Alexandria Engineering Journal 80 (2023) 110–121

Contents lists available at ScienceDirect

Alexandria Engineering Journal

journal homepage: www.elsevier.com/locate/aej

A survey on hate speech detection and sentiment analysis using machine learning and deep learning models

Malliga Subramanian [a], Veerappampalayam Easwaramoorthy Sathiskumar [b], G. Deepalakshmi [a], Jaehyuk Cho [b,*], G. Manikandan [c]

[a] Department of Computer Science and Engineering, Kongu Engineering College, Perundurai, Erode, India
[b] Department of Software Engineering, Jeonbuk National University, Jeongu-si, Republic of Korea
[c] School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India

ARTICLE INFO

Keywords:
Hate speech detection
Sentiment analysis
Machine learning
Deep learning
Inclusive online

ABSTRACT

In today's digital era, the rise of hate speech has emerged as a critical concern, driven by the rapid information-sharing capabilities of social media platforms and online communities. As the internet expands, the proliferation of harmful content, including hate speech, presents considerable obstacles in ensuring a secure and inclusive online environment. In response to this challenge, researchers have embraced machine learning and deep learning methods to create automated systems that can effectively detect hate speech and conduct sentiment analysis, offering potential solutions to address this pressing issue. This survey article provides a comprehensive overview of recent advancements in hate speech detection and sentiment analysis using machine learning and deep learning models. We present an in-depth analysis of various methodologies and datasets employed in this domain. Additionally, we explore the unique challenges faced by these models in accurately identifying and classifying hate speech and sentiment in online text. Finally, we outline areas where more study is needed and suggest potential new avenues for exploration in the field of hate speech identification and sentiment analysis. Using the results of this survey, we hope to encourage the development of more effective machine learning and deep learning-based solutions to curb hate speech and promote a more inclusive online environment.

1. Introduction

The advancement in internet technology and tremendous growth of users in online activities, and social media networks leads to the generation of an unprecedented volume of data. The data that users generate through their online activities, whether it be in the form of text, images, music, videos, log files, reviews, etc., is typically generated from a variety of sources, voluminous and includes structured as well as unstructured data. Performing and analysing these types of unstructured and structured data has a greater impact on the big data field [1]. Such type of data can be analysed for decision making using machine learning, data mining, web mining and text mining techniques. Also, since these types of data can be voluminous and extracting the patterns from this data is quite a difficult process. And, further, microblogging services like Twitter, YouTube, Instagram, Facebook, Snapchat, WhatsApp, LinkedIn, blogs, Wikis etc., support a variety of data formats with/without the proper grammatical rules and also short texts which are written without concerning the grammars [2]. Fig. 1 shows the percentage of users on social network platforms. From these platforms the amount of information (opinions) [3,4] which is shared by the users can be used for analysing the opinions about the products, political movements, financial and political forecasting, monitoring the company strategies, marketing analysis, disseminating news, crime forecasting, product preferences, tracing the terrorist activities, e-health and e-tourism, monitoring reputations, detecting the hate speech in the public forms etc. To find meaningful information from the text (corpus) or data coming from public forums, Natural Language Processing (NLP) techniques is used [5].

The advent of social media and online forums has revolutionized the way people communicate and express their opinions. However, this newfound freedom of expression has also given rise to the proliferation of hate speech, cyberbullying, and offensive content, which can have severe implications on individuals and society as a whole. Identifying and curbing such harmful contents has become a critical task for maintaining a respectful and safe online space. For instance, Modha et. al. [6] dealt with the identification of the aggression types of texts in the

* Corresponding author.
E-mail address: chojh@jbnu.ac.kr (J. Cho).

https://doi.org/10.1016/j.aej.2023.08.038
Received 13 May 2023; Received in revised form 28 July 2023; Accepted 12 August 2023
Available online 24 August 2023

online platforms and divided the texts into aggressive and non–aggressive. Fig. 2 depicts the percentage of hate speech texts posted in Instagram during the four quarters of the years 2020 and 2021. Kaur et. al. [7] mentions the concepts of abusive content detection based on four categories of features namely, activity based, user based, context-based, and network-based features. This survey has also mentioned many parameters to identify the abusive contents such as posts per day, age, gender, etc and helps to build the researchers with fundamental concepts and key insight areas including the recent trends and techniques. The relationship between hate speech, aggressiveness and offensive speech is discussed in [8].

Traditional rule-based methods for hate speech detection and sentiment analysis often lack the scalability and adaptability to handle the vast amount of user-generated content on social media platforms. In contrast, machine learning and deep learning techniques have shown promising results in automating the process of identifying hate language and analyzing sentiments expressed in text data. The primary objective of this survey is to present an in-depth analysis of hate speech detection and sentiment analysis techniques, focusing on the application of machine learning and deep learning models. By exploring the challenges faced by the present approaches, this paper aims to provide researchers with insights into the evolving landscape of hate speech detection and sentiment analysis. By investigating a wide range of methodologies, and datasets, this survey seeks to shed light on the advancements made in this crucial field and highlight the challenges that lie ahead. The main contributions of this survey include:

1. Review of Datasets: We present a detailed list of datasets used for hate speech detection and sentiment analysis.
2. Review of machine learning approaches: We begin by exploring the early efforts in hate speech detection and sentiment analysis, which relied on machine learning algorithms such as Support Vector Machines (SVM), Naive Bayes (NB), Decision Trees (DT), and Logistic Regression (LR).
3. Emergence of deep learning models: Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based architectures like BERT and GPT, which have shown remarkable capabilities in text classification and sentiment analysis have also been reviewed.
4. Challenges in hate speech detection: We then delve into the unique challenges faced by hate speech detection models, including the dynamic nature of language, context-dependent interpretations, and the subtleties involved in identifying sarcastic or disguised hate speech.

1.1. 1.1 Review methodology

The main aim of this work is to review the articles that show and describe the importance of text mining and NLP to online social media networks.

1. Therefore, this survey process is started and gathered numerous numbers of papers from the standard academic and research search engines such as ScienceDirect, Springer, IEEE Explore, Francis, Taylor and Google Scholar DBLP. The key terms which are used to perform the search process are:1. “Text classification” and “NLP”, 2. “sentiment analysis” and “NLP”, 3. “Hate speech” and “NLP” 4. “Hate speech” and “Machine Learning”, 5: “Offensive message” and “Deep Learning”, 6. “Online social media” and “hate speech”. 7. “Online Sources” and “Hate speech”. 8. “Sentiment analysis” and “Hate Speech”, 9. “Abusive content” and “Social Networks”. This search strategy is used to collect the initial set of articles which are published in the research platforms recently and further, this search strategy was expanded by identifying the new set of articles that were cited from this initial set of articles.

2. Further, we extracted and refined our survey by exploring more than 100 articles that were published from a decade to till date. Also, we have explored and investigated the importance of hate speech detection through the Statista web platform for knowing the activities, posts, and classification of messages on online platform networks such as Twitter, Facebook, Instagram, etc., In the same way, this survey article covers artificial intelligence techniques, in particular with, deep learning and machine learning approaches for hate speech detection and sentiment analysis.

3. An in-depth investigation of deep learning and machine learning

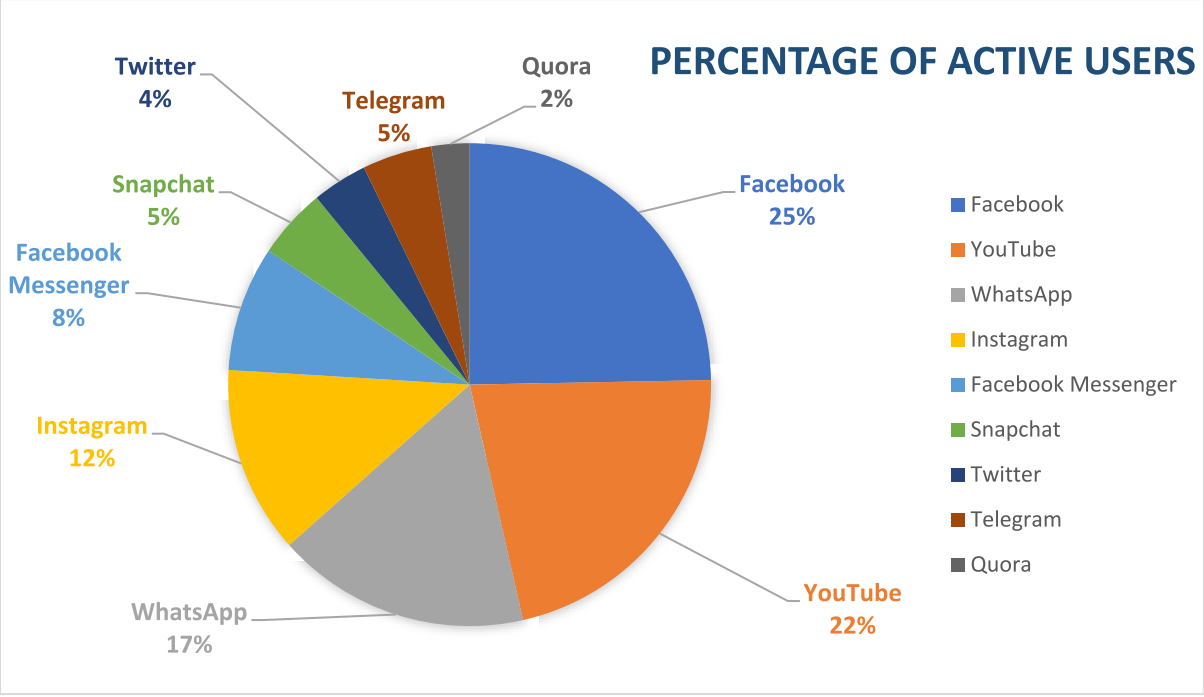

Fig. 1. Active users and their percentage in social networks.

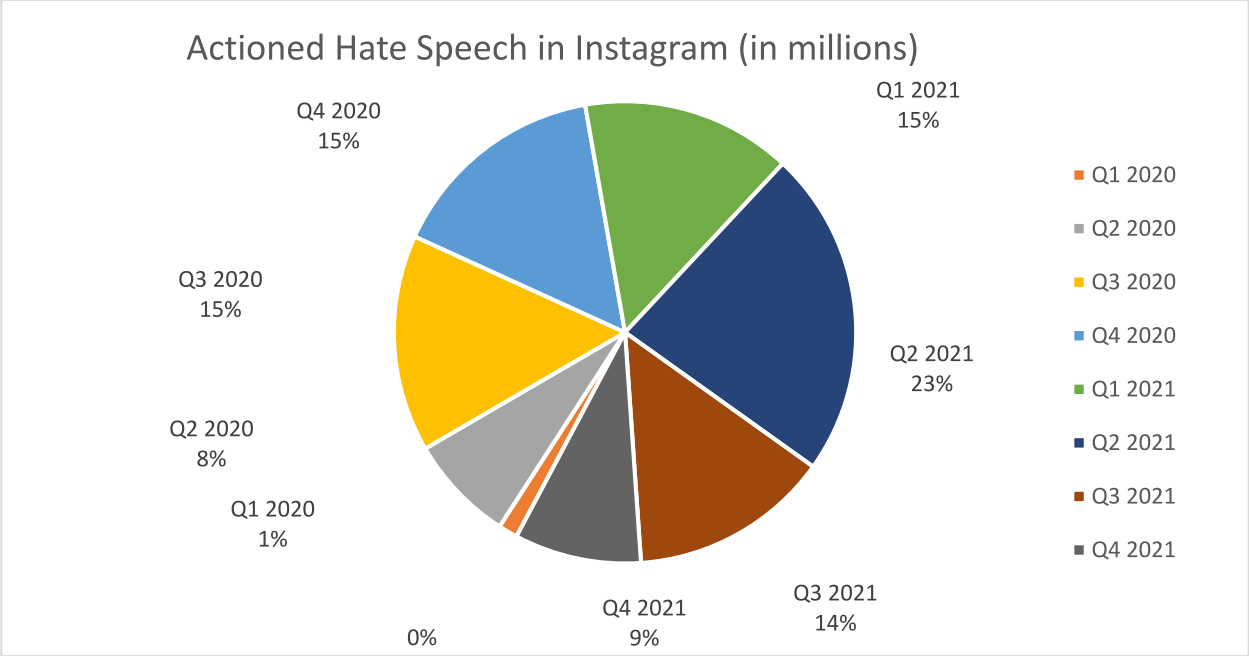

Fig. 2. Actioned Hate Speech on Instagram from 2020 to 2021.

approaches to hate speech detection is carried out by presenting datasets. This study focused on showing the issues present in hate speech detection from the users' posts, blogs, etc.

4. Additionally, this survey covers the importance and need of hate speech detection and sentiment analysis in day-to-day life and subsequently, addressed the need for highly sophisticated machine learning and deep learning approaches to analyse and classify the data from online social media.

The rest of the article is organized as follows: Section 2 presents an overview of data preprocessing and datasets used for hate speech detection and sentiment analysis. This section enumerates the datasets with their class labels and languages. In Section 3, we present the role of machine learning and deep learning algorithms in the present study. In addition, we tabulate the models developed for hope speech detection and sentiment analysis along with the details of the datasets used in those models. The challenges in the objectives of the proposed study are enumerated in Section 4. Finally, we conclude our work and present future research directions of the proposed study in Section 5.

2. Preliminary steps for detecting hate speech in text

2.1. Data acquisition

Comments on the social media websites such as Twitter, Facebook, YouTube, etc., are not always good for the users, in some cases the posts may be rude or hateful words. On social media, offensive remarks might include indiscriminate slang, abusive language and vulgarity. Because of the drastic increase in online resources, data collection is extremely dependent on the type of media used to share the contents and also the data format is important to analyse the data. Twitter, Sina-Wiebo and other microblogging services have made their Application Programming Interface (API) available to extract public data from the sites. Twitter provides a REST API for static data such as user profiles and a Streaming API2 for streaming data such as tweets [1]. Twitter4J API3 [2] is used to extract the streaming tweets. Facebook Graph API4 and Tancent API5 are also made available by Facebook and Sina-Wiebo, respectively. These APIs are also used to collect articles as well as other data from their site for further analysing the data.

2.2. Data pre-processing

Data pre-processing is the first and foremost step and it includes data cleaning, tokenization, stop word removal, normalization etc, Data cleaning processes the links, punctuation marks, hashtags, and numeric characters are all regarded as non-essential in NLP. However, eliminating punctuation and hashtags, for example, may not be the most effective technique to clean up text information. Punctuation marks can be used as alternative emojis to represent the users' feelings, and hashtags typically contain extensive semantic meaning that could be useful for detecting abusive comments. As a result, the pre-processing step has been tested with or without the data cleaning step to see how it is affected by the outcomes.

The process which divides the text data into words and sentences, which are referred to as tokens is called tokenization. These tokens aid in the comprehension of the context or the development of the NLP model. By evaluating the sequence of words, tokenization aids in understanding the context of the text. The comments can be tokenized based on punctuation marks, whitespaces, etc. Stop words like formatting tags, numerals, pronouns, prepositions, conjunctions, and auxiliary verbs can be eliminated from the comments. Text is normalized to lessen its unpredictability and move it closer to a defined standard. As a result, the amount of variation in the data is reduced, and efficiency could be enhanced.

2.3. Datasets for hate speech detection

Before surveying the methods and techniques that can detect hate speech and sentimental comments from social media comments, we carried out a detailed survey on the datasets used by the research communities to develop and evaluate their models. Table 1 presents the details of the datasets about hate speech.

3. Review of Machine learning models for hate speech detection and sentiment analysis

3.1. Machine learning approaches

Hope speech detection and sentiment text classification can be

Table 1
Dataset about the Hate Speech Detection and Sentiment Analysis.

References	Year	Dataset	Dataset Description	Language
[9]	2022	Newspaper articles Data from News Channels	Data from CNN News, FOX News, NPR News.	English
[10]	2022	Unstructured Data from the Social Media comment-Facebook	The comments are classified as Positive, Neutral, Negative	English
[11]	2022	Social media comments-Twitter	Data Set 1 [balanced] - abusive, hateful, normal, or spam.Data Set 2 [Unbalanced]-abusive, hateful, normal, or spam.	English
[12]	2022	Social Media comments	The comments are classified as Positive, Neutral, Negative	English
[13]	2022	TripAdvisor dataset	44,217 multiple ratings for overall service of restaurants, food, price with 17,795 text reviews.	English
[14]	2022	Laptop DatasetRestaurant Dataset	1. Reviews − 1326 990 615 comments2. Reviews − 4131 1535 927 comments	English
[15]	2022	Thomas Davidson, Zeerak Waseem	The first dataset contains 24,783 tagged cases divided into three categories: “Hate,” “Offensive,” and “non-offensive.”. Another has 16,907 labeled cases divided into three categories: “Racism”, “Sexism” and “Neither.”	English
[15]	2022	Sina Weibo Sexism Review (SWSR)	The comments are classified as (i) non-sexism or sexism, (ii) target type, and (iii) sexism category with 8969 comments.	Chinese
[16]	2022	Online reviews of travel websites	Total number of online reviews for five different hotels is 3486, 3846, 4549, 1960, and 2575 and these were grouped into five review datasets.	English
[17]	2022	Online reviews	Four benchmark datasets - SemEval2014 (laptop domain& restaurant), 2015(restaurant), and 2016(restaurant). Datasets contain review sentences with annotated labels from the laptop and restaurant domains	English
[18]	2022	Melancholic sentences from melancholicarticle	The PsychPark website − 2,300 sad sentences from a melancholic essay with the class Depressed moodSuicide, Insomnia, Work and activity, Anxiety, Physical symptom	English
[19]	2022	Social Media comments	4 datasets: Waseem Hovy, Waseem, Davidson, HatEval: The corpus contains data collected from Twitter over the months	English
[5]	2021	Spanish Tweets Dataset	The comments are classified as Positive, Negative, Neutral	Spanish-English
[20]	2021	Offensive Language Identification Dataset (OLID)	This is the dataset from SemEval-2019 Task 6: Identifying and Categorising Offensive Language in social media and classified the texts as Offensive, Non-Offensive and Neutral	English
[21]	2021	Social media comment-Twitter	Three sets of datasets namely Storm Front Dataset, Twitter White Supremacy Dataset and Balanced and Combined Datasets from Twitter and StormFront with hate and Non-hate as labels.	English
[22]	2021	Social media comment-Twitter	The comments are classified as Highly Offensive, Offensive, Neutral, Positive, Highly Positive Messages	Urdu-English
[23]	2021	Social Media comments	The comments are classified as Positive, Negative, Neutral	English
[24]	2021	Unstructured Data from the Twitter	The comments are classified as Positive-True, Negative-False	English
[25]	2021	Hatebase Twitter HatEval Waseem A Waseem BMLMA	Hate/ Offensive/ Neither Women /Immigrants Sexist /Racist Sexist /RacistGender /Sexist /Religion / Disability	English
[26]	2021	Internet Movie Database (IMDB), Amazon	Positive and Negative comments	English
[27]	2021	L-HSAB DKhate,	Hate speech and abusive language, 5846 comments Offensive speech, target, and grade, 3600 comments	Arabic DanishEnglish
[28]	2021	AskFm	10,000 comments to detect cyberbullying	English
[29]	2021	Waseem and Hovyf	Dataset of racist and sexist tweets sampled from Twitter and labeled by a mix of expert annotators and activists.	English
[30]	2021	HaterNet	Collected tweets between February and December 2017 on several random dates. A total of 2 million tweets from Spain were retrieved in the end.	Spanish
[31]	2021	VLSP corpus	A dataset created in the VLSP project contains 10,000 sentences that have been manually segmented by linguists	Vietnamese
[32]	2021	Original ATC ATC	Original ATC −10,528 abusive 19,826 non-abusive commentsOversampled − 19,826 abusive 19,826 non-abusive comments	Turkish
[33]	2021	Devanagari Hindi Offensive Tweet (DHOT)	The survey was taken from 150 Hindi-speaking people and collected	Hindi
[34]	2021	Twitter dataset	Various tweets having less than 280 characters are taken with text, photos, videos and URLs.	English
[7]	2021	Internet Argument Corpus, Hate Speech,Twitter Corpus, Wikipedia Talklabels, Insult Dataset, Toxic Comment Classification	Classified as Not insulting, Insulting, Toxic, Severe toxicObscene, Threat Insult, Identity_hate.	English
[35]	2021	Weibo	38,183,194 microblog entries posted by 2,239,472 users	Chinese-English
[36]	2021	Online Review Dataset	Classified as Positive and Negative	Chinese-English
[37]	2021	RuEthnoHateRuEthnics	5,594 texts,12,052 instances with 2,040 instances are negative 8,697 are neutral 1,315 are positive	Russian
[38]	2021	Wikipedia Talk Labels	100,000 commentsaggressive, neutral, friendly	English

(*continued on next page*)

Table 1 (*continued*)

References	Year	Dataset	Dataset Description	Language
[39]	2021	IMDB NAVER-Harassment-Yellow NAVER-Harassment-Red NAVER-Posneg	Text dataset50,000 movie reviews with binary labels and 31,392 comments	EnglishKorean
[40]	2021	Airline1, Airline 2, Airline 3, Airline 4, Property agent, Homebuilders	Dataset are collected from Tripadvisor.com-airline and productreview.com website which consists of reviews with positive and negative labels	English
[41]	2021	Stream of-consciousness essays (SoCE) and YouTube (YoTB)	Classified as neuroticism (NEU), extroversion (EXT), agreeableness (AGR),openness(OPN), conscientiousness (CON)	English
[42]	2021	Restaurant, Laptop, and Twitter data sets.	Restaurant and laptop Labels: Positive and neutral, negative, Twitter data: celebrities, businesses, and products, with positive, neutral, and negative labels	English
[43]	2021	Aspect-Oriented Long Text Dataset (AOLTD).	Collected data from public news websites. Positive, Neutral and Negative	Chinese/ English
[44]	2021	Online reviews	The internet reviews - Ctrip.com & Dazhong.com, 4 travel destinationsHua Mountain (X1), Emei Mountain (X2), Taishan Mountain (X3), Huang Mountain (X4)	Chinese
[45]	2021	SemEval2014: Laptop and Restaurant	The dataset consists of aspect terms and their corresponding sentiment polarities, positive,neutral and negative.	English
[46]	2021	Social media comments	Clean and Offensive/Hateful	EnglishArabic
[1]	2020	Kannada Codemixed Dataset (KanCMD)	Classified as Positive, Negative, Mixed feelings, Neutral, Other language	Kannada-English
[2]	2020	Social media comment	Classified as PositiveNegative, Neutral	Chinese-English
[3]	2020	Social media comment-Twitter	600 tweets per second and 500million tweets per day are taken as input	English
[47]	2020	Hindi-English code-mixedHinglish Offensive Tweet (HOT) Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC)	Class labels are - Offensive Speech or Normal Speech, Abusive and Hate, hate and non-offensive	Hindi English
[48]	2020	Hindi-English code-mixed	Classified as Anger, Happiness, Sadness	Hindi, English
[49]	2020	English, Arabic,Multilingual dataset	Rated as - obscene, offensive, and clean.Tweets - English texts, French texts, Arabic texts.	EnglishArabic
[6]	2020	Trolling Aggression and Cyberbullying (TRAC) Dataset sampled from Facebook/Twitter plugins.	15,001 aggression-annotated Facebook and Twitter posts	English, Hindi
[50]	2020	The dataset is made by the opinion of the public.	Classified as Positive, Negative, Neural, Not Mentioned	
[51]	2020	Chinese Corpus is the dataset.	The corpus has 2,105 documents from SINA news. Labels as Disgust, Anger, Sadness, HappinessFear, Surprise	English
[52]	2020	Synthesized gold standard Commercial anti-spam measures	Classified as truthful reviews, deceptive reviews	English
[53]	2020	310 million COVID 19 Tweets Dataset and the geo version of the dataset Geo COVID 19Tweet Dataset	Classified as Positive Sentiment, Negative Sentiment, Neutral Sentiment	English
[54]	2019	MRD (Mobile Phone Reviews), HRD (Hotel Reviews)	1. Reviews − 413,840 Sentences per review − 3.035 Words per review − 40.53 2. Reviews − 20,491 Sentences per review − 12.03Words per review − 104.36	English
[55]	2019	Cornell movie review Amazon product reviews	Both are classified as Positive Reviews and Negative Reviews	English
[56]	2019	Movie reviews, Stanford Sentiment Treebank SST-1, SST- 2 datasets	Labels: positive or negative.people's reviews about the movie. very negative, negative, neutral, positive, and very positive.	English
[57]	2019	Obama-McCain Debate (OMD) dataset and the Sentiment140 Twitter dataset	The OMD dataset Source: first presidential TV debate in the United States.Sentiment140 Twitter dataset-one million negative and positive tweets.	English
[58]	2019	Social Media comments	Classified as Negative and Neutral	English
[59]	2018	Social media comment-Twitter	2010 tweets classified as Clean, Offensive, Hateful Tweets	English
[60]	2018	Social media comment-Twitter	Classified as Neutral, Racism, sexism from 16 K tweets	English
[61]	2018	The Stanford Sentiment Treebank is a sentiment dataset, Sanders Twitter Corpus (STC)	The dataset includes hate crimes based on race, religion, sexuality, feminism, disability, and nationality. STC is evaluated as positive, negative, neutral, or irrelevant messages	English
[62]	2018	Foursquare dataset - New York City (NYC) and Los Angeles (LA).	POI category based on the venue id derived from check-in records. The dataset contains Users, POIs, check-ins and reviews	English
[63]	2015	ISEAR	7666 sentences (positive and negative product reviews.)	English
[64]	2015	Social Media comments	Classified as Positive and Negative	Chinese-English
[65]	2014	Feature Polarity dataset	Strong Positive, Positive Neutral,Negative, Strong Negative are the labels	English
[66]	2013	600 Corpuses Movie Reviews	300 Positive reviews300 Negative reviews	English
[67]	2019	Synthetic / Facebook	Binary class as Islamophobic or not, multi-topic related class such as Culture, Crimes, Economics, History, Terrorism, Rapism, Women Oppression.	English
[68]	2018	Twitter	Classified as hate speech, aggressiveness, offensiveness, irony, stereotype, and intensity.	Italy

broadly categorized into three groups based on the labeling of training samples employed by machine learning techniques namely supervised learning methods, semi supervised learning methods, and unsupervised learning methods [69].

3.1.1. *Supervised learning*

The most up-to-date algorithms for detecting abusive content use supervised learning to classify text and are oriented on identifying a good set of characteristics to perform the classification. The supervised machine learning algorithms looked at a variety of factors such as

comment content, user profile, user behaviour, and social graph structure. Experts or crowdsourcing platforms are used to label the training data. However, the effectiveness of these approaches is highly dependent on training data, necessitating the use of a large volume of labeled data for training.

3.1.2. Semi-Supervised learning

To generate models, semi-supervised machine learning techniques blend labeled and unlabeled data. To minimize the manual annotation work, Xiang et. al. [70] suggested a semi-supervised technique for identifying hate speech content in the Twitter corpus. This work supplies a tiny batch of seed data using the bootstrapping technique to classify unlabeled data. By automatically extracting variables from language regularities in profane language, the authors utilized a statistical topic model called Latent Dirichlet Allocation (LDA) [4,70] to locate offending tweets.

3.1.3. Unsupervised learning

Unsupervised learning algorithms work out how data might be grouped into clusters. As a result, data does not require labeling. It entails learning to discriminate the provided input data from unlabelled data. Growing Hierarchical Self-Organizing Maps (SOMs) is an unsupervised method proposed by Capua et. al. [71] that can effectively cluster documents, including bully traces. [27,72] investigated the sentiment of Twitter messages using the clustering approach based on a machine learning algorithm.

3.1.4. Machine learning models for hate speech detection and sentiment analysis

The anonymity of social networks attracts hate speech, which presents a problem for the entire world, to hide their unlawful online behaviour. Detecting hate speech is crucial given the growing volume of social media data since it can have negative impacts on society [44]. The most recent machine learning algorithms for detecting hate speech are covered in the discussion that follows.

Classical Machine Learning methods: The term "shallow detection" refers to word encoding techniques used by classical word representation hate speech detectors. After that, shallow classifiers can be used to perform the classification. The tagged dataset is used to train the learning algorithms, resulting in a model that can be used to detect and classify hate speech and non-hate speech in texts. Two examples of feature representation strategies that can be applied are TF-IDF and N-grams. Traditionally, supervised machine learning methods like Naive Bayes (NB) [73–76], Decision Tree (DT), Support Vector Machine (SVM) [77–81], Linear Regression [75,81], and Logistic Regression (LR) [82] have been used to detect hope speech and sentiment analysis.

Ensemble approach: The ensemble technique was developed to overcome the limitations of several individual machine learning algorithms while enhancing their strengths [50,83]. Each model has its own set of flaws; thus, no model is ideal. But, ensemble approaches attempt to combine the benefits of multiple models to provide better performance than any single model can provide. Combining two or more machine learning algorithms can minimize variance and increase learning capacity greatly, according to statistics. Bagging methodology, Random Forest (RF) [81,84–88], and boosting method [81,89,90] are some of the ensemble techniques.

Word-embeddings based methods: Word embedding learns the vectorized representations from scattered representations, which are then employed in downstream text mining activities. The embeddings make it possible for semantically related phrases to share the same vector representation. Many word embedding algorithms have been developed over the years, including Glove, word2vec, and FastText [90–93]. The representations from the word embedding techniques are fed into various classifiers.

3.1.5. Deep learning model for hate speech detection and sentiment analysis

Deep learning introduces a multi-layer structure in the neural network's hidden layers, enabling it to attain more intricate outcomes. Unlike conventional machine learning methods where features are manually specified or obtained through feature selection techniques such as TF-IDF, Word2Vec etc., deep learning models autonomously learn and extract information, resulting in enhanced accuracy and overall performance. To predict and categorize hate speech and sentimental texts, deep learning techniques have been utilized in a range of studies in the fields of data mining and text classification [69,94]. Below, we present a summary of the deep learning models used for sentiment analysis and hate speech detection.

3.1.5.1. Recurrent neural networks (RNNs). RNN is a subclass of artificial neural networks and assesses time series or sequential data. The sole purpose of common feed forward neural networks is to process unrelated pieces of data. If, however, we have data in a sequence where one data point depends on the data point before it, we will need to adjust the neural network to account for these dependencies. RNNs can remember the states or specifics of previous inputs to use when constructing subsequent outputs. RNNs are well-suited for tasks like sentiment analysis and hate speech detection [74,95] where the context and order of words in a text are crucial for making accurate predictions.

Long Short-Term Memory (LSTM) is a type of RNN that addresses the vanishing gradient problem, making it more effective in capturing long-range dependencies in sequential data. LSTM [73,77] can be used for sentiment analysis and hate speech detection like standard RNNs, but with the advantage of handling longer texts and preserving context over longer sequences. LSTM can effectively capture the sequential dependencies between words, allowing it to understand the context and sentiment expressed in the sentence. For hate speech detection, LSTM works [80,81,95–98] similarly to sentiment analysis. The LSTM processes the input text word by word and updates its hidden state at each time step, capturing the contextual information and dependencies between words. To enhance the performance of LSTM for hate speech detection, additional techniques like attention mechanisms are also incorporated. Attention mechanisms allow the model to focus on specific parts of the text that are more indicative of hate speech, leading to improved accuracy.

Gated Recurrent Unit (GRU) is another type of RNN that, like LSTM, addresses the vanishing gradient problem and captures long-range dependencies in sequential data. GRU is used for sentiment analysis and hate speech detection like LSTM and standard RNNs. To perform sentiment analysis using GRU, the model processes the input sentence word by word, updating its hidden state at each time step [74,92]. GRU can effectively capture the sequential dependencies between words, allowing it to understand the context and sentiments expressed in the sentence. For hate speech detection, GRU works similarly to sentiment analysis. The GRU processes the input text word by word and updates its hidden state at each time step, capturing the contextual information and dependencies between words.

3.1.5.2. Convolution neural networks (CNNs). CNNs are powerful deep learning models that have been widely used for various computer vision tasks, such as image classification and object detection. But in recent times, CNNs are also adapted for NLP tasks, including sentiment analysis and hate speech detection. CNNs are used for sentiment analysis by treating the text as a one-dimensional vector and applying 1D convolutions to capture local patterns and features within the text. As in sentiment analysis, CNNs are also used for hate speech detection [79–81,95,97,98]. In such cases, the convolutional layer applies filters to the sequences, capturing local patterns and features in the text. In both sentiment analysis and hate speech detection, CNNs excel at capturing local patterns and features from text data, making them effective tools for various NLP tasks.

3.1.6. Transformer based models

Transformer-based models have brought about a remarkable transformation in the field of NLP since they were introduced by Vaswani et. al. [99] in their article titled "Attention is All You Need". These models have emerged as the leading approach for various NLP tasks, mainly due to their capability to effectively handle long-range dependencies and process text in parallel. This parallel processing makes them highly efficient and scalable, outperforming traditional RNNs and CNNs. The core concept of transformer-based models lies in their attention mechanism, allowing the model to focus on relevant parts of the input text while making predictions. This attention mechanism enables the model to consider the contextual relationships between each word and all other words in the sentence, effectively capturing comprehensive contextual information.

Transformer-based models, including Bidirectional Encoder Representations from Transformers (BERT) [78,92,93,100–102], Generative Pre-trained Transformer (GPT), and A Robustly Optimized BERT Pretraining Approach (RoBERTa) [100,103], have demonstrated their potential across diverse NLP tasks. These tasks encompass sentiment analysis, named entity recognition, machine translation, question-answering, and various others. These models are typically pre-trained on extensive text corpora and subsequently fine-tuned for specific downstream tasks, enabling them to deliver exceptional performance even with limited training data. Contextual understanding, bidirectional context, pre-training on large corpora, transfer learning, and the attention mechanism make transformer models highly effective for sentiment analysis and hate speech detection tasks.

3.1.7. Hybrid models

In addition to the above models, hybrid models have also been proposed for sentiment analysis and hate speech detection. In an attempt by Mathew et. al. [104] CNN has been integrated with GRU to combine their strengths for speech detection tasks. Kahn et. al. [105] proposed a CNN-LSTM framework for Roman Urdu and English dialect sentiment analysis. Such combinations allow the models to capture both local patterns and long-term dependencies in the text data. Recently, Graph-based Neural Networks (GNNs) have emerged as powerful models for NLP tasks, offering new ways to represent and process textual data based on graph structures. In GNNs for NLP, the input text is transformed into a graph representation, where nodes represent words or entities, and edges encode the relationships between them. The graph can be constructed using various techniques, such as dependency parsing, co-occurrence information, or entity relationships in knowledge graphs. These models leverage the inherent connections and relationships between words or entities in a sentence to extract meaningful information and achieve state-of-the-art performance in various NLP applications.

In an attempt by Sarracén and Rosso [106], an unsupervised approach using Graph Auto-Encoders (GAE) has been proposed to represent texts as nodes of a graph and use a transformer layer together with a convolutional layer to encode these nodes in a low-dimensional space. As a result, embeddings can be decoded into a reconstruction of the original network. The authors employed this strategy to detect hate speech in multi-domain and multilingual sets of texts. By leveraging graph-based representations and convolutional operations, CGNNs have shown potential in capturing the structural information and linguistic patterns within the texts. A study by Sarracén and Rosso [107] has demonstrated the effectiveness of CGNNs in detecting hate speech by constructing graphs from texts and utilizing convolutional graph neural networks to learn meaningful embeddings. Additionally, in their work, Sarracén and Rosso [108] introduced an unsupervised hybrid method that merges BERT's multi-head self-attention with a word graph reasoning approach. The attention mechanism enables the model to grasp word relationships within a context while simultaneously learning a language model. Fig. 3 shows the general framework used for detecting hate speech and sentiment analysis.

Table 2 summarizes the recent attempts that perform sentiment analysis and hate speech detection using the above-mentioned approaches.

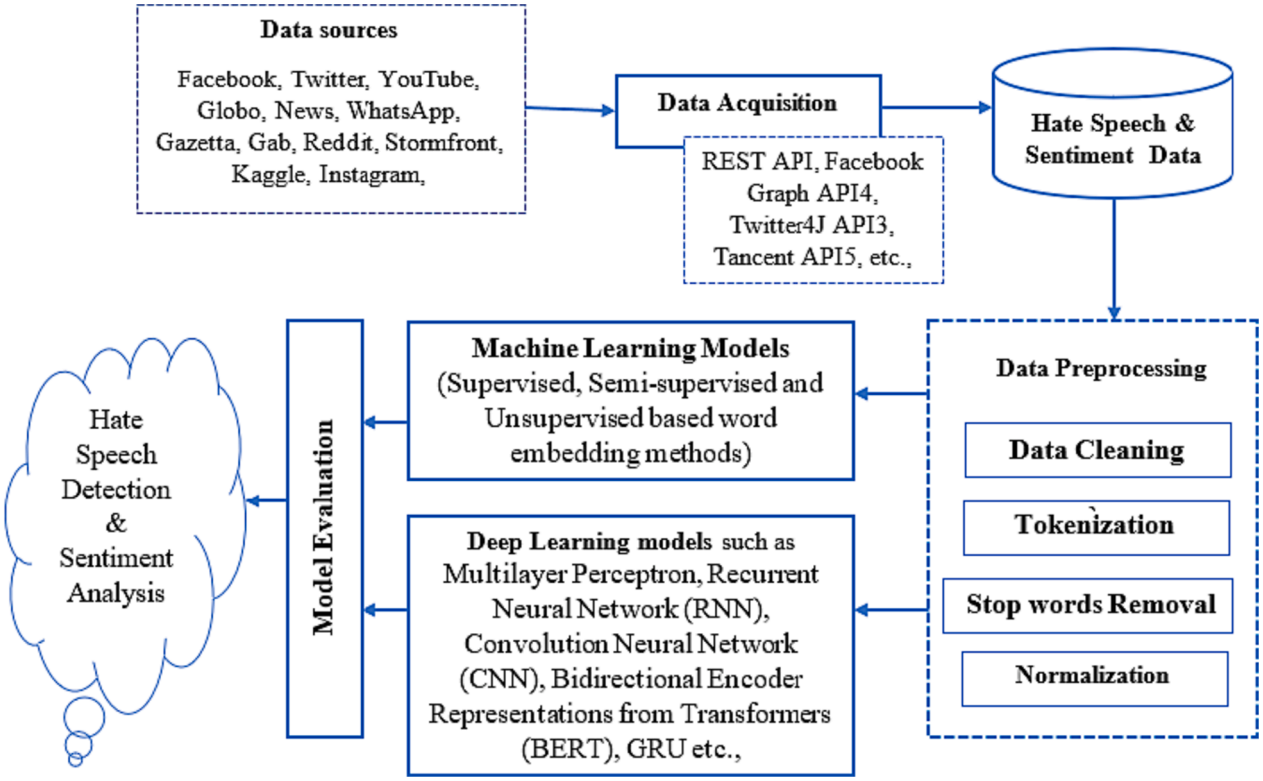

Fig. 3. General Approach for Hate Speech Detection and Sentiment Analysis.

Table 2
Review of the models proposed for Hate Speech Detection and Sentiment Analysis.

Ref	Dataset Size	Language	Description of Datasets	Models
[73]	11,874	Albanian	Offensive or not, untargeted or targeted; person, group or other, gives the abuse Language detection in social media data	NB BiLSTM BERT
[74]	6136	Arabic	Hate or Not Religious and its subcategories	RNN, GRU
[91]	30,000	Bengali	Hateful or not sports, politics, crime, religion, etc.,	Word2Vec, FastText and BengFastText with
[77]	10,496	Chinese	Sexism detection in the Chinese language on social media in. Classes contain Sexist, Non-sexist, Stereotype based on Cultural Background & Appearance, Micro Aggression, etc.,	SVM, LSTM, BiLSTM
[78]	4185	English	Directedness explicit & implicit Target group, Abuse severity sexist, homophobic, sexual harassment, racist, ableist, transphobic, intellectual. Presents the Abuse detection in conversational AI	SVM, BERT, MLP, Random Forest
[100]	50,000	English	Hate speech measurement on social media in English. Subcategories include race, ethnicity, religion, citizenship, gender, sexual violence, orientation, age, national origin, (dis)respect, political ideology, disability, insult, sentiment, humiliation, dehumanization inferior status, attack/defense, genocide, hate speech	Universal Sentence Encoders, BERT, and RoBERTa
[101]	3000	English	Hate or Offensive or not Data collected from Twitter for the election and the supporters of Trump or Biden	BERT
[102]	14,100	English	Presents the annotations of the offensive and abusive contents with labels.	BERT
[84]	1320	English	Binary class as abusive swear words and non-abusive swear words from the Twitter.	linear support classifier (LSVC), LR, RF
[75]	743	English	Binary class as non-offensive and offensive, which is related to the presidential election in the U.S in 2016	LR, NB, DNN, Stcked LSTM, BiLSTM, CNN, VGG16, Stacked LSTM + VGG16, BiLSTM + VGG16, CNNText + VGG16
[109]	5912	English	Binary class Hate and Not Hate. Hate consists of Type and Target.	DeBERTA, DeBERTa model with 100x upsampling
[110]	3728	English	Binary class with Hate and Not Hate. Seven targets in Hate such as Trans people, Women, Muslims, Gay people, Black people, Disabled people and Immigrants.	Google Jigsaw's Perspective (P) and Two Hat's SiftNinja (SN), Fine-tuned BERT
[103]	10,629	English	Binary toxic spans which include toxic & non-toxic. Predicts the spans of incorrect toxic posts.	RoBERTa with FLAIR and FastText
[111]	5003	English	Binary class with hateful or not. Classified based on race, religion, sexual orientation, disability, country of origin, and gender.	GPT-2
[104]	20,148	English	Level of hate are labeled as hate, offensive and normal), concerning the target class race, gender, religion, sexual orientation, rationales, miscellaneous	CNN-GRU, BiRNN, BiRNNwitH Attention, BERT
[79]	14,100	English	Branching of the structure of tasks. A as offensive or not, C as individual, B as targeted insult or untargeted, group and other.	SVM, BiLSTM, CNN
[76]	33,458	English	Hate speech, offensive but not hate speech, or neither offensive nor hate speech	Logistic Regression, NB, decision trees, random forests, and linear SVMs
[80]	10,568	English	Hate speech or not	SVM, CNN, LSTM
[112]	16,914	English	3-topic as the class of Sexist, Racist and Not Racism and Sexism	n-gram with LR
[113]	27,665	English	Classified as Assault on human Dignity or Call for Violence or offense (with subclasses)	TF-IDF and LIWC with SVM
[83]	3977 and 4138	English and Spanish	Binary as misogyny or not, And five categories as Stereotype, derailing, dominance, discredit, target of misogyny as active or passive Sexism, sexual harassment.	SVM and Ensemble of classifiers
[67]	1288 (6654 after augmentation)	English	Binary class as Islamophobic or not, multi-topic related class such as Culture, Crimes, Economics, History, Terrorism, Racism, Women Oppression.	Moved to Table 1
[89]	4972	English	Binary class as hateful or normal	GradBoost, GraphSage with Glove
[95]	33,776 22,324	English	Binary class as hateful or not hate speech	LR, SVM,CNN, RNN, Seq2Seq, Variable AutoEncoder, Reinforcement Learning
[96]	5647 4014 3353	English, French, Arabic	Detailed taxonomy with attributes such as Hostility, and Directness. Target, Group, Annotator	LR, LSTM, BiLSTM with BoW
[97]	14,100	English	1.Binary with Offensive or Not, 2. Targeted insult and untargeted 3.Within Target consists of Individual/Group/ others	CNN, BiLSTM, and SVM
[92]	13,000 6000	English Spanish	Within Hate as Hateful/ non-hateful, within target Group/Individual, Within Aggressive as Aggressive/Not)	SVM, MFC, BiLSTM, GRU, BERT, BiGRU with TF-IDF and GloVe
[81]	998 433	English	Binary as Hate/ Not. Multi class as 'violence', 'directed_vs generalized', 'gender', 'race', 'national_origin', 'disability', 'sexual_orientation', 'religion	LR, SVMs, RF and Gradient Boosting, CNN, LSTM, BiLSTM
[90]	5143 137,098	English	Binary with Hate /Not, Multinomial classification with 21 categories classified as 'hate targets' 'hateful language', and 'hate sub-targets'	Logistic Regression, Decision Tree, Random Forest, Adaboost, and Linear SVM with TF-IDF, Word2Vec.
[114]	85,000	German	Multi-class label as sexism, racism, profane language, threats, insults, etc.	Multinomial Naive Bayes, Logistic Regression, Gradient Boosted Trees and AutoML fine-tune three BERT models
[82]	8541	German	3 classes of offensive texts such as profanity, abuse and insult.	Logistic Regression with n-gram
[98]	4669 with subtasks	Hindi, German, English	Includes classes such as Hate speech, Offensive, Profane, person-directed or neither	LSTM, BERT, CNN

(*continued on next page*)

Table 2 (*continued*)

Ref	Dataset Size	Language	Description of Datasets	Models
[85]	8192	Hindi	Multi-tags in Hostile as Fake News, Defame, Binary as Hostile/ Not Hostile, Hate, Offense.	SVM, LR, Decision Tree, RF, MLP
[86]	3189	Hindi-English	Hierarchy as Not Offensive/ Abusive, Hate inducing	SVM, RF with N-gram, TF-IDF, BoWV, CNN, LSTM
[87]	4575	Hindi-English	Binary class as Hate/ Not	SVM with RBF, RF and with Char N-Gram and word N-Gram
[88]	13,169	Indonesian	There are numerous categories, such as no hate speech, no hate speech but abusive, hate speech but no abuse, and hate speech with abuse. There are also numerous subcategories.	SVM, NB, Random Forest Decision Tree
[115]	2016	Indonesian	Classified as Not abusive, Offensive, Abusive but not offensive and offensive	SVM, NB, Random Forest Decision Tree
[116]	9381	Korean	Ternary (Gender bias, other biases, None), Ternary (Hate, Offensive, None) Person/Group-directed, Binary (Gender bias, No gender bias), Gender/ Sexual orientation, Sexism, Harmfulness/Toxicity	CharCNN, BiLSTM, and BERT
[93]	6,000 1,961	Italian	Misogyny, non misogyny, aggressive and not aggressive	Word2Vec, GloVe, FastText, CNN, BERT, Finetuned sentence-BERT
[117]	10,336	Portuguese	Classified as racism, sexism, homophobia, xenophobia, religious intolerance, or cursing	NB and SVM
[118]	100,000	Russian	Lookism, sexism, nationalism, threats, harassment, homophobia, and other	BiLSTM, CNN with self-attention
[119]	11,000	Spanish	Binary class as Aggressive or Not	NB, LR, SVM
[107]	14,100 30,000	English	Offensive and Not-offensive	LSTM with GNN, Attention mechanism of BERT

4. Findings and discussion

In this comprehensive survey, we have given a critical review of how sentiment analysis and hate speech detection have developed over recent years. We have explored the publicly available hate speech and sentiment analysis corpora in different languages and presented a brief description in Table 1. From an in-depth analysis of the datasets, we identified four major problems such as (i) noisy/imbalance nature of the data, (ii) presence of highly skewed data in both multi-label and multi-class, (iii) the representation of the feature vector in the machine learning and deep learning models, (iv) the sparse data representation. Also, based on the comprehensive analysis of the dataset, getting them labelled from the experts or linguistics is important in the dataset preparation. Concerning hate speech, it is found that there are some attempts where abusive speech and hate speech have been interchangeably used. But, the authors of [8] have clearly distinguished between toxic/abusive language, hate speech, and offensive language. Based on the results of our systematic review, we understood that the data in online hate speech is skewed, and that further research is needed to find effective ways to annotate new posts and communications.

Next, we provided an overview of various models utilized in hate speech detection and sentiment analysis, accompanied by the datasets used. Our examination of the surveyed articles revealed that certain attempts categorize offensive texts as hate speech. Fig. 4 illustrates that while deep learning models have been introduced, machine learning models continue to be more prevalent in research efforts.

Most of the work focused on comments in English and only a few works in other Languages like Spain, Chinese etc. When training models to find hate speech, it is very important to have access to labeled datasets. English is one of the most commonly spoken languages, so there is more information about hate speech and sentiment analysis in English than in many other languages. Because there aren't enough data for other languages, it's hard to make and test models that can find hate speech. Hate speech identification is a difficult task, because models must understand the complexities and nuances of each language. Because different languages have different sentence patterns, vocabularies, and cultural references, adapting English-based models to other languages is more difficult. To address these issues, even though the original BERT model developed by Google was pretrained on English text, since its release, researchers have extended BERT's pretraining to support multiple languages. This leads to models like Multilingual BERT (mBERT), BERT-Base, Multilingual Cased etc. By leveraging the knowledge learned during pretraining, these models can be fine-tuned on specific downstream tasks, such as sentiment analysis, named entity recognition, and machine translation, in various languages.

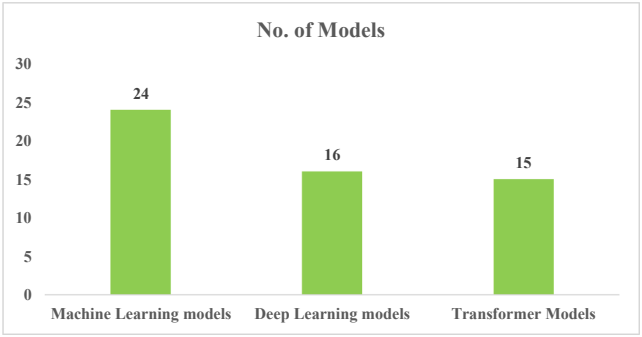

Fig. 4. Number of Models for Hate Speech Detection and Sentiment Analysis.

From the survey, we understand that in recent times, researchers have started to use transformer-based models for NLP tasks. The transformers generate contextual embeddings and relationships using attention mechanisms for each word in a sentence and these embeddings encode rich contextual information that captures the meaning of words based on their surrounding words. Hence, the transformers have achieved state-of-the-art results in various NLP tasks, including hate speech detection and sentiment analysis. Their advanced architecture and capabilities make them the preferred choice for many researchers and practitioners in real-time applications. Finally, we found that finding the datasets, annotation of the datasets, datasets with different languages and open-source availability of the code for the diverse languages needs to improve for languages other than English. Further, it can be undoubtedly identified that the field of hate speech detection and sentiment analysis would results in the impact in societal applications with many research challenges.

In this review, we have considered only the models where hate speech was conveyed explicitly. Often, this is not the case and it would be important to say that hate speech can be also conveyed in an implicit way [120], in the form of stereotypes [121], or figurative language, e.g. sarcasm [122] or humor [123]. Implicit hate speech, being insidious, might escape the attention of conventional models. Nevertheless, its effects can be as harmful as explicit hate speech, resulting in emotional distress, discrimination, and potential violence. Thus, detecting and addressing such content at an early stage is vital to prevent harm and minimize its consequences.

5. Challenges and issues of hate speech detection

Hate speech detection and sentiment analysis are essential NLP tasks, but they come with various challenges and issues. Some of the main challenges and issues include:

(1) Usually, social media messages contain poorly written texts which do not reside in the formal structure to find out the patterns in the text.
(2) Developing high-quality labeled datasets for hate speech and sentiment analysis, especially in languages other than English, can be time-consuming and expensive. The scarcity of diverse and annotated data limits the performance of models, particularly for low-resource languages.
(3) Extending hate speech detection and sentiment analysis to multiple languages introduces language-specific complexities, including varying grammatical structures, sentiment lexicons, and cultural expressions.
(4) In hate speech identification and sentiment analysis, the data distribution and imbalance nature are one of the issues for finding a meaningful pattern in the data.
(5) Hate speech and sentiment expression can be highly subjective and context-dependent. What may be considered hateful in one context and may not be so in another. Detecting hate speech accurately requires understanding the context and cultural nuances, which can be challenging for algorithms. For example, the sentence: "I'm dying to meet you!", in some English-speaking regions, this phrase might be interpreted as a positive expression of eagerness or excitement to meet someone. However, in other places, the use of "dying" might be considered inappropriate or negative due to the literal meaning of the word.
(6) The interpretation of implicit hate speeches heavily depends on the context in which they are used. Without a proper understanding of the context, it can be challenging to distinguish between harmful and innocuous statements.

6. Conclusion

In this review article, we have provided a comprehensive overview of the current state-of-the-art methods in hate speech detection and sentiment analysis. We investigated a vast array of methodologies and datasets, highlighting the advancements made in this vital field. The analysis of conventional machine learning approaches, such as SVM, Decision Tree, NB, LR etc. revealed the foundation upon which subsequent research has built more complex models. Deep learning models, such as CNNs, RNNs, and Transformer-based architectures such as BERT and GPT, have demonstrated remarkable abilities in detecting hate speech and analyzing sentiment in online content since their introduction.

Several promising research directions can further advance the field of hate speech detection and sentiment analysis. It is important to explore transfer learning approaches that can utilize knowledge from high resource languages to enhance hate speech detection in languages with limited resources. Further, exploring the transferability and application of methodologies derived from hate speech detection and sentiment analysis to address other NLP tasks, like cyberbullying detection, depression detection, or identifying fake news may be focused.

Funding Support

This work was supported the Korea Environmental Industry & Technology Institute (KEITI), with a grant funded by the Korea government, Ministry of Environment (The development of IoT-based technology for collecting and managing big data on environmental hazards and health effects), under Grant RE202101551 and partially supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) funded by the Korean Government, Ministry of Science and ICT (MSIT) (Implementation of Verification Platform for ICT Based Environmental Monitoring Sensor), under Grant 2019–0-00135 and partially supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) funded by the Korea Government, Ministry of Science and ICT (MSIT) (Building a Digital Open Lab as open innovation platform) under Grant 2021–0-00546.

CRediT authorship contribution statement

Malliga Subramanian: Conceptualization, Methodology, Investigation, Writing – original draft. **Veerappampalayam Easwaramoorthy Sathiskumar:** Conceptualization, Methodology, Investigation, Writing – original draft, Visualization, Supervision, Funding acquisition. **G. Deepalakshmi:** Project administration, Methodology, Investigation. **Jaehyuk Cho:** Methodology, Formal analysis, Writing – review & editing, Visualization. **G. Manikandan:** Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Hande, R. Priyadharshini, B.R. Chakravarthi, KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection, in: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media. 2020.
[2] J. Cao, et al., A risky large group emergency decision-making method based on topic sentiment analysis, Expert Systems with Applications 195 (2022), 116527.
[3] P.K. Roy, et al., A framework for hate speech detection using deep convolutional neural network, IEEE Access 8 (2020) 204951–204962.
[4] H. Liu, et al., A fuzzy approach to text classification with two-stage training for ambiguous instances, IEEE Transactions on Computational Social Systems 6 (2) (2019) 227–240.
[5] F.M. Plaza-Del-Arco, et al., A multi-task learning approach to hate speech detection leveraging sentiment analysis, IEEE Access 9 (2021) 112478–112489.
[6] S. Modha, et al., Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance, Expert Systems with Applications 161 (2020), 113725.
[7] S. Kaur, S. Singh, S. Kaushal, Abusive content detection in online user-generated data: a survey, Procedia Computer Science 189 (2021) 274–281.
[8] F. Poletto, et al., Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation 55 (2021) 477–523.
[9] M. Luo, X. Mu, Entity sentiment analysis in the news: A case study based on negative sentiment smoothing model (nssm), International Journal of Information Management Data Insights 2 (1) (2022), 100060.
[10] A. Rodriguez, Y.-L. Chen, C. Argueta, FADOHS: framework for detection and integration of unstructured data of hate speech on facebook using sentiment and emotion analysis, IEEE Access 10 (2022) 22400–22419.
[11] S. Khan, et al., HCovBi-caps: hate speech detection using convolutional and Bi-directional gated recurrent unit with Capsule network, IEEE Access 10 (2022) 7881–7894.
[12] H. Wu, et al., Phrase dependency relational graph attention network for Aspect-based Sentiment Analysis, Knowledge-Based Systems 236 (2022), 107736.
[13] M. Hong, J.J. Jung, Sentiment aware tensor model for multi-criteria recommendation, Applied Intelligence 52 (13) (2022) 15006–15025.
[14] A. Kumar, et al., BILEAT: a highly generalized and robust approach for unified aspect-based sentiment analysis: BILEAT, Applied Intelligence 52 (12) (2022) 14025–14040.
[15] R.M. Cruz, W.V. de Sousa, G.D. Cavalcanti, Selecting and combining complementary feature representations and classifiers for hate speech detection, Online Social Networks and Media 28 (2022), 100194.
[16] L.-L. Tao, T.-H. You, A multi-criteria decision-making model for hotel selection by online reviews: Considering the traveller types and the interdependencies among criteria, Applied Intelligence 52 (11) (2022) 12436–12456.
[17] R. Wang, et al., Post-processing method with aspect term error correction for enhancing aspect term extraction, Applied Intelligence 52 (14) (2022) 15751–15763.
[18] J.-L. Wu, W.-Y. Chung, Sentiment-based masked language modeling for improving sentence-level valence–arousal prediction, Applied Intelligence 52 (14) (2022) 16353–16369.
[19] F.R. Nascimento, G.D. Cavalcanti, M. Da Costa-Abreu, Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on

social media using ensemble learning, Expert Systems with Applications 201 (2022), 117032.
[20] M. Sharma, I. Kandasamy, V. Kandasamy, Deep learning for predicting neutralities in offensive language identification dataset, Expert Systems with Applications 185 (2021), 115458.
[21] H.S. Alatawi, A.M. Alhothali, K.M. Moria, Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT, IEEE Access 9 (2021) 106363–106374.
[22] M.Z. Ali, et al., Improving hate speech detection of Urdu tweets using sentiment analysis, IEEE Access 9 (2021) 84296–84305.
[23] A. Zhao, Y. Yu, Knowledge-enabled BERT for aspect-based sentiment analysis, Knowledge-Based Systems 227 (2021), 107220.
[24] S.H. Biradar, J. Gorabal, G. Gupta, Machine learning tool for exploring sentiment analysis on twitter data, Materials Today: Proceedings 56 (2022) 1927–1934.
[25] K.A. Qureshi, M. Sabih, Un-compromised credibility: Social media based multi-class hate speech classification for text, IEEE Access 9 (2021) 109465–109477.
[26] J. Chen, et al., A classified feature representation three-way decision model for sentiment analysis, Applied Intelligence (2022) 1–13.
[27] F.E. Ayo, et al., A probabilistic clustering model for hate speech classification in twitter, Expert Systems with Applications 173 (2021), 114762.
[28] D.R. Beddiar, M.S. Jahan, M. Oussalah, Data expansion using back translation and paraphrasing for hate speech detection, Online Social Networks and Media 24 (2021), 100153.
[29] S. Agarwal, C.R. Chowdary, Combating hate speech using an adaptive ensemble learning model with a case study on COVID-19, Expert Systems with Applications 185 (2021), 115632.
[30] F.M. Plaza-del-Arco, et al., Comparing pre-trained language models for Spanish hate speech detection, Expert Systems with Applications 166 (2021), 114120.
[31] P. Le-Hong, Diacritics generation and application in hate speech detection on Vietnamese social networks, Knowledge-Based Systems 233 (2021), 107504.
[32] H. Karayiğit, Ç.İ. Acı, A. Akdağlı, Detecting abusive Instagram comments in Turkish using convolutional Neural network and machine learning methods, Expert Systems with Applications 174 (2021), 114802.
[33] V.K. Jha, et al., DHOT-repository and classification of offensive tweets in the Hindi language, Procedia Computer Science 171 (2020) 2324–2333.
[34] D. Antonakaki, P. Fragopoulou, S. Ioannidis, A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks, Expert Systems with Applications 164 (2021), 114006.
[35] C. Liu, et al., Emoji use in China: popularity patterns and changes due to COVID-19, Applied Intelligence 52 (14) (2022) 16138–16148.
[36] F. Huang, et al., Multi-granular document-level sentiment topic analysis for online reviews, Applied Intelligence (2022) 1–11.
[37] E. Pronoza, et al., Detecting ethnicity-targeted hate speech in Russian social media texts, Information Processing & Management 58 (6) (2021), 102674.
[38] J. Kocoń, et al., Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach, Information Processing & Management 58 (5) (2021), 102643.
[39] J. Jang, et al., Sequential targeting: A continual learning approach for data imbalance in text classification, Expert Systems with Applications 179 (2021), 115067.
[40] AL-Sharuee, M.T., F. Liu, and M. Pratama, Sentiment analysis: dynamic and temporal clustering of product reviews. Applied Intelligence, 2021. 51: p. 51-70.
[41] X. Xue, J. Feng, X. Sun, Semantic-enhanced sequential modeling for personality trait recognition from texts, Applied Intelligence (2021) 1–13.
[42] Y. Wu, W. Li, Aspect-level sentiment classification based on location and hybrid multi attention mechanism, Applied Intelligence 52 (10) (2022) 11539–11554.
[43] Z. Wu, et al., Make aspect-based sentiment classification go further: step into the long-document-level, Applied Intelligence (2021) 1–20.
[44] J. Wu, et al., A group consensus-based travel destination evaluation method with online reviews, Applied Intelligence 52 (2) (2022) 1306–1324.
[45] D. Zhang, et al., Syntactic and semantic analysis network for aspect-level sentiment classification, Applied Intelligence 51 (8) (2021) 6136–6147.
[46] S. Alsafari, S. Sadaoui, Semi-supervised self-training of hate and offensive speech from social media, Applied Artificial Intelligence 35 (15) (2021) 1621–1645.
[47] K. Sreelakshmi, B. Premjith, K. Soman, Detection of hate speech text in Hindi-English code-mixed data, Procedia Computer Science 171 (2020) 737–744.
[48] T.T. Sasidhar, B. Premjith, K. Soman, Emotion detection in hinglish (hindi+ english) code-mixed social media text, Procedia Computer Science 171 (2020) 1346–1352.
[49] S. Alsafari, S. Sadaoui, M. Mouhoub, Hate and offensive speech detection on Arabic social media, Online Social Networks and Media 19 (2020), 100096.
[50] W. Liao, et al., An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa, Applied Intelligence 51 (2021) 3522–3533.
[51] M. Li, et al., Emotion-cause span extraction: a new task to emotion cause identification in texts, Applied Intelligence (2021) 1–13.
[52] J. Li, et al., Identifying ground truth in opinion spam: an empirical survey based on review psychology, Applied Intelligence 50 (2020) 3554–3569.
[53] R. Lamsal, Design and analysis of a large-scale COVID-19 tweets dataset, applied intelligence 51 (2021) 2790–2804.
[54] B. Bansal, S. Srivastava, Hybrid attribute based sentiment classification of online reviews for consumer intelligence, Applied Intelligence 49 (1) (2019) 137–149.
[55] J. Khan, et al., EnSWF: effective features extraction and selection in conjunction with ensemble learning methods for document sentiment classification, Applied Intelligence 49 (2019) 3123–3145.
[56] J. Xie, et al., Enhancing sentence embedding with dynamic interaction, Applied Intelligence 49 (2019) 3283–3292.
[57] Y. Zhang, et al., A quantum-inspired sentiment representation model for twitter sentiment analysis, Applied Intelligence 49 (2019) 3093–3108.
[58] Y. Tang, N. Dalzell, Classifying hate speech using a two-layer model, Statistics and Public Policy 6 (1) (2019) 80–86.
[59] H. Watanabe, M. Bouazizi, T. Ohtsuki, Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection, IEEE Access 6 (2018) 13825–13835.
[60] G.K. Pitsilis, H. Ramampiaro, H. Langseth, Effective hate-speech detection in Twitter data using recurrent neural networks, Applied Intelligence 48 (2018) 4730–4742.
[61] N. Zainuddin, A. Selamat, R. Ibrahim, Hybrid sentiment classification on twitter aspect-based sentiment analysis, Applied