of 80.85% on a benchmark dataset [36]. Another study by Mathur et al. [37] created a self-made Hindi-English code mix dataset with annotations and applied machine learning (ML) models, including a baseline model. They proposed a Multi-Channel Transfer Learning based model (MIMCT) and observed that the proposed model outperformed state-of-the-art methods. In another paper, a novel tweet dataset titled "Hindi English Offensive Tweet (HEOT)" was introduced by Mathur et al. [38]. The tweets were manually annotated into three categories: non-offensive, abusive, and hate speech. They used a CNN model and reported an accuracy of 83.90%, precision of 80.20%, recall of 69.98%, and F1-score of 71.45%.

In the field of hate speech detection, racism and sexism are major subsets that have been studied as well. In one study [39], the authors used pre-trained word embeddings [40] and applied max/mean pooling to extract features using these embeddings. Next, This feature is fed to a neural network with 2 layers, followed by ReLU activation functions and a soft-max output layer. This method achieved an impressive F1 score of 0.9241. In another study [41], the researchers proposed a classification model that utilized pre-trained word2vec features applied to a multi-layer CNN. They achieved an F1 score of 78.30% with their approach.

One of the major aspects of this study is to provide the examples of real life applications considering the proposed methods. Quoc et. al. [42] made a real time hate speech detection system using Spark Streaming for their country Vietnam. The paper [43] suggested a generizable model, named PEACE, which takes inherent casual cues to detect hate speeches and it performed better in this task across five different social media platforms and two different targets. The papers [9, 44, 45] have done interesting research on hate speech detection or classification (e.g.- [1]) in Arabic and the later one [45] is first time ever, to apply psychology theories to build a computational hate speech detection system. The class imbalance problem in online platforms and hate speech detection in short text format [46] paved one new avenue of research. Some papers have done hate speech detection from social media in multiple language. The works of Chopra et.al. [47], Kamble et. al. [36] focused on mixed code words (in Hindi and English) whereas Ranasinghe et. al. in German, English and Hindi and Bilal et.al. in Roman Urdu. Gupta et. al. [48] experimented with 12 model architectures to train their model on Character Level Embedding for multiclass labeling of Hinglish tweets into three categories(non-offensive, abusive, hate-inducing tweets). The paper [1] shows how to utilize deep neural network architectures for hate speech detection on the platforms, like Twitter and can drastically enhance the efficiency of content moderation, promoting healthier online interactions. As platforms grow, integrating user network features can further refine the accuracy of these detection systems, ensuring safer online communities. Implementing automated hate speech detection tools for Urdu on social media as suggested in [49], can significantly improve content moderation, fostering more positive online interactions in the Urdu-speaking community. By combining machine learning and transfer learning techniques, these tools can effectively identify and mitigate hateful comments, enhancing the overall user experience and minimizing the spread of offensive content. On the other hand due to limited training data, Kovács et al. [50] explored unlabeled data with labeled corpora for better analysis of hate speech with limited hateful words available in social media. In [51], HateNet and t-HateNet present valuable tools for social media platforms to automatically detect and mitigate hateful content. By automating the process of identifying and categorizing hate speech, these technologies protecting users from harmful online interactions and reduce the emotional toll on human moderators. In another study, Nagar et. al. [52] implemented the problem in a real-world setting, where social media platforms can integrate the proposed model into their content moderation systems. By continuously updating the model with fresh data and feedback, the system can learn and adapt new forms of hate speech. Additionally, users can be given an option to report false positives or negative cases, enhancing the model's accuracy and efficiency over time. Collaboration with organizations focusing on digital safety, ethics and also help to refine and optimize the model for diverse online communities. In another paper, Saleha et. al. [1], implemented this model in real-life applications using social media platforms and online communities can integrate it into their content moderation systems. Continuous monitoring and user feedback can be used to improve the model's performance and adapt to evolving hate speech trends. Tables 1, 2 and 3 provide summaries of articles related to hate speech detection using deep learning, showcasing the methods, techniques, and results achieved in these studies.

Table 1 Article summary of hate speech detection using deep learning

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Quoc et. al. [42], 2023	Wikipedia, 223k	BERT base features Representation	PhoBERT, Text-CNN	Proposed using combined PhoBERT and Text-CNN for hate speech detection.	Need high censoring Need large social network.	0.62	0.47	-	-	-
Mazari et. al. [43], 2023	ViHSD, HSD-VLSP	GloVe FastText	LSTM, BiLSTM, BERT Bi-GRU	Methodology combines pre-trained BERT with DL models for ensemble architectures.	a higher error rate in the offensive hate type than the violence type.	0.73	-	-	0.98	-
Elzayady et. al.[45], 2023	Twitter (Egyptian)	TF-IDF	LSTM, BiLSTM, AraBERT Joint CNN and RNN models	Strategy focuses on using personality traits to detect Arabic hate speech.	Does not tell about multi-personality trait features.	-	-	-	0.82	-
Al-Hassan et. al. [9], 2021	Twitter, 11k (Arabic)	Keras word embeding	LSTM, GURU, CNN + GRU,CNN + LSTM	The objective of this study is to categorize Arabic tweets into five distinct classifications, namely: none, religious, racial, sexism, or general hate	Not applied to real time stream of tweets.	0.75	0.72	0.75	0.73	0.74
Rizos et. al. [46], 2019	Twitter, 24k	FastText, Word2Vec, GloVe	CNN, LSTM, GRU	Provide three methods for augmenting text-based data that are designed to address the issue of class imbalance.	Limited by data quality, Requires training one generative model per class.	-	-	49.5	0.74	-
Kamble et. al.[36], 2018	Twitter, 3.8k	Word2Vec	LSTM, BiLSTM, CNN	Study improves state-of-the-art in code-mixed English-Hindi hate speech detection.	Code-switched tweets in Hindi, Series of swear words, Possibly incorrect labels.	-	0.83	0.78	0.80	0.80

Table 1 continued

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Ranasinghe et al. [14], 2019	Facebook Twitter HASOC 2019	FastText	LSTM, GRU, BERT	Study aims to detect hate speech in multilingual social media via deep learning.	Limited to only three languages(German, English, Hindi).	-	-	0.75	0.78	-
Faris et al.[44], 2020	Twitter, (Arabic)	Word2Vec	Aravec CNN + LSTM	Paper analyzes Twitter hate speech detection using the NLTK library dataset.	Not applied for large benchmark datasets.	0.66	0.69	0.79	0.71	0.70

Table 2 Continued

Author, Year	Context/Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Badjatiya et. al. [1], 2021	Twitter, 16k	TF-IDF, Bag of Words, GloVe	LSTM+Random Embedding+GBDT, LSTM+GLoVe, CNN+GloVe+GBDT	The focus of this study is to explore the use of deep neural network structures in hate speech detection.	The significance of user network features is not considered here.	-	0.93	0.93	0.93	-
Duwairi et al. [53], 2021	Twitter, (ArHS dataset) 9k, 2k (Arabic)	SG, MUSE, CBOW	CNN, CNN + LSTM, BiLSTM + CNN	This article addresses the challenge of detecting hate speech in Arabic by undertaking three specific tasks.	Does not tell about sources and targets of hate speech.	-	0.74	-	-	-
Ali et al. [49], 2022	Twitter (10k) (Urdu)	FastText, FastText + BiGRU	BERT, DistilBERT, XLM-Roberta	This study aims to create a lexicon for identifying hateful language in Urdu, which has been developed by the authors.	Need for more research on automated hate lexicons.	0.73	0.76	0.65	0.67	0.68
Kovács et al. [50], 2021	Twitter(HASOC 2019) (Hindi-English)	FastText, GloVE	RoBERTa, CNN-BiLSTM , DistillBERT	Proposed an NLP model combining convolutional and recurrent layers for automatic hate speech detection in social media.	Limited to only one dataset due to computational limitation.	-	-	-	0.85	-
Chopra et. al. [47], 2020	Twitter (Hindi-English)	Keras Tokenizer	FT + CNN+ BiLSTM + Attn + PV + DW + Debias	This study introduces a three-step method to detect hate speech in Hinglish on platforms like Twitter, employing profanity modeling, deep graph embeddings, and author profiling.	Confidentiality, Prejudice, Potential Misrepresentation, Obstacles to Deployment	0.78	-	-	0.73	-
Gupta et al. [48], 2021	Twitter, 3k (Hindi English)	Character-Level Embeddings Generation	GRU + Attention, CNN + GRU, Bi-LSTM + GRU	The author used 12 model architectures to categorize Hinglish tweets into non-offensive, abusive, and hate-inducing classes, leveraging character-level embedding for training. .	resolving the shortcomings of the character level approach	0.87	0.87	0.87	0.87	0.86

Table 2 continued

Author, Year	Context/Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Bilal et al. [54], 2023	Twitter,Facebook 173k (Roman Urdu)	BERT based Embeddings Generation	BERT-RU, LSTM, BiLSTM, 123BiLSTM + Attention Layer, and CNN	This study focuses on employing a transformer-based model to classify hate speech in Roman Urdu, and introduces the first pre-trained BERT model for Roman Urdu, named BERT-RU. .	Varities of word embedding approaches(like-FastText, Glove etc.) were not tested	0.96	0.97	0.96	0.97	0.97

Table 3 Continued

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Yuan et. al. [51], 2023	Twitter, 22k	GloVe	BiLSTM, CNN	Introduced a method combining multiple datasets to reduce bias in sentiment analysis.	Reliance on bias-rich, human annotated data affects the model's objectivity in hate speech detection.	-	-	-	-	-
Jahan et. al. [55], 2023	social media platforms (Facebook, Youtube, Twitter)	FastText, GloVe, Word2Vec	LSTM, BiLSTM RoBERTa, DistilBERT, ALBERT, RNN	provides a systematic review of literature in this field, with a focus on natural language processing and deep learning technologies.	Limited coverage of newer methodologies and absence of hands-on case studies.	-	-	-	-	-
Nagar et. al. [52], 2023	Twitter, (Founta,Ribeiro)	BERT embeddin	LSTM, Variational Graph Auto-Encoder	Present a novel approach to detecting hate speech on Twitter	The solution may not account for evolving linguistic nuances and sarcasm in hate speech.	0.84	0.85	0.84	0.84	0.90
Saleha et. al. [56], 2023	Davidson-ICWSM, Waseem-EMNLP, Waseem-NAACL	Word2vec,Glove, HSW2V	BiLSTM, BERT Base, BERT Large	Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model.	BERT's limitations in recognizing domain-specific hate terms, abbreviations, and intentional misspellings.	0.96	0.96	0.96	0.96	0.96

Table 3 continued

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Islamia et. al. [57], 2023	Twitter (IEHate)	Word2vec,								
FastText, Doc2vec	Bi-LSTM + CNN XLM-Roberta, RoBERTa	Introduce a new dataset named IEHate and Uncovering Political Hate Speech During Indian Election Campaign	The dataset focuses only on Hindi language tweets, limiting linguistic diversity.	0.92	-	0.72	-	-		

2.2 Hate speech detection using classical method

The Classical machine learning methods for hate speech detection are applied to train a dataset that is labeled either manually or automatically. Support Vector Machines (SVM)[58], Naive Bayes (NB)[59], Logistic Regression (LR)[60], Decision Trees (DT)[61], K-Nearest Neighbor (KNN)[62], Random Forest (RF)[63], and others[64] are used to build the model that can detect and classify text as hate speech or non-hate speech.

Tables 4 and 5 summarizes the research papers that describe classical machine learning approaches used for hate speech detection and showcasing different algorithms. It's worth to note that the performance of these algorithms may vary depending on the specific dataset and features used for training, as well as the evaluation metrics employed for performance analysis. In addition to hate speech in general, another important type of hate speech includes comments about race and sex, which are racism and sexism, respectively. In the paper [65], the authors explore works that attempted to classify racism and sexism in text. In one approach [66], researchers used a simple method by applying word uni-gram features to a Naive Bayes classifier. However, this model resulted in a large number of false-positive predictions because the uni-gram features did not consider the relation between the words, leading to tweets with certain keywords being classified as racist irrespective of the context. In another study [67], Waseem and Hovy compared various combinations of features and applied them to a logistic regression classifier. They used a combination of uni-gram, bi-gram, tri-gram, and quad-gram models, along with the gender of the tweet writer, which resulted in the highest F1 score of 73.93%. In a different approach [68], researchers used a combination of 1-3 word n-grams, 1-7 char n-grams, and sexism-related lexicons to detect sexism in text. This combination of features helped in identifying sexist content more accurately.

SVM classifier was used to classify tweets using the features mentioned above, related to race and sex. For example, in one study [69], a combination of pre-processed 512-word embedding, TF-IDF, and 300-dimensional Bag of Word Vector (BoWV) as features were used with a logistic regression classifier, achieving an F1 score of 70.04%. In another study [68], researchers worked with three datasets, as reported in [70] and [71]. On dataset [70], char n-grams were used as features and SVM was used as the classifier, achieving an accuracy of 78.77% and 75.44%, respectively. On dataset [71], bag of words and sequences of words features with SVM as the classifier provided the best accuracy of 89.32%. A few more practical work, like the work of B.Vidgen et. al. [72] describes the multi-class Islamophobic hate speech classifier to provide minute insight into online Islamophobia. The paper [73], applying the proposed machine learning model to popular Arabic social media platforms can actively detect and mitigate hate speech, fostering safer online communities. By continuously refining the model to understand colloquial and dialectic nuances, platforms can respond promptly to emerging hate speech patterns and deter users from engaging in harmful behavior. K. Nugroho et. al. [74] reported that by implementing the Random Forest method on major online platforms, a more robust system is developed for detecting and mitigating hate speech and offensive content. Its superior performance over AdaBoost and Neural Network can aid in establishing a safer online environment for users. The paper [75] contributes towards the nascent study of intersectionality in hate crime and the paper [76] has shown a way to automatically detect hate speech patterns and classify them in binary (whether a tweet is offensive or not) and ternary (whether a tweet is hateful, offensive, or clean). The works of Megdy et. al. [77] and Kaati et. al. [78] can be used to study the activities, concerned with the security of nations. Within the scope of detecting Arabic hate speeches and cyberbullying, a few papers ([79–83]) exhibited some classical methods, may lead to real time hate speech detection in social media. By detecting German hate speech, the work of Jaki et. al. [84]

Table 4 Article summary of hate speech detection using classical method

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
B.Vidgen et.al. [72], 2020	Twitter, 73K	Word Embedding	NB, RF, LG, DT, SVM, DL	Improvement on islamophobia detection	Not tested against increasing the size of the training dataset.	-	0.77	0.74	0.73	-
Aljarah et.al [73], 2021	Twitter	TF-IDF ,Bag of Word(BoG)	SVM,NB, DT,RF	Addresses Code-switch	Subjectivity, dialect variations, lack of datasets, and misspellings challenge.	0.90	0.90	0.95	0.90	0.90
K. Nugroho et. al. [74], 2019	Twitter, 14K	Count Vectors	RF	Improved RF for HS detection	Neural Network and AdaBoost underperform compared to Random Forest.	-	0.71	0.72	0.71	-
R. Martins et. al. [87], 2018	Twitter, 24K	N-gram	SVM,NB,RF	Hate speech classification in social media using emotional analysis	Does not tell the issue of users characterisation to overcome anti-hate speech policies.	-	0.76	0.73	-	-
H. Watanabe et. al [76], 2018	Twitter	Unigram	SVM, J48graft	Combination 3 different datasets which give a wider coverage Unigrams	Limited to mainly unigram dictionary of uni-gram hate speech pattern.	-	0.79	0.78	0.78	-
P. Burnap et. al [75], 2016	Twitter	Bag of words, N-gram	SVM, RBF	Identifying cyber Hate	Does not cover intersectional dimensions(like-religion intersects with sexual orientation etc.)	-	0.96	0.97	0.96	-

Table 4 continued

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Abozinadah et al. [88], 2015	Twitter	Bag of words, N-gram	NB,SVM,DT	Detecting abusive accounts with Arabic tweets by using text classification.	Only limited to arabic tweets.	0.86	0.87	0.86	0.86	0.86
Magdy et al. [77], 2015	Twitter	Temporal patterns, Hashtags	SVM	Terrorism (Pro-ISIS and Anti-ISIS)	Data privacy	-	0.87	0.87	0.87	-
Kaati et al. [78], 2015	Twitter	Data dependent features and data independent features	AdaBoost	Terrorism (Support or Oppose Jihadism)	Ageing factor (AF) is not included.	0.99	0.98	0.98	-	-
Abozinadah et. al. [79], 2017	Teitter	Page Rank (PR) algorithm, Semantic Orientation (SO)algorithm, statistical measures	SVM	Abusive Hate Speech detection	Dictionary, used here, has limitations on dealing with slang and dialect words.	0.96	0.96	0.96	0.96	0.96
Mubarak et al. [80], 2017	Twitter	Unigram and bigram, Log Odds Ratio (LOR)	Compute Log Odds Ratio (LOR)	Abusive, Ofensive	Limited to only Arabic dataset.	-	0.98	0.45	0.60	-

Table 5 Continued

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Jaki et. al. [84], 2018	Twitter	Skip grams and Character tri-grams	K-means, single-layer averaged perceptron	Radicalization (Muslim, Terrorist, Islamo fascistoid)	Multimodal tweets are not considered as dataset.	-	0.84	‘0.83	0.84	0.83
Alakrot et al. [81], 2018	YouTube	N-gram	SVM	Offensive, In-offensive	Combination between stemming and N-gram features has negative effect on precision and recall.	-	0.88	0.80	0.82	-
Ozel et al. [85], 2017	Instagram and Twitter	Bag of words	SVM,DT,NB,kNN	Develop a dataset to detect cyberbullying on social media messages written in Turkish.	The dataset used, is not a large one.	-	-	-	0.84	-
Alfina et al. [59], 2017	Twitter	Bag of words, N-gram	Random Forest	Built a new dataset of tweets in the Indonesian language for hate speech detection	The BOW model is inadequate to detect hate speech.	-	-	-	0.93	-
Haidar et al. [83], 2017	Twitter	Feature Vector	SVM, NaÏve Bayes	Detecting Cyber- bullying	Performance is not tested against deep-learning methods.	-	0.93	0.94	0.92	0.93
Abdelfatah et al. [82], 2017	Twitter	Morphological features Vector Space Model	K-means clustering	Introduced an unsupervised framework for detecting violence in Arabic Twitter.	Compared against less number of experiments.	-	0.58	0.55	0.57	-
Fernandez et. al. [86], 2018	Twitter	Semantic Context	SVM	Building a representation of the semantic context of the terms that are linked to radicalised rhetoric	Relation between the term and the entity has not been considered.	-	0.84	0.85	0.84	0.82

Table 5 continued

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Wiegand et al. [89], 2018	Twitter	Word embedding	SVM	Propose novel features employing information from both corpora and lexical resources	Potential challenges in detecting subtle abuse, domain-specific nuances not fully addressed, and varying effectiveness in different scenarios.	-	0.82	0.80	0.81	0.80
Warner et. al. [90], 2012	, Yahoo (3k)	part-of-speech tagging	SVM	Detecting hate speech in online text.	Sensitivity to labeling definitions, low recall, and the need for further research on classification methods.	0.94	0.68	0.60	0.63	0.69
Davidson et. al. [91],2017	Twitter (25k), English	Part-of-Speech (POS) tag, unigrams, bigrams, and trigrams	SVM	Automatic hate-speech detection on social media	Challenges in distinguishing hate speech, understanding biases, and contextual algorithm nuances.	-	0.91	0.90	0.90	0.95

reported a way to be restrictive in social media and to reduce the reproduction of stereotypes and discrimination in the future. In the realm of different languages as its domain, the classical techniques excel in hate speech detection, ensuring a more inclusive online environment. For instance, the work of Ozel et. al. [85] on Turkish language, Alfina et. al. [59] on Indonesian languages etc. Fernandez. et. al. [86] play the role of basis of future works within and across the Semantic Web and the Social Web.

3 Negative opinion analysis

Social media sentiment analysis method 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. Negative Emotion 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. Emotion can be categorized into positive, negative, and neutral class labels. 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 plays an important role to suggest how to operate business houses, from opinion polls to creative marketing strategies.

Transformers are superior for text classification because of their ability to handle large amount of sequential data effectively. The transformer models consist of 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 Negative opinion analysis using deep learning method

In this sub-section, we conducted a comprehensive review of several documents focusing on the architecture and features used in the models. Among the popular algorithms employed, BERT, LSTM, and CNN were extensively studied. 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 deep learning layers, which was the core of the architecture. The algorithms are allowed for creation of complex models that effectively analyzed and interpreted large amount of text data.

Future prospects encompass implementing the Sayyida et al. model [92] to analyze sentiment in under-resourced languages and extend its application to multi-class classification tasks. This approach finds utility in industries for product sentiment analysis and obtain valuable insights from textual data, across various domains. Moreover, the forthcoming potential lies in amalgamating the lexical algorithm proposed by Alwakid et al. [93] with machine learning techniques, aiming to enhance classification accuracy, particularly in the domain like "hate speech." This approach can be broadened to encompass neutral text analysis, offering valuable applications in linguistic and sentiment analysis tasks. It contributes to areas such as social media monitoring, customer feedback analysis, and content moderation across Arabic dialects. Expanding the model introduced by Chandra et al. [94] involves enlarging the image dataset and exploring multi-modal sentiment analysis. This extension enables applications in diverse areas like brand perception analysis, content moderation, and market research by providing deeper insights into sentiments, expressed in social media images. Furthermore, the evolving landscape of visual sentiment analysis, as indicated by Hassan et al. [95], includes considerations such as multi-modal datasets, annotators' demographics, and incorporation of more intricate visual cues. The applications span disaster monitoring, social media content moderation, and enhanced insights by extracting features from images and videos across diverse domains. Lastly, the future trajectory, as outlined by Sufi et al. [96], revolves around enhancing disaster monitoring through advanced NLP and AI technologies. This expansion encompasses broader language support, improved disaster type recognition, and accessibility through mobile applications. The real-world applications extended to crisis management, disaster response, and the development of comprehensive analytical intelligence for a more profound understanding of global disasters. In another study, Vatambeti et. al. [97] integrate the findings from their customer relationship management strategies, focusing on areas highlighted by customers' sentiments to improve their offerings. By paying attention to real-time feedback from users on platforms like Twitter, businesses can swiftly address concerns and capitalize on positive feedback. Additionally, by expanding analysis to include other languages and geospatial data, brands can fine-tune their localized marketing strategies and improve their global footprint. Integrating temporal patterns can also help companies to identify seasonal trends or time-sensitive issues, allowing them to proactively manage their services. In another study Bello et. al. [98] implementing the combined BERT with traditional models in sentiment analysis tools that can provide businesses with more accurate insights into customer feedback from various platforms. This deeper understanding can guide product development, marketing strategies, and customer service approaches. To further enhance the reliability of insights, businesses should also consider integrating data from offline sources or specialized forums, ensuring a more holistic understanding. Furthermore, by expanding the analysis to capture specific emotions, businesses can gain a more nuanced perspective on customer sentiments, allowing for tailored responses to different emotional triggers.

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 strengthens the ability of these models to handle and analyze large amount of text data. Table 6 summarizes the emotion analysis researches using deep learning techniques.

3.2 Negative emotion analysis using classical method

This approach refers to detection of emotion 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 suggest that deep learning algorithms give superior results than classical machine learning algorithms, in general. For example,

Table 6 Article summary of negative opinion analysis using deep learning

Author, Year	Context /Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Sayyida et. al. [92], 2022	Facebook	Word2vec, GloVe, FastText, BERT	CNN, LSTM, CBRNN	Suggesting a sentiment analysis framework with the capability of processing noisy data in a comprehensive manner.	BERT's data dependency and applicability need deeper exploration.	0.90	0.96	0.91	0.94	0.94
Alwakid et. al. [93], 2022	Twitter	Part of speech tagging	Lexical base approach	Propose a unique method for analyzing the sentiment of Arabic language tweets, using lexical analysis techniques	Challenges in domain feature recognition and neutral text evaluation.	0.89	0.86	0.87	0.85	0.85
Qian et. al. [99], 2022	Twitter	glove embeddings	deep neural net(DNN)	Analyzing neural networks and sentiments to understand NFTs' rising popularity.	Lack of quantitative parameters, reliant on public sentiments.	-	-	-	-	-
Zhigang Jin et. al. [100], 2022	Chinese Weibo and English SST2 dataset	Part of speech tagging	MH-GAT combines co-occurrence, syntactic graphs, BiLSTM, Att-BLSTM, and RCNN.	Analyzing social media sentiment using both dependency and co-occurrence graphs	Dependence on word segmentation, lengthy co-occurrence graph construction.	0.90	-	-	0.89	-
Chandra et. al. [94], 2022	Facebook, Instagram, Youtube	Image to pixel	VGG-19, ResNet50V2, and DenseNet-121	Fine-tuned transfer learning models effectively analyze image sentiment challenges..	The study deals with visual sentiment analysis but overlooks multimodal content challenges.	-	0.86	0.88	-	-
Hassan et. al. [95], 2022	Twitter	Image to pixel	VGGNet (places + ImageNet), Inception-v3 (ImageNet), ResNet-101 (ImageNet)	Suggest a deep learning framework for disaster image sentiment analysis.	Article emphasizes disaster image sentiment, misses broader applications and annotation biases.	0.92	0.89	0.89	0.89	0.88

Table 6 continued

Author, Year	Context /Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Sufi et. al. [96] , 2022	Twitter	-	convolution neural network (CNN) based anomaly detection, automated regression and Getis-Ord Gi algorithm	Automated Disaster Monitoring From Social Media Posts	System's accuracy in disaster monitoring may fluctuate with post quality and diversity.	0.97	0.93	0.88	0.90	0.90
Vatambeti et. al. [97], 2023	Twitter, 13k	Word2Vec, GloVe	Bi-LSTM, ConvBiLSTM, CNN	Twitter sentiment analysis for Swiggy, Zomato, UberEats shows industry insights.	The study solely focused on Twitter data, neglecting insights from other major social media platforms.	0.92	0.92	0.92	0.92	0.92
Bello et. al. [98], 2023	Twitter, 16k	Word2vec	BERT+CNN, BERT+RNN, BERT+BiLSTM	Study proposes BERT-based text classification, using NLP and its variants.	Traditional NLP approaches often miss deeper word context, while online data sources may lack reliability.	0.93	0.96	0.95	-	0.95

the model based on Idan et. al. [101] work, can be implemented for political campaigns, targeting the voters more effectively by understanding their behavior. By augmenting it with additional traits, improving data completeness, and considering temporal features, this model could enhance political prediction and engagement efforts in the future. The work of Laszlo et. al [102] has been implemented using RNN model and enhancing its interface can help to analyze the organizations and categorize the tweets based on emotions more effectively. Expanding the analysis, classification, and data visualization capabilities can provide deeper insights and support various applications beyond emotional analysis in different domains. The study [103] highlights the potential to using social media sentiment analysis that may complement traditional polling methods, offering more efficient and diverse data collection. Organizations conducting surveys for polling, can integrate sentiment analysis from platforms like Twitter to gather additional insights, potentially making their assessments more comprehensive and reflective of public opinion. The BPEF[104] framework can be implemented by businesses and organizations for precise sentiment analysis on Twitter, providing insights into public perceptions and reactions. The refined and balanced sentiment polarity metrics produced by BPEF can guide marketing strategies, PR responses, and product development that aligning them better with public sentiment and ensuring more effective communication on social media platforms. Implementing the proposed method[105] can enhance sentiment analysis on Twitter data by addressing its unique challenges. It combines lexicon-based and learning-based approaches to effectively identify and classify opinionated tweets, which can be valuable for businesses and researchers to analyzing public sentiment on the platform. In another study [106], then proposed framework can revolutionize how businesses perceive public sentiment on Twitter, by classifying emotions beyond just positive and negative. By using this method, businesses can gain deeper insights into specific sentiments, leading to more tailored marketing or public relations strategies.

The highest accuracy obtained among the classical records is 88.34% [107]. Tables 7 and 8 summarizes emotion analysis using classical methods.

4 Profanity detection

Profanity and offensive language on social media platforms, including Twitter, have gained attention in recent years due to concerns about their impact on online discourse and interactions. Profanity refers to the use of language that is considered socially offensive, including cursing, cussing, swearing, bad language, foul language, obscenities, expletives, or vulgarism. It can take various forms, such as explicit words, slurs, insults, and derogatory comments.

The prevalence of profanity on social media platforms has raised concerns about the impact it may have on online communication and the overall tone of online discourse. Profanity can contribute to the escalation of conflicts, online harassment, and the spread of negativity in online interactions. It can also create barriers to constructive dialogue and impede meaningful conversations on important topics. Profanity is relatively common on Twitter as observed by the researchers. Some studies have estimated that as much as 10-20% of all tweets contain profanity. The use of profanity on Twitter can be influenced by various factors, including the topic of discussion, the tone of the conversation, the demographics of the users, and the cultural and societal norms.

Efforts have been made by social media platforms to regulate and moderate the use of profanity and offensive language by applying automated tools and human moderators. However, striking the right balance between freedom of expression and addressing offensive

Table 7 Article summary of negative opinion analysis using classical method

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Read at. al. [108], 2005	Newswire dataset	N-Gram	Naive Bayes and SVM	Sentiment classification can be influenced by factors like domain, topic, time, and language style.	Addressing various factors and specific word influences is crucial for effective sentiment classification.	0.82	-	-	-	-
Go et al. [109], 2009	Twitter	N-gram and Part -of - Speech tagging	Naive Bayes, Maximum Entropy, and SVM	Using emoticons as approximate labels proves effective for distant supervised learning.	Relies on emoticons, potential bias, not applicable universally.	0.83	-	-	-	-
Davidov et al. [106], 2010	Twitter	N-grams, patterns, and tweet-based features	KNN	A framework was created to detect and categorize sentiments in short text snippets using Twitter data.	Relies on Twitter-specific labels, potential bias in hashtag sentiment.	0.86	-	-	-	-
Zhang et al. [105], 2011	Twitter	N-gram	SVM	Suggested a new approach to address the limitations of existing sentiment analysis methods based on lexicons and machine learning techniques.	Relies on initial lexicon-based method for training data.	-	0.68	0.82	0.74	-
Agarwal et al. [110], 2011	Twitter	Part -of -Speech tagging	SVM	The author enhanced the previously state-of-the-art unigram model, achieving over a 4% improvement in classifying positive, negative, and neutral outcomes.	Limited use of follower graph, undirected relationships, link context. 0.60	-	-	0.62	-	

Table 7 continued

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Speriosu et al. [111], 2011	Twitter	N-gram	Lexicon-based approach	The authors showed that using distant supervision with a maximum entropy classifier outperforms the lexicon-based ratio predictor.	Limited use of follower graph, undirected relationships, link context.	0.71	-	-	-	-
Saif et al. [112], 2013	Twitter	N-gram and Part -of - Speech tagging	SVM, NB	The study proposed using semantic features in Twitter sentiment analysis and evaluated three integration methods: replacement, augmentation, and interpolation.	Reliance on Alchemy API for coarse semantic concept mappings.	0.88	0.77	0.76	0.76	0.76
Lin et. al. [113], 2012	Twitter	Feature Hashing	Logistic Regression classifier	Twitter integrated machine learning with its Hadoop and Pig-centric analytics platform; the paper presents a relevant case study.	No consensus on best practices for predictive analytics.	-	-	-	-	-
Clark et al. [114], 2013	Twitter	N-gram, lexicon and polarity strength	Naive Bayes	This supervised system combines many features to classify positive and negative emotion at the phrase level.	Bug in preprocessing removed emoticon features, potentially affecting results.	0.89	-	-	-	-
Hassan et al. [104], 2013	Twitter	A combination of unigrams and bigrams of simple words, part-of-speach and semantic features derived from WordNet and SentiWordNet 3.0	RBF Neural Network, Random Tree, REP Tree, Naive Bayes, Bayes Net, LR and SVM.	proposed and evaluated a robust ensemble framework capable of effectively classifying Twitter sentiments.	Reliance on three parameter components, potential for search method refinement.	0.71	-	0.77	-	-

Table 8 Continued

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Yuliyanti et. al. [115], 2017	Twitter	TF-IDF	Principal Component Analysis SVM	Success level of the community development program	Limited tweet sample, model influenced by SVM parameters, preprocessing.	-	0.82 -	-	-	-
Mansour et. al. [116], 2018	Twitter	TF-IDF	Lexicon base approach	User’s Responses to terrorism using sentiment analysis and text mining	Analysis limited to specific countries, needs more diverse data.	-	-	-	-	-
Saragih et. al [117], 2017	Facebook and Twitter comments	TF-IDF	Lexicon base approach	Sentiment Analysis of Customer Engagement on Social Media	Study limited to Indonesia, two platforms, three companies.	-	-	-	-	-
Hassan et. al. [118], 2017	Twitter and newsgroup	POS Tagger, N-Gram, Unigram	SVM, NB, Maximum Entropy(ME)	Comparison among SVM, NB and ME classifiers regarding sentence level sentiment analysis for depression measurement	Only three classifiers tested; broader machine learning algorithms unexplored.	0.91	0.83	0.85	-	-
Joyce et. al [103], 2017	Twitter	-	- Naive Bayes, Lexicon base approach	Sentiment Analysis of Tweets for the 2016 US Presidential Election	Study limited to Trump and Clinton; may not generalize.	0.85	-	-	-	-
Ikoro et. al. [119], 2018	Twitter	-	Lexicon base approach	Analyzing Sentiments Expressed on Twitter by UK Energy Company Consumers	Focused only on Britain’s energy providers, may not generalize.	-	-	-	-	-
Hao et. al. [120], 2016	Twitter	-	- Lexicon base approach	Social media content and sentiment analysis on consumer security breaches	Data sample collected from a short period on Twitter.	-	-	-	-	-

Table 8 continued

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Shayaa at. al [121], 2017	Multiple channel social media	-	Lexicon base approach	Negative emotion score on employment	Sentiment analysis based on limited time frame, data sources.	-	-	-	-	-
Ali et. al. [122], 2017	Twitter, Reddit, Instagram	Geo-Tagging	Directed Acyclic Graph	Sentiment Analysis as a Service	Limited data tested, performance scalability, and data fusion challenges.	-	-	-	-	-
Laszlo et. al [102], 2020	Twitter	-	TextBlob, RNN	Social media sentiment analysis based on COVID-19	Neutral results in TextBlob, interface improvement, data expansion.	-	-	-	-	-
Idan et. al. [101], 2020	Facebook	-	Naive Bayes	Predicting Voting Behavior in Social Networks	Model limited to static attributes, lacks temporal behavioral dynamics.	-	0.84	0.87	0.82	-
Arias et. al. [123], 2022	Twitter, Instagram, Youtube	TF-IDF, GloVe, word2vec, n-gram	ML techniques SVM, LR, NB, KNN, RF	Sentiment Analysis of Public Social Media as a Tool for Health-Related Topics	Dependence on text, lacking multilingual and multimedia adaptability.	-	-	-	-	-

language remains a challenge. It is important for the users to be mindful of their language use on social media and strive to engage in respectful and constructive communication online. Deep learning models have shown promising results in detecting profanity and offensive language on Twitter and other social media platforms, compare to traditional machine learning approaches.

4.1 Profanity detection using deep learning method

Deep learning techniques have indeed been increasingly used in recent years to analyze profanity and offensive language on social media platforms like Twitter, as well as other online forums. These techniques involve the use of artificial neural networks, which are trained on large datasets of annotated text to recognize patterns and features associated with profanity and offensive language. By leveraging large amounts of data, deep learning models can achieve high accuracy in detecting and classifying different types of abusive language. This has led to the development of various profanity detection models that can be used for blocking, filtering, or alerting users about the use of abusive language on social media platforms. In addition to detecting profanity and offensive language, deep learning models have also been applied to other use cases, such as identifying hate speech, cyberbullying, sarcasm, irony, and even detecting potential mental health issues based on language patterns. These models can help in creating a safer and more inclusive online environment, by allowing moderators and administrators to quickly identify and take action against abusive behavior.

Bilal et al. [54] emphasize the potential of transformer-based models on social media platforms for South Asian users to effectively combat hate speech in Roman Urdu. For a safer online environment for Roman Urdu speakers, platforms are encouraged to refine the lexical normalization process and enhance annotation guidelines. Ajlan et al. [124] advocate for the adoption of firefly-CDDL on mainstream social media platforms as it automates cyberbullying detection. This promotes rapid response to threats, with the system's self-optimization ensuring resilience against changing online harassment patterns. According to Chaudhari et al. [125], the technology they discussed provides video platforms with tools for proactively moderating offensive audio content. By neutralizing both the audio and the speaker's lip movements, a comprehensive solution emerges, particularly useful for platforms inundated with user-generated content. Bhowmick and colleagues [126] suggest that their model, when applied to social media platforms, can detect and limit derogatory content targeting known individuals. When the framework includes multiple languages, it can provide global protection against targeted harassment. Kumari et al. [127] demonstrate that their model, once implemented on popular platforms, can aid real-time content moderation. Enhanced refinement can allow platforms to serve diverse audiences, fostering respectful interactions across languages and settings. Dadvar et al. [128] highlight the adaptability of DNN models to new datasets, noting their superior performance compared to traditional ML models. By including users' profile data and demographics, there's potential for more robust cyberbullying detection. Basavraj et al. [129] advocate for the integration of the HateBERT architecture into social media platforms. This aids in screening and flagging potential hateful content, promoting a respectful digital community. Continuous model training and updates are essential to keep pace with evolving language trends. In a study by Levent et al. [130], the LinearSVC model stands out for real-time applications due to its efficiency in detecting profanity in Turkish search engine queries. For greater accuracy, particularly in complex linguistic scenarios, transformer models like BERT and Electra are preferable. Dandeniya and team [] have developed a model that can be integrated across various digital platforms, ensuring content adheres to community standards by identifying and censoring offensive content. Enhancing the dataset diversity and refining onset detection can improve content monitoring without sacrificing user experience. Lastly, Galinato et al. [131] underscore the effectiveness of the Tagalog BERT model in context-based profanity classification and censorship for Tagalog texts. This can be beneficial for cleaner online spaces in social media platforms and communication tools. Future adaptions of this model for multilingual profanity detection, coupled with performance enhancements, could revolutionize global profanity detection solutions.

Tables 9 and 10 summarizes the works we studied on profanity detection using deep learning methods. Based on the information provided in Tables 9 and 10, we summarize the works studied on profanity detection using deep learning methods as follows: Methods and tools used: NLP, a mixed approach like machine learning and deep learning are used and also used LSTM, BLSTM, BERT, Attention, Bi-GRU, LR, SVM, FastText, Char-CNN, HybridCNN, WordCNN, NB, KNN, DT, RF, Bagging, AdaBoost, GloVe, ERNIE 2.0, TwitterRoBERTaOffensive, HateBERT, Logistic-Regression, SGDClassifier, LinearSVC, Random-ForestClassifier, CNN, RNN, and one-hot. Platforms and languages analyzed: Twitter, movie reviews, Turkish, Chinese, and others. Highest F1 score achieved: 0.93 by Levent Soykan et al.[130] (2022) in Turkish Profanity Detection using Logistic Regression, SGD Classifier, Linear SVC, Random Forest Classifier, LSTM, BERT, Electra, and T5 techniques. Highest Recall achieved: 0.99 by Dadvar, Maral, and Eckert Kai (2018) for Cyberbullying detection in social networks using deep learning-based models. Highest Precision achieved: 0.99 by Dadvar, Maral, and Eckert Kai (2018) for Cyberbullying detection in social networks using deep learning-based models. Ensembling models can provide better accuracy and F1 scores than individual models. Overall, the studies suggest that deep learning models can be used successfully for profanity detection, and ensembling models can improve the accuracy and F1 scores of the models.

4.2 Profanity detection using classical machine learning method

Social media is an open platform and people often misuse the freedom. A major display of profanity is cyberbullying. Cyberbullying is a huge phenomenon among teenagers as a victim or predator or bystander [5]. Authors in [144] 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 [145], 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. Table 11 represent some article summaries of profanity detection using classical method. Employing Sood et. al. [146] model for an enhanced profanity detection system can assist online platforms in automatically moderating and ensuring a respectful digital environment. With the added adaptability of the system to specific online communities through crowdsourced labeling, platforms can ensure tailored and context-aware content moderation, leading to more meaningful interactions and reduced manual oversight. The work of Chin et. al. [147] provide a valuable tool for music industry to automatically filter and screen explicit lyrics, saving time and effort in manual reviews and ensuring the appropriateness of songs, particularly for younger audiences. By integrating these models into their workflows, musicians and recording companies can preemptively assess whether their songs meet the

Table 9 Work summary of profanity detection using deep learning

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Phoey et. al [132], 2020	Twitter	BERT base features Representation	LSTM, BiLSTM and BERT	Profanity and Hate Speech Detection	Profanity-based hate speech detection varies across cultural contexts.	-	-	-	0.84	-
Pratik et. al. [133], 2019	Twitter	PoS tagging, Tf-Idf and GloVe	Attention and Bi-GRU	An attention ensemble based approach for multi-label profanity detection	Acquisition of specifically-labeled abusive language datasets is challenging.	0.97	0.82	0.84	0.76	0.75
JiHo et. al. [134], 2017	Twitter	-	LR, SVM, FastText, Char-CNN, HybridCNN, Word-CNN, HybridCNN	One-step and Two-step Classification for Abusive Language Detection on Twitter	Acquisition of specifically-labeled abusive language datasets is challenging.	-	0.88	0.85	0.86	-
Basavraj et. al. [129], 2021	Twitter	GloVe,TF-IDF	KNN, SVM, DT, RF, Bagging, AdaBoost, Voting, Twitter RobertaOffensive, HateBERT	Classification of Hate, Offensive and Profane content from Tweets using an Ensemble of Deep Contextualized and Domain Specific Representations	Universal hate speech definition lacking, dependent on multiple factors.	0.81	0.81	0.79	0.81	0.85
Levent et. al. [130], 2022	Twitter	N-gram, TF-IDF	Logistic-Regression, SGDClassifier, LinearSVC, RandomForest Classifier, LSTM, BERT, Electra, T5	A Comparison of Machine Learning Techniques for Turkish Profanity Detection	Model might miss profane words with uncommon suffixes or joined with other words.	0.98	0.98	0.87	0.93	0.98
Dadvar et. al [128], 2018	Twitter, Wikipedia	Glove	CNN, LSTM, BLSTM attention, BLSTM	Cyberbullying detection in social networks using deep learning based models	Imbalanced cyberbullying datasets may affect model performance.	0.99	0.97	0.98	0.99	0.9

Table 9 continued

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
BaWazir et. al. [135], 2021	MMUTM foul language dataset	-	CNN, RNN, LSTMA, lexnet, VGG16, GoogLeNet, and Resnet50	Design and Implementation of Fast Spoken Foul Language Recognition with Different End-to-End Deep Neural Network Architectures	Accuracy reduction in noisier environments; limited to speaker-independent mode.	-	0.97	-	0.97	-
Kim et. al. [136], 2022	Twitter, Facebook	-	LSTM	A Study of Profanity Effect in Sentiment Analysis on Natural Language	Limited scope, sample size, cultural bias, lacks external validity.	0.83	-	-	-	-
Dandeniya et. al. [137], 2023	MMUTM foul language dataset	word2vec	RNN,CNN	To develop a model to identify the F-words in a speech audio file using advanced deep neural network technologies	Computational challenges, imperfect synchronization, onset detection issues.	0.98	0.98	0.98	0.98	0.98

Table 10 Continued

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Kumari et. al. [127], 2019	Facebook and Twitter	One-hot, GloVe, fastText Embeddings followed by CNN	CNN	Deep Learning Approach for Identification of Abusive Content	Limited embeddings tested, challenges with multi-modal and code-mixed languages.	-	-	-	0.78	-
HsuYang et. al. [138], 2020	TOCP, a larger dataset of Chinese profanity	Word2Vec, fastText	CNN, BiLSTM,	TOCP, A Dataset for Chinese Profanity for detection and rephrasing	Limited coverage of rule-based systems, dataset specific to Chinese.	0.77	0.85	0.87	0.86	0.86
Woo et. al. [139], 2022	Dataset developed by XLGames	Grapheme and syllable separation-based word embedding	CNN	Improving Korean Profane Detection using Deep Learning	Limited to Korean, dictionary reliance, lacking advanced model evaluation.	0.90	0.92	0.93	0.92	0.91
Sazzed et. al. [140], 2021	YouTube, BengSentiLex	part-of-speech (POS) taggers	LR, SVM, SGD, CNN, LSTM, BiLSTM	Creating lexicons for sentiment analysis and profanity detection in low-resource Bengali language	Lexicon size, expanding to multi-domain training corpora.	0.93	0.93	0.90	0.92	0.92
Al-Hashedi et. al. [141], 2019	Cyberbullying detection model using Kaggle dataset	word2vec, GloVe, Reddit and ELMO	GRU, LSTM and BLSTM	Cyberbully detection using deep learning	Model overfitting with binary dataset.	-	-	0.98	-	-
Bhowmick at. al. [126], 2021	Facebook, Twitter	FaceNet Embidding, BERT Base Embidding	Distil-BERT, ELECTRA, XLM-RoBERTa, FaceNet	A multimodal deep framework for derogatory social media post identification of a recognized person	Limited dataset size, reliance on pre-trained models, language specificity.	0.90	-	-	-	-

Table 10 continued

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Malik et. al. [142], 2021	Twitter and Facebook (ALONE and HASOC'20)	fastText (non-context based) and BERT (context based)	LR, SVM, DT, RF, XGBoost, CNN, MLP, LSTM	Toxic speech detection	Applicability to diverse platforms, real-time implementation.	0.82	0.83	0.82	0.81	0.82
Marwa et. al. [143], 2018	Tweets	GloVe and Word2ve	SVM, NB, CNN, LSTM, BLSTM	Deep learning for online harassment detection	Limited data size, lack of detailed user analysis.	-	0.80	0.80	0.71	0.76
Chaudhari et. al. [125], 2021	iBUG 300-W	GloVe, fastText, and Word2Vec	CharCNN, WordCNN and HybridCNN	Profanity Detection and Removal in Videos using Machine Learning	Focuses only on audio; ignores visual profane indicators.	0.82	-	-	-	-
Bilal et. al. [54], 2023	Largest Roman Urdu , (173,714)	BERT embeddings	LSTM, BiLSTM, BiLSTM + Attention Layer, and CNN	They employed a transformer-based model for Roman Urdu hate speech classification	Limited BERT training, context challenges, and inadequate lexical normalization.	0.96	0.97	0.96	0.97	0.96
Al Ajlan et. al. [124], 2023	20-UCI	Word embedding	Deep CNN	A Firefly-Based Algorithm for Cyberbullying Detection Based on Deep Learning	Dependent on algorithm optimization, potential overfitting with high accuracy.	0.98	0.87	0.76	0.81	0.85
Galinato et. al. [131], 2023	TTP dataset 14000 tweet	WordPiece	BERT	Context-based Profanity Detection and Censorship using BERT	Relies on a single pre-trained model for detection.	-	0.84	0.83	0.84	-

Table 11 Article summary of profanity detection using classical method

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Sood et. al. [146], 2012	Social news site	-	SVM	Using crowd sourcing to improve profanity detection	High recall optimization remains challenging; relies heavilyon crowd-sourcing.	0.93	0.62	0.64	0.63	0.85
Chin et. al. [147], 2018	South Korean Broadcasting System (KBS),song screening result data	TF-IDF	Naive Bayes, Decision Tree, SVM, MCES	Explicit content detection in music lyrics	Difficulty detecting metaphorically expressed explicit content in lyrics.	-	0.94	0.84	0.88	0.84
Gottipati et. al. [145], 2020	Facebook and Twitter	N-grams, FastText	NaÏve Bayes, Logistic Regression, Stochastic Gradient Descent (SGD)	Leveraging Profanity for Insincere Content Detection-A Neural Network Approach	Model not tested across diverse social media platforms.	-	-	-	0.95	-
Nobata et. al. [148], 2016	Yahoo!, WWW2015 Set	N-gram, POS, word2vec	Lexicon based approach	Abusive language detection in online user content	Abusive language detection in online user content	-	0.83	0.84	0.83	-
Baby et. al. [149], 2023	Twitter and MySpace datasets	TF-IDF	KNN, NB, SVM, DT, RF, LR	Psychosomatic Study of Criminal Inclinations with Profanity on Social Media	Requires improved accuracy, execution time, and evaluation matrices.	0.88	0.99	0.88	0.93	0.90

screening criteria, reducing the likelihood of issues during official screening reviews and creating more standardized and objective content evaluation processes.

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 detection

Targeted insults on social media encompass directed abusive or offensive language and behaviors towards specific individuals or groups. Such behaviors can manifest as name-calling, harassment, or even violent threats. While statistical data on targeted insults can fluctuate based on platforms, study populations, and time frames, research consistently indicates that women and marginalized communities bear the brunt of these targeted insults. Anonymity and the resultant lack of accountability on social media platforms further exacerbate the intensity and frequency of such insults.

5.1 Targeted insult detection using deep learning method

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. Tables 12 and 13 represent some article summary of targeted insult detection using the deep learning method. In recent years, a plethora of studies have emerged emphasizing the necessity and effectiveness of implementing advanced algorithms and models to enhance content moderation on various digital platforms. Ensuring safer and more respectful online interactions has been at the forefront of these endeavors. C Raj et. al. [150] pioneered an innovative cyberbullying detection system. By harnessing the power of bidirectional RNNs combined with attention-based models, they paved the way for automated detection and mitigation of online harassment on social media. A separate exploration by A Kalaivani et. al. [151] delved into the BERT pre-trained model, facilitating automated detection and categorization of offensive language in English, Danish, and Greek on digital platforms. K Shanmugavadivel et. al. [152] took a slightly different approach, integrating Adapter BERT into sites like YouTube. Their primary objective was to gauge user sentiments and pinpoint offensive language, ultimately enhancing content moderation. Similarly, M Alotaibi et. al. [153] employed a multi-architecture deep learning model on platforms such as Twitter, honing in on real-time cyberbullying detection. Lazaro et. al [154] emphasized the significance of refined deep learning models, particularly when combined with the likes of BERT, for efficient content processing. R Sodhi et. al. [155] showcased a model that specifically targets indirect insults, suggesting the potential of unsupervised text style transfer methods for advanced content moderation. Z Zhang et. al. [156] combined CNN-GRU models to tackle real-time hate speech detection on social media. They

Table 12 Article summary of targeted insult detection using deep learning method

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Thenmozhi et. al. [165], 2019	Social Media (OLID)	TF-IDF	BiLSTM	Offensive language identification in social media using traditional and deep machine learning approaches	Reliance on specific vectorization methods limits model generalizability.	0.83	-	-	0.53	-
S Alsafari et. al. [157], 2020	Twitter	N-grams, Fasttext	NB, SVM, LR, CNN, LSTM, GRU	Hate and offensive speech detection on Arabic social media	Model struggles to distinguish profane, hateful, and offensive posts.	-	0.81	0.84	0.82	0.82
A Parikh et. al. [162], 2019	Social Media (HASOC 2019)	GloVe	CNN, Navie Bayes, Logistic Regression	Identification of Hate Speech using Machine Learning and Deep Learning approaches	Limited accuracy; struggled differentiating profane, hateful, offensive posts.	-	-	-	0.64	-
S Thara et. al. [166], 2022	Twtter	Word2Vec, FastText	CNN, BiLSTM, GRU, XLM-Roberta	Offensive language identification in social media	Scarce training data; transfer learning for multilingual model needed.	0.83	-	-	0.53	-
TL Sutejo et. al. [167], 2018	Facebook, Twitter, Youtube	N-grams, GloVe BOW(Bag-of-Words), TF-IDF, FastText	LSTM,	Indonesia hate speech detection using deep learning	Acoustic model underperformed compared to textual and multimodel.	-	-	-	0.87	-
Z Zhang et. al. [156], 2018	Twitter, Yahoo!	Out-of-vocabulary (OOV)	SVM, CNN, LSTM	Detecting hate speech on twitter using a convolution-gru	Detecting abstract hate concepts solely from text is challenging.	0.91	0.91	0.91	0.94	0.92

Table 12 continued

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
R Sodhi et. al. [155], 2021	Jibe and Delight Corpus (JBC) Dataset	GloVe , Bert based Embedding	Logistic Regression, SVM, BERT, RoBERTa, and XLNet	A Dataset of Targeted Insults and Compliments to Tackle Online Abuse	Unsupervised text style transfer for negativity remains challenging.	0.87	0.98	0.89	0.88	0.87
Chakravarthi et. al.[158], 2023	Dravidian CodeMix[168]	Bert based Embedding	Newly proposed fusion of MPNet and CNN model	Categorize code-mixed social media comments and posts in Tamil, Malayalam, and Kannada into offensive or not offensive at different levels.	Models excel in specific languages but lack generalization.	-	0.98	0.97	0.98	-

Table 13 Continued

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
Lazaro et. al [154], 2020	Offensive Language Identification dataset (OLID)	GloVe	BiLSTM, BiGRU, LSTM,RNN, CNN,LSTM	Identifying and categorizing offensive language	Neural network configuration is challenging; training process slow.	0.89	-	-	0.66	-
M Alotaibi et. al. [153], 2021	Twitter	of words (BoW)	BiGRU, CNN	A multichannel deep learning framework for cyberbullying detection	Method may underperform without larger datasets and diverse languages	0.89	0.89	0.89	0.89	0.89
M Zampieri et. al. [169], 2019	Social Media (OLID)	FastText	SVM, BiLSTM, CNN	Predicting the type and target of offensive posts using proposed OLID dataset	Lacks cross-corpus comparison and multilingual dataset extensions currently.	-	0.79	0.89	0.80	0.80
S Sharifirad et. al. [163], 2019	Twitter	Glove	BiLSTM,BERT,RNN,CNN	Detect online harassment on social networking platforms	Cultural biases and vague boundaries challenge harassment annotations.	0.84	0.83	0.84	0.84	0.84
K Shanmugavadivel et. al. [152], 2022	Twitter	BERT and GLOVE embedding	BiLSTM,BERT, RoBERTa, Adapter-BERT	Offensive language identification on multilingual code-mixed data	Model accuracy needs enhancement for sentiment and offensiveness.	0.79	-	-	0.80	-
A Kalaivani et. al. [151], 2020	Social Media(OLID)	Word2vec	BERT	Offensive language identification in English, Danish, Greek using BERT	Limited exploration of alternative models and feature sets for potential performance enhancement.	-	0.80	0.81	0.77	0.82

Table 13 continued

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
C Raj et. al. [150], 2021	Wikipedia Attack Dataset	TF-IDF,GloVe	LSTM, Bi-LSTM,GRU, and Bi-GRU	Cyberbullying Detection: Hybrid Models	Reliance on shallow neural networks may overlook deeper contextual nuances	0.83	-	-	0.98	-
A Cojocaru et. al. [159], 2022	Romanian local news website (4,052)	BERT embedding	RoBERT + MLP, RoBERT + CNN multiBERT + CNN	Propose a novel Romanian language dataset for offensive message detection	Dataset limited, single annotator, relies heavily on BERT models.	0.74	0.63	0.67	0.60	0.69
Abbasi et. al. [160], 2022	Toxic Comment Classification Challenge	GloVe, Word2vec, and FastText	CNN, NN, BiLSTM, GRU, BiGRU	This research analyzes and compares modern deep learning algorithms for multilabel toxic comments classifcation.	Data imbalance, time-consuming training of individual models.	0.94	0.96	0.96	0.96	0.95

argue that a user-centric approach could be instrumental in refining this model. Similarly, TL Sutejo et. al. [146] championed the application of word embeddings, particularly CBOW architecture, in detecting hate speech in the Indonesian language. Further highlighting the importance of multilingualism, S Thara et. al. [157] discussed the potential of deep learning models in sentiment analysis on Malayalam-English code-mixed platforms. S Alsafari et. al. [148] proposed the hierarchical CNN approach as a preliminary measure to detect and categorize hate speech. Chakravarthi et. al.[158] suggests the fusion of MPNet and CNN models for content moderation in Dravidian languages, hinting at the broader potential of diverse neural network architectures. A Cojocaru et. al. [159] highlighted the RoBERTa-based model combined with CNN layers for the Romanian language, emphasizing its application in improving online discourse. Finally, Abbasi et. al. [160] presented a deep learning model that specifically targets toxic comments online. Their research underscores the importance of addressing data imbalance and optimizing training time to enhance detection efficiency. Mostly LSTM and NLP has been used to predict whether the text in the papers are hateful or not. Authors in [128, 153, 156, 161–163] have used a Twitter dataset. Out of all the papers [161] has the highest accuracy, precision and F1-score of all with 0.97, [164] has the second highest accuracy of 0.95 and [128] has the second highest precision 0.93.

5.2 Targeted insult detection using classical methods

This approach is also known as the shallow method. This method depends on an automatically or manually coded dataset, used to train the learning models to detect and classify the text as targeted insult or non. These classical Machine Learning algorithms include support vector machines (SVM), Naive Bayes (NB), Logistic Regress (LR) etc Table 14 summarizes targeted insult detection models using classical method.

Mostly SVM has been used to classify the text whether targeted insult or not with good accuracy. Authors in [170] have used a dataset which is multi-lingual and has both Hindi and English Language. The work [154] shows highest accuracy 0.89 and [155] shows the highest precision of 0.98 and F1-score of 0.883.

Recent findings indicate that certain techniques can be effectively integrated into our daily processes. For instance, the approach put forward by Bharathi et al. [164] aligns with the idea of employing machine learning models that leverage TFIDF and count vectorizer features. Such models are adept at detecting and flagging code-mixed content in Dravidian languages across social media platforms. In another strategy [171], there's the potential for creating an advanced content moderation system for Arabic social media content. Based on handpicked datasets, this system can autonomously sieve out hate speech, classify content, and gauge sentiment. Deploying this on platforms like Facebook, Twitter, Instagram, and WhatsApp could uplift user engagement, cultivate constructive conversations, and limit the dissemination of harmful content. Additionally, a study [172] indicates the feasibility of a digital moderation tool designed for sites like Formspring. Through linguistic analysis, this tool can auto-identify and assess posts abundant in derogatory language, potentially decreasing cyberbullying occurrences. Moreover, a distinct research [173] has brought forth an AI-infused moderation tool tailored for Twitter. This innovation is set to automatically flag content that's politically sensitive, especially during pivotal moments. Another research [173] has fashioned a machine learning mechanism to spot hate speech in Indonesian tweets using SVM and word unigrams. The same research also underscores the prospective benefits of adopting deep learning techniques with more extensive datasets for optimized outcomes. Arnisha et al. [174] have pioneered an AI-driven cyberbullying detection mechanism for social media

Table 14 Article summary of targeted insult detection using classical method

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
S Kurniawan et. al. [173], 2020	Facebook, Twitter	ROC N-grame, TF-IDF	NB, SVM	Indonesian tweets hate speech target classification	The study is constrained by a limited dataset size, potentially hindering deep learning effectiveness.	0.77	-	-	0.84	-
Md Fahim et. al. [172], 2021	Twitter, Private Dataset	N-grams, TF-IDF	Logistic Regression, Naive Bayes, SVM	Detecting Offensive Content on Twitter During "Proud Boys Riots"	Lacks capability to detect sarcasm and differentiate between various politically motivated rhetorics.	0.87	0.88	0.93	0.88	0.89
Kelly et. al. [171], 2011	Private Dataset	-	SVM	Using Machine Learning to Detect Cyberbullying	limiting detection capabilities in a small Formspring sample	0.81	-	-	-	-
A Omar et. al. [161], 2021	Twitter	TFIDF, N-gram, BoW	LinearSVC, Logistic Regression	Multi-label arabic text classification	Limited dataset sources and features for comprehensive Arabic text classification in OSNs.	0.97	0.97	0.97	0.97	0.97
MSA et. al. [177], 2022	Facebook	Word2Vec, Doc2Vec, and Fasttext	LR, SVM,RF,KNN	Detection of Hate Speech Texts Using Machine Learning Algorithm	Lack of guidelines complicates comparing hate speech methods.	0.95	-	-	-	-
HA Nayel et. al. [178], 2019	Social Media (Hasoc2019)	TF-IDF	Stochastic Gradient Descent (SGD), SVM Linear Classifier	A Machine Learning Framework for Hate Speech and Offensive Language Detection	The study doesn't explore deep learning methods, which may limit its adaptability	-	0.91	0.90	0.90	0.90

Table 14 continued

Author, Year	Context / Dataset	Features Representation	Algorithm	Contribution	Limitation	A	P	R	F1	ROC
B Bharathi et. al. [164], 2021	Social Media (HASOC2019)	N-gram, TF-IDF	SVM, LR, RF	Offensive language identification on multilingual code mixing text	The study relies primarily on basic TFIDF and count vectorizer features and may not generalize well with deeper embeddings due to insufficient training data.	0.95	0.87	0.85	0.95	0.88
Arnisha et. al. [174], 2023	Facebook Bangla text dataset (44,001)	FastText TF-IDF	DT, RF, LR, MLP	A robust hybrid machine learning model for Bengali cyber bullying detection	Focused solely on Bengali text, lacks deep learning exploration.	0.98	0.98	0.98	0.98	0.98
Emad et. al. [175], 2023	38K-Tweet	FastText GloVe	SVM, LR	Persian Hate and Ofensive language using keyword-based data selection strategies	Model biases persist; human intervention still required.	-	0.851	0.851	0.851	-
Bharadwaj et. al.[176], 2023	Twitter and Facebook (20000)	TF-IDF	NaÏve Bayes, LinearSVC, LR, KNN	Machine Learning Algorithm for Detecting Cyberbullying Activities Automatically	Machine learning's dependency on quality data may limit effectiveness.	0.94	0.95	0.93	0.94	-

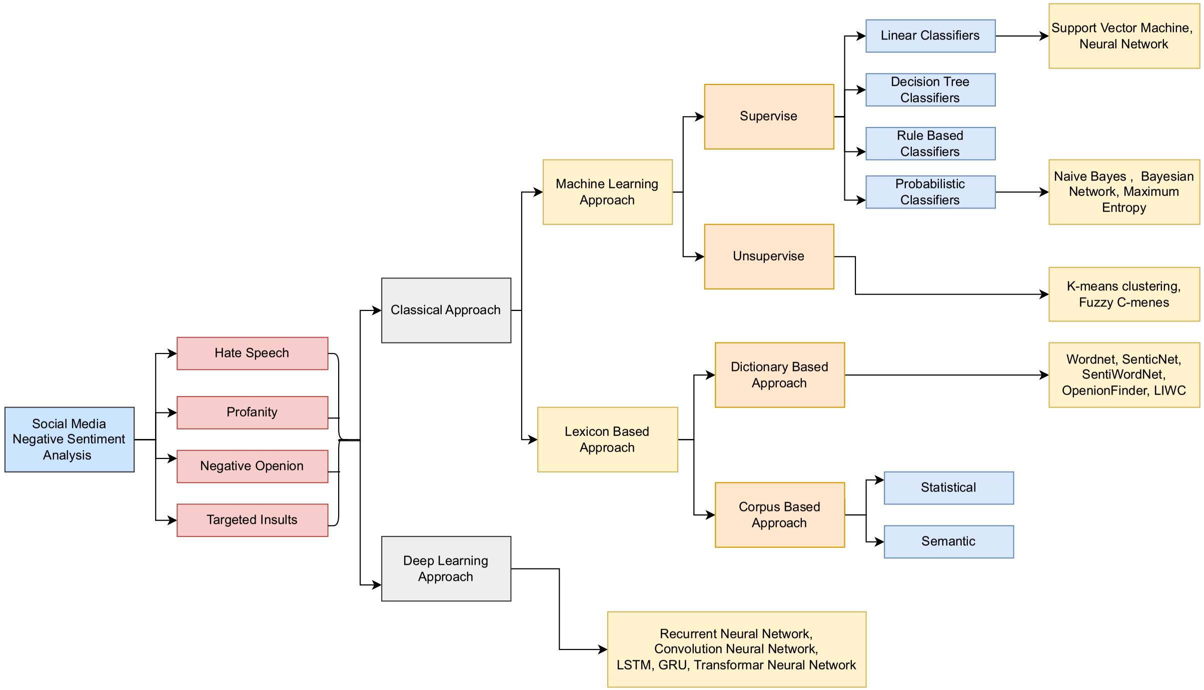

Fig. 1 Social media negative sentiment analysis methods

platforms, particularly targeting Bengali speakers. This mechanism proactively spots, flags, and manages adverse content, bolstering user security and fostering healthier online engagements. Emad et al. [175] present an innovative AI-based moderation system for Persian social media platforms, adept at auto-identifying and filtering out offensive materials. Such tools not only pave the way for a more secure online milieu, along with considerably cut down manual content moderation efforts. Lastly, Bharadwaj et. al.[176] by incorporating machine learning algorithms in social media platforms can actively monitor and flag potentially harmful content, providing a safer online environment for users. By scaling these algorithms, even large platforms with billions of posts can effectively counteract cyberbullying, prioritizing user safety and mental well-being.

6 Discussion

In-depth study of the review papers, exploring sentiment analysis methods on social media, which is our primary focus on the intricate landscape of negative sentiment. This encompasses hate speech, negative opinion, profanity, and targeted insults, with a specific emphasis on their methodologies and technologies tailored to deciphering the complexities of human sentiment in the online domain. Figure 1 illustrates various models used for negative sentiment analysis in social media, incorporating both classical and deep learning-based machine learning algorithms. Two predominant methodological paradigms surfaced in our review: classical machine learning and deep learning. Classical approaches, encompassing machine learning and lexicon-based methods, leverage statistical techniques and sentiment dictionaries. The classical methods like Naive Bayes and Support Vector Machines (SVM) are explored however, these methods may struggle with the evolving nuances, slang, or code-mixed languages prevalent in negative sentiment expressions on social media. On the other hand, deep learning models, represented by architectures like LSTM, GRU, CNN and Transformer models, excel in capturing intricate patterns in vast datasets. This adaptive capability allows for a nuanced interpretation of negative sentiment, particularly vital in deciphering the complexities of online communication characterized by evolving language patterns, symbols, emojis, memes, and new slang. Despite the advantages, the success of any model, whether classical or deep learning, hinges on the quality and relevance of the dataset on which the model has been trained. The rapid evolution of online language, necessitates consistent updating and retraining of the models in order to obtain appropriate outcome and relevance in the context of negative sentiment expressions. Ensuring unbiased, fair, and transparent analysis, especially when dealing with sensitive content like hate speech or targeted insults, the crucial consideration in all research endeavors focusing on negative sentiment in social media data.

7 Conclusion

In conclusion, this comprehensive review paper systematically surveys contemporary techniques for analyzing diverse forms of negative sentiment in social media data, spanning hate speech, profanity, negative opinion, and targeted insults. Our exploration traverses crucial aspects of sentiment analysis, including data collection, pre-processing, feature extraction, classical machine learning algorithms, deep learning algorithms, and evaluation metrics. We scrutinize various data pre-processing techniques such as text normalization, stop-word removal, and stemming, aiming to provide a robust understanding of the methods employed in handling unstructured social media data. Additionally, our review encompasses several feature extraction methods, including bag-of-words, n-grams, and word embeddings, shedding light on the diverse strategies employed to represent textual information. A vital aspect of our examination involves studying the strengths of each methodology as contributions to the field. Classical machine learning approaches offer interpretability and computational efficiency, while deep learning models excel in capturing intricate patterns, providing nuanced interpretations of sentiment. Simultaneously, we identify inherent weaknesses, such as the adaptability challenges faced by classical methods and the resource-intensive nature of deep learning models. Beyond the academic realm, the real-life implementation of negative sentiment analysis methodologies are equally studied in the paper. The comparative analysis of classical machine learning and deep learning based methods highlight their unique strengths, and challenges in the context of negative sentiment analysis. Furthermore, we critically review the evaluation metrics pivotal for assessing sentiment analysis model performance, including accuracy, precision, recall, F1-score, and ROC score. This evaluation framework serves as a guide for researchers and practitioners in selecting appropriate metrics tailored to their specific objectives.

Acknowledgements The authors express deep gratitude to Abhijit Mitra and Anuska Roy for their significant contributions to the research. Abhijit Mitra provided valuable technical assistance and insightful feedback in the writing process, while Anuska Roy played a crucial role in enhancing the rigor of data analyses. The authors appreciate their dedication and unwavering support, acknowledging that their expertise was instrumental in the successful completion of the work.

Data Availability Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflicts of interest The authors declare that there is no conflicts of interest regarding the publication of this paper.

References

1. Badjatiya P, Gupta S, Gupta M, Varma V (2017) Deep learning for hate speech detection in tweets, pp 759–760
2. Sun F, et al. (2019) Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer, pp 1441–1450
3. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780
4. Albawi S, Mohammed TA, Al-Zawi S (2017) Understanding of a convolutional neural network, pp 1–6 (Ieee)
5. Tifrea A, Bécigneul G, Ganea O-E (2018) Poincar\'e glove: Hyperbolic word embeddings. arXiv preprint arXiv:1810.06546
6. Selva Birunda S, Kanniga Devi R (2020) A review on word embedding techniques for text classification. Innovative data communication technologies and application: Proceedings of ICIDCA pp 267–281
7. Jahan MS, Oussalah M (2021) A systematic review of hate speech automatic detection using natural language processing. arXiv preprint arXiv:2106.00742
8. Zhou Y, Yang Y, Liu H, Liu X, Savage N (2020) Deep learning based fusion approach for hate speech detection. IEEE Access 8:128923–128929
9. Al-Hassan A, Al-Dossari H (2021) Detection of hate speech in arabic tweets using deep learning. Multimedia Systems pp 1–12
10. Kapil P, Ekbal A (2020) A deep neural network based multi-task learning approach to hate speech detection. Knowl-Based Syst 210:106458

11. Shruthi P, KM AK (2020) Novel approach for generating hybrid features set to effectively identify hate speech. Intel Artif 23:97–111
12. Kumar A, Abirami S, Trueman TE, Cambria E (2021) Comment toxicity detection via a multichannel convolutional bidirectional gated recurrent unit. Neurocomputing 441:272–278
13. Nikolov A, Radivchev V (2019) Nikolov-radivchev at semeval-2019 task 6: Offensive tweet classification with bert and ensembles, pp 691–695
14. Ranasinghe T, Zampieri M, Hettiarachchi H (2019) Brums at hasoc 2019: Deep learning models for multilingual hate speech and offensive language identification, pp 199–207
15. Saleh Alatawi H, Maatog Alhothali A, Mustafa Moria K (2020) Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. arXiv e-prints arXiv–2010
16. Dowlagar S, Mamidi R (2021) Hasocone@ fire-hasoc2020: Using bert and multilingual bert models for hate speech detection. arXiv preprint arXiv:2101.09007
17. Velankar A, Patil H, Gore A, Salunke S, Joshi R (2022) L3cube-mahahate: A tweet-based marathi hate speech detection dataset and bert models. arXiv preprint arXiv:2203.13778
18. Joshi R (2022) L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources. arXiv preprint arXiv:2202.01159
19. Kannan RR, Rajalakshmi R, Kumar L (2021) Indicbert based approach for sentiment analysis on code-mixed tamil tweets
20. Muller B, Anastasopoulos A, Sagot B, Seddah D (2020) When being unseen from mbert is just the beginning: Handling new languages with multilingual language models. arXiv preprint arXiv:2010.12858
21. Ziehe S, Pannach F, Krishnan A (2021) Gcdh@ lt-edi-eacl2021: Xlm-roberta for hope speech detection in english, malayalam, and tamil, pp 132–135
22. Polignano M, Basile P, De Gemmis M, Semeraro G, Basile V (2019) Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets, vol 2481, pp 1–6 (CEUR)
23. Alami H, El Alaoui SO, Benlahbib A, En-nahnahi N (2020) Lisac fsdm-usmba team at semeval-2020 task 12: Overcoming arabert's pretrain-finetune discrepancy for arabic offensive language identification, pp 2080–2085
24. Sai S, Sharma Y (2020) Siva@ hasoc-dravidian-codemix-fire-2020: Multilingual offensive speech detection in code-mixed and romanized text, pp 336–343
25. Wang S, Liu J, Ouyang X, Sun Y (2020) Galileo at semeval-2020 task 12: Multi-lingual learning for offensive language identification using pre-trained language models. arXiv preprint arXiv:2010.03542
26. Antoun W, Baly F, Hajj H (2020) Arabert: Transformer-based model for arabic language understanding. arXiv preprint arXiv:2003.00104
27. Kuratov Y, Arkhipov M (2019) Adaptation of deep bidirectional multilingual transformers for russian language. arXiv preprint arXiv:1905.07213
28. de Vries W, et al. (2019) Bertje: A dutch bert model. arXiv preprint arXiv:1912.09582
29. Araci DF, Genc Z (2019) Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063
30. Martin L, et al (2019) Camembert: a tasty french language model. arXiv preprint arXiv:1911.03894
31. Le H, et al. (2019) Flaubert: Unsupervised language model pre-training for french. arXiv preprint arXiv:1912.05372
32. Souza F, Nogueira R, Lotufo R (2019) Portuguese named entity recognition using bert-crf. arXiv preprint arXiv:1909.10649
33. Nguyen DQ, Vu T, Nguyen AT (2020) Bertweet: A pre-trained language model for english tweets. arXiv preprint arXiv:2005.10200
34. Kumar R, Reganti AN, Bhatia A, Maheshwari T (2018) Aggression-annotated corpus of hindi-english code-mixed data. arXiv preprint arXiv:1803.09402
35. Ravi K Ravi V (2016) Sentiment classification of hinglish text, pp 641–645 (IEEE)
36. Kamble S, Joshi A (2018) Hate speech detection from code-mixed hindi-english tweets using deep learning models. arXiv preprint arXiv:1811.05145
37. Mathur P, Sawhney R, Ayyar M, Shah R (2018) Did you offend me? classification of offensive tweets in hinglish language, pp 138–148
38. Mathur P, Shah R, Sawhney R Mahata D (2018) Detecting offensive tweets in hindi-english code-switched language, pp 18–26
39. Kshirsagar R, Cukuvac T, McKeown K, McGregor S (2018) Predictive embeddings for hate speech detection on twitter. arXiv preprint arXiv:1809.10644
40. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation, pp 1532–1543
41. Gambäck B, Sikdar UK (2017) Using convolutional neural networks to classify hate-speech, pp 85–90

42. Quoc Tran K et al (2023) Vietnamese hate and offensive detection using phobert-cnn and social media streaming data. Neural Comput Appl 35:573–594
43. Mazari AC, Boudoukhani N, Djeffal A (2023) Bert-based ensemble learning for multi-aspect hate speech detection. Cluster Computing pp 1–15
44. Faris H, Aljarah I, Habib M, Castillo PA (2020) Hate speech detection using word embedding and deep learning in the arabic language context, pp 453–460
45. Elzayady H, Mohamed MS, Badran KM, Salama GI (2023) A hybrid approach based on personality traits for hate speech detection in arabic social media. Int J Electric Comput Eng 13:1979
46. Rizos G, Hemker K, Schuller B (2019) Augment to prevent: short-text data augmentation in deep learning for hate-speech classification, pp 991–1000
47. Chopra S, Sawhney R, Mathur P, Shah RR (2020) Hindi-english hate speech detection: Author profiling, debiasing, and practical perspectives, vol 34, pp 386–393
48. Gupta V, Sehra V, Vardhan YR, et al. (2021) Hindi-english code mixed hate speech detection using character level embeddings, pp 1112–1118 (IEEE)
49. Ali R, Farooq U, Arshad U, Shahzad W, Beg MO (2022) Hate speech detection on twitter using transfer learning. Comput Speech Lang 74:101365
50. Kovács G, Alonso P, Saini R (2021) Challenges of hate speech detection in social media. SN Comput Sci 2:1–15
51. Yuan L, Wang T, Ferraro G, Suominen H, Rizoiu M-A (2023) Transfer learning for hate speech detection in social media. J Computat Soc Sci 1–21
52. Nagar S, Barbhuiya FA, Dey K (2023) Towards more robust hate speech detection: using social context and user data. Soc Netw Anal Min 13:47
53. Duwairi R, Hayajneh A, Quwaider M (2021) A deep learning framework for automatic detection of hate speech embedded in arabic tweets. Arab J Sci Eng 46:4001–4014
54. Bilal M, Khan A, Jan S, Musa S, Ali S (2023) Roman urdu hate speech detection using transformer-based model for cyber security applications. Sensors 23:3909
55. Jahan MS, Oussalah M (2023) A systematic review of hate speech automatic detection using natural language processing. Neurocomputing 126232
56. Saleh H, Alhothali A, Moria K (2023) Detection of hate speech using bert and hate speech word embedding with deep model. Appl Artif Intell 37:2166719
57. Jafri FA, et al. (2023) Uncovering political hate speech during indian election campaign: A new low-resource dataset and baselines. arXiv preprint arXiv:2306.14764
58. Del Vigna12 F, Cimino23 A, Dell'Orletta F, Petrocchi M, Tesconi M (2017) Hate me, hate me not: Hate speech detection on facebook, pp 86–95
59. Alfina I, Mulia R, Fanany MI, Ekanata Y (2017) Hate speech detection in the indonesian language: A dataset and preliminary study, pp 233–238 (IEEE)
60. Sajjad M, Zulifqar F, Khan MUG, Azeem M (2019) Hate speech detection using fusion approach, pp 251–255 (IEEE)
61. Abro S, et al. (2020) Automatic hate speech detection using machine learning: A comparative study. Int J Adv Comput Sci Appl 11
62. Briliani A, Irawan B, Setianingsih C (2019) Hate speech detection in indonesian language on instagram comment section using k-nearest neighbor classification method, pp 98–104 (IEEE)
63. Mujadia V, Mishra P, Sharma DM (2019) Iiit-hyderabad at hasoc 2019: Hate speech detection, pp 271–278
64. Mullah NS, Zainon WMNW (2021) Advances in machine learning algorithms for hate speech detection in social media: a review. IEEE Access 9:88364–88376
65. Istaiteh O, Al-Omoush R, Tedmori S (2020) Racist and sexist hate speech detection: Literature review, pp 95–99 (IEEE)
66. Kwok I, Wang Y (2013) Locate the hate: Detecting tweets against blacks
67. Waseem Z, Hovy D (2016) Hateful symbols or hateful people? predictive features for hate speech detection on twitter, pp 88–93
68. Frenda S, Ghanem B, Montes-y Gómez M, Rosso P (2019) Online hate speech against women: Automatic identification of misogyny and sexism on twitter. J Intell Fuzzy Syst 36:4743–4752
69. Saha P, Mathew B, Goyal P, Mukherjee A (2018) Hateminers: Detecting hate speech against women. arXiv preprint arXiv:1812.06700
70. Fersini E, Nozza D, Rosso P, et al. (2018) Overview of the evalita 2018 task on automatic misogyny identification (ami) (Accademia University Press, 2018)
71. Andreas J, Choi E, Lazaridou A (2016) Proceedings of the naacl student research workshop
72. Vidgen B, Yasseri T (2020) Detecting weak and strong islamophobic hate speech on social media. J Inform Technol Politics 17:66–78

73. Aljarah I et al (2021) Intelligent detection of hate speech in arabic social network: A machine learning approach. J Inf Sci 47:483–501
74. Nugroho K, Noersasongko E, Fanani AZ, Basuki RS, et al. (2019) Improving random forest method to detect hatespeech and offensive word, pp 514–518 (IEEE)
75. Burnap P, Williams ML (2016) Us and them: identifying cyber hate on twitter across multiple protected characteristics. EPJ Data Sci 5:1–15
76. 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 access 6:13825–13835
77. Magdy W, Darwish K, Weber I (2015) # failedrevolutions: Using twitter to study the antecedents of isis support. arXiv preprint arXiv:1503.02401
78. Kaati L, Omer E, Prucha N, Shrestha A (2015) Detecting multipliers of jihadism on twitter, pp 954–960 (IEEE)
79. Abozinadah EA (2016) Improved micro-blog classification for detecting abusive arabic twitter accounts. Int J Data Mining Knowled Manag Process (IJDKP) 6
80. Mubarak H, Darwish K, Magdy W (2017) Abusive language detection on arabic social media, pp 52–56
81. Alakrot A, Murray L, Nikolov NS (2018) Towards accurate detection of offensive language in online communication in arabic. Procedia Comput Sci 142:315–320
82. Abdelfatah KE, Terejanu G, Alhelbawy AA, et al. (2017) Unsupervised detection of violent content in arabic social media. Comput Sci Inform Technol (CS IT) 7
83. Haidar B, Chamoun M, Serhrouchni A (2017) A multilingual system for cyberbullying detection: Arabic content detection using machine learning. Adv Sci Technol Eng Syst J 2:275–284
84. Jaki S, De Smedt T (2019) Right-wing german hate speech on twitter: Analysis and automatic detection. arXiv preprint arXiv:1910.07518
85. Özel SA, Saraç E, Akdemir S, Aksu H (2017) Detection of cyberbullying on social media messages in turkish, pp 366–370 (IEEE)
86. Fernandez M, Alani H (2018) Contextual semantics for radicalisation detection on twitter
87. Martins R, Gomes M, Almeida JJ, Novais P, Henriques P (2018) Hate speech classification in social media using emotional analysis, pp 61–66 (IEEE)
88. Abozinadah EA, Mbaziira AV, Jones J (2015) Detection of abusive accounts with arabic tweets. Int J Knowl Eng -IACSIT 1:113–119
89. Wiegand M, Ruppenhofer J, Schmidt A, Greenberg C (2018) Inducing a lexicon of abusive words–a feature-based approach, pp 1046–1056
90. Warner W, Hirschberg J (2012) Detecting hate speech on the world wide web, pp 19–26
91. Davidson T, Warmsley D, Macy M (2017) Weber I. Automated hate speech detection and the problem of offensive language vol 11, pp 512–515
92. Tabinda Kokab S, Asghar S, Naz S (2022) Transformer-based deep learning models for the sentiment analysis of social media data. Array 14:100157. https://www.sciencedirect.com/science/article/pii/S2590005622000224
93. Alwakid G, et al. (2022) Muldasa: Multifactor lexical sentiment analysis of social-media content in nonstandard arabic social media. Appl Sci 12. https://www.mdpi.com/2076-3417/12/8/3806
94. Chandrasekaran G, Antoanela N, Andrei G, Monica C, Hemanth J (2022) Visual sentiment analysis using deep learning models with social media data. Appl Sci 12. https://www.mdpi.com/2076-3417/12/3/1030
95. Hassan SZ, et al. (2022) Visual sentiment analysis from disaster images in social media. Sensors 22. https://www.mdpi.com/1424-8220/22/10/3628
96. Sufi FK, Khalil I (2022) Automated disaster monitoring from social media posts using ai-based location intelligence and sentiment analysis. IEEE Trans Computat Soc Syst 1–11
97. Vatambeti R, Mantena SV, Kiran K, Manohar M, Manjunath C (2023) Twitter sentiment analysis on online food services based on elephant herd optimization with hybrid deep learning technique. Cluster Computing 1–17
98. Bello A, Ng S-C, Leung M-F (2023) A bert framework to sentiment analysis of tweets. Sensors 23:506
99. Qian C, et al. (2022) Understanding public opinions on social media for financial sentiment analysis using ai-based techniques. Informa Process Manag 59:103098 . https://www.sciencedirect.com/science/article/pii/S0306457322001996
100. Zhigang Jin XZ, Manyue Tao Hu Y (2022) Social media sentiment analysis based on dependency graph and co-occurrence graph
101. Idan L (2020) Feigenbaum J (2020) Show me your friends, and i will tell you whom you vote for: Predicting voting behavior in social networks, ASONAM '19, 816–824 (Association for Computing Machinery, New York, NY. USA. https://doi.org/10.1145/3341161.3343676

102. Nemes L, Kiss A (2021) Social media sentiment analysis based on covid-19. J Inform Telecommun 5:1–15. https://doi.org/10.1080/24751839.2020.1790793
103. Joyce D (2017) Sentiment analysis of tweets for the 2016 us presidential election
104. Hassan A, Abbasi A, Zeng D (2013) Twitter sentiment analysis: A bootstrap ensemble framework, pp 357–364 (IEEE)
105. Zhang L, Ghosh R, Dekhil M, Hsu M, Liu B (2011) Combining lexicon-based and learning-based methods for twitter sentiment analysis. HP Laboratories, Technical Report HPL 89:1–8
106. Davidov D, Tsur O, Rappoport A (2010) Enhanced sentiment learning using twitter hashtags and smileys, pp 241–249
107. Mohammad SM, Kiritchenko S, Zhu X (2013) Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. arXiv preprint arXiv:1308.6242
108. Read J (2005) Using emoticons to reduce dependency in machine learning techniques for sentiment classification, pp 43–48
109. Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. CS224N project report, Stanford 1:2009
110. Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau RJ (2011) Sentiment analysis of twitter data, pp 30–38
111. Speriosu M, Sudan N, Upadhyay S, Baldridge J (2011) Twitter polarity classification with label propagation over lexical links and the follower graph, pp 53–63
112. Saif H, He Y, Alani H (2012) Semantic sentiment analysis of twitter, pp 508–524 (Springer)
113. Lin J, Kolcz A (2012) Large-scale machine learning at twitter, pp 793–804
114. Clark S, Wicentwoski R (2013) Swatcs: Combining simple classifiers with estimated accuracy, pp 425–429
115. Yuliyanti S, Djatna T, Sukoco H (2017) Sentiment mining of community development program evaluation based on social media. TELKOMNIKA (Telecommunication Computing Electronics and Control) 15(24):1858–1864
116. Mansour S (2018) Social media analysis of user's responses to terrorism using sentiment analysis and text mining
117. Saragih MH, Girsang AS (2017) Sentiment analysis of customer engagement on social media in transport online
118. Hassan AU, Hussain J, Hussain M, Sadiq M, Lee S (2017) Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. In: 2017 international conference on information and communication technology convergence (ICTC). IEEE, pp 138–140
119. Ikoro V, Sharmina M, Malik K, Batista-Navarro R (2018) Analyzing sentiments expressed on Twitter by UK energy company consumers. In: 2018 Fifth international conference on social networks analysis, management and security (SNAMS). IEEE, pp 95–98
120. Hao J, Dai H (2016) Social media content and sentiment analysis on consumer security breaches. J Financ Crime 23(4):855–869
121. Shayaa S, Wai PS, Chung YW, Sulaiman A, Jaafar NI, Zakaria SB (2017) Social media sentiment analysis on employment in Malaysia. In: the Proceedings of 8th Global Business and Finance Research Conference, Taipei, Taiwan
122. Ali K, Dong H, Bouguettaya A, Erradi A, Hadjidj R (2017) Sentiment analysis as a service: a social media based sentiment analysis framework. In: 2017 IEEE international conference on web services (ICWS). IEEE, pp 660–667
123. Arias F, Zambrano Núñez M, Guerra-Adames A, Tejedor-Flores N, Vargas-Lombardo M (2022) Sentiment analysis of public social media as a tool for health-related topics. IEEE Access 10:74850–74872
124. Al-Ajlan M, Ykhlef M (2023) Firefly-cddl: A firefly-based algorithm for cyberbullying detection based on deep learning. CMC-Computers Materials Continua 75:19–34
125. Chaudhari A, Davda P, Dand M, Dholay S (2021) Profanity detection and removal in videos using machine learning, pp 572–576 (IEEE)
126. Bhowmick RS, Ganguli I, Paul J, Sil J (2021) A multimodal deep framework for derogatory social media post identification of a recognized person. Trans Asian Low-Resource Lang Inform Process 21:1–19
127. Kumari K, Singh JP (2019) Ai ml nit patna at hasoc 2019: Deep learning approach for identification of abusive content. FIRE (working notes) 2517:328–335
128. Dadvar M, Eckert K (2018) Cyberbullying detection in social networks using deep learning based models; a reproducibility study. arXiv preprint arXiv:1812.08046
129. Chinagundi B, Singh M, Ghosal T, Rana PS, Kohli GS (2021) Classification of hate, offensive and profane content from tweets using an ensemble of deep contextualized and domain specific representations
130. Soykan L, Karsak C, Elkahlout ID, Aytan B (2022) A comparison of machine learning techniques for turkish profanity detection, pp 16–24

131. Galinato V, Amores L, Magsino GB, Sumawang DR (2023) Context-based profanity detection and censorship using bidirectional encoder representations from transformers. Available at SSRN 4341604
132. Teh PL, Cheng C-B (2020) Profanity and hate speech detection. Int J Inf Manage Sci 31:227–246
133. Ratadiya P, Mishra D (2019) An attention ensemble based approach for multilabel profanity detection, pp 544–550 (IEEE)
134. Park JH, Fung P (2017) One-step and two-step classification for abusive language detection on twitter. arXiv preprint arXiv:1706.01206
135. Ba Wazir AS et al (2021) Design and implementation of fast spoken foul language recognition with different end-to-end deep neural network architectures. Sensors 21:710
136. Kim C-G, Hwang Y-J, Kamyod C (2022) A study of profanity effect in sentiment analysis on natural language processing using ann. J Web Eng 751–766
137. Dandeniya D (2023) Profanity filtering in speech contents using deep learning algorithms. Ph.D. thesis
138. Yang H, Lin C-J (2020) Tocp: A dataset for chinese profanity processing, pp 6–12
139. Woo J, Park SH, Kim HK (2022) Profane or not: Improving korean profane detection using deep learning. KSII Trans Internet Inform Syst (TIIS) 16:305–318
140. Sazzed S (2021) Bengsentilex and bengswearlex: creating lexicons for sentiment analysis and profanity detection in low-resource bengali language. PeerJ Comput Sci 7:e681
141. Al-Hashedi M, Soon L-K, Goh H-N (2019) Cyberbullying detection using deep learning and word embeddings: An empirical study, pp 17–21
142. Malik P, Aggrawal A, Vishwakarma DK (2021) Toxic speech detection using traditional machine learning models and bert and fasttext embedding with deep neural networks, pp 1254–1259 (IEEE)
143. Marwa T, Salima O, Souham M (2018) Deep learning for online harassment detection in tweets, pp 1–5 (IEEE)
144. Perera A, Fernando P (2021) Accurate cyberbullying detection and prevention on social media. Procedia Comput Sci 181:605–611
145. Gottipati S, et al. (2020) Leveraging profanity for insincere content detection-a neural network approach, pp 0041–0047 (IEEE)
146. Sood SO, Antin J, Churchill E (2012) Using crowdsourcing to improve profanity detection
147. Chin H, Kim J, Kim Y, Shin J, Yi MY (2018) Explicit content detection in music lyrics using machine learning, pp 517–521 (IEEE)
148. Nobata C, Tetreault J, Thomas A, Mehdad Y, Chang Y (2016) Abusive language detection in online user content, pp 145–153
149. Baby A, Jose J, Raj A (2023) Psychosomatic study of criminal inclinations with profanity on social media: Twitter, pp 611–627 (Springer)
150. Raj C, Agarwal A, Bharathy G, Narayan B, Prasad M (2021) Cyberbullying detection: hybrid models based on machine learning and natural language processing techniques. Electronics 10:2810
151. Kalaivani A, Thenmozhi D (2020) Ssn_nlp_mlrg at semeval-2020 task 12: Offensive language identification in english, danish, greek using bert and machine learning approach, pp 2161–2170
152. Shanmugavadivel K et al (2022) Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data. Sci Rep 12:21557
153. Alotaibi M, Razaque Alotaibi B A (2021) A multichannel deep learning framework for cyberbullying detection on social media. Electronics 10:2664
154. Viñas Redondo B (2020) Identifying and categorizing offensive language in tweets using Machine Learning. B.S. thesis, Universitat Politècnica de Catalunya
155. Sodhi R, Pant K, Mamidi R (2021) Jibes delights: A dataset of targeted insults and compliments to tackle online abuse, pp 132–139
156. Zhang Z, Robinson D Tepper J (2018) Detecting hate speech on twitter using a convolution-gru based deep neural network, pp 745–760 (Springer)
157. Alsafari S, Sadaoui S, Mouhoub M (2020) Hate and offensive speech detection on arabic social media. Online Social Networks and Media 19:100096
158. Chakravarthi BR, Jagadeeshan MB, Palanikumar V, Priyadharshini R (2023) Offensive language identification in dravidian languages using mpnet and cnn. Intern J Inform Manag Data Insights 3:100151
159. Cojocaru A, Paraschiv A Dascalu M (2022) News-ro-offense-a romanian offensive language dataset and baseline models centered on news article comments, pp 65–72
160. Abbasi A, Javed AR, Iqbal F, Kryvinska N, Jalil Z (2022) Deep learning for religious and continent-based toxic content detection and classification. Sci Rep 12:17478
161. Omar A, Mahmoud TM, Abd-El-Hafeez T, Mahfouz A (2021) Multi-label arabic text classification in online social networks. Inf Syst 100:101785
162. Parikh A, Desai H Bisht AS (2019) Da master at hasoc 2019: Identification of hate speech using machine learning and deep learning approaches for social media post, pp 315–319

163. Sharifirad S (2019) Nlp and machine learning techniques to detect online harassment on social networking platforms
164. Bharathi B, et al. (2021) Ssncse_nlp@ dravidianlangtech-eacl2021: Offensive language identification on multilingual code mixing text, pp 313–318
165. Thenmozhi D, Sharavanan S, Chandrabose A et al. (2019) Ssn_nlp at semeval-2019 task 6: Offensive language identification in social media using traditional and deep machine learning approaches pp 739–744
166. Thara S, Poornachandran P (2022) Social media text analytics of malayalam-english code-mixed using deep learning. J Big Data 9:45
167. Sutejo TL, Lestari DP (2018) Indonesia hate speech detection using deep learning, pp 39–43 (IEEE)
168. Chakravarthi BR et al (2022) Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. Lang Resour Eval 56:765–806
169. Zampieri M, et al. (2019) Predicting the type and target of offensive posts in social media. arXiv preprint arXiv:1902.09666
170. Yasaswini K, et al. (2021) Iiitt@ dravidianlangtech-eacl2021: Transfer learning for offensive language detection in dravidian languages, pp 187–194
171. Reynolds, K., Kontostathis, A. Edwards, L (2011) Using machine learning to detect cyberbullying, vol 2, pp 241–244 (IEEE)
172. Fahim M, Gokhale SS (2021) Detecting offensive content on twitter during proud boys riots, pp 1582–1587 (IEEE)
173. Kurniawan S, Budi I (2020) Indonesian tweets hate speech target classification using machine learning, pp 1–5, (IEEE)
174. Akhter A, Acharjee UK, Talukder MA, Islam MM, Uddin MA (2023) A robust hybrid machine learning model for bengali cyber bullying detection in social media. Natural Lang Process J 4:100027
175. Kebriaei E, et al. (2023) Persian offensive language detection. Machine Learning pp 1–21
176. Bharadwaj VY, et al. (2023) Automated cyberbullying activity detection using machine learning algorithm, vol 430, 01039 (EDP Sciences)
177. Sanoussi MSA, et al. (2022) Detection of hate speech texts using machine learning algorithm, pp 0266–0273 (IEEE)
178. Nayel HA, Shashirekha H (2019) Deep at hasoc2019: A machine learning framework for hate speech and offensive language detection, pp 336–343

Authors and Affiliations

Jayanta Paul[1] · **Ahel Das Chatterjee**[1] · **Devtanu Misra**[1] · **Sounak Majumder**[1] · **Sayak Rana**[1] · **Malay Gain**[1] · **Anish De**[1] · **Siddhartha Mallick**[1] · **Jaya Sil**[1]

Ahel Das Chatterjee
aheldc@gmail.com

Devtanu Misra
devtanumisra@gmail.com

Sounak Majumder
sounakmajumder472@gmail.com

Sayak Rana
sayak.rana2001@gmail.com

Malay Gain
malaygain10@gmail.com

Anish De
anishde85@gmail.com

Jaya Sil
js@cs.iiests.ac.in

1 CST, Indian Institute of Engineering Science and Technology, Shibpur, Botanic Garden, Howrah 711103, West Bengal, India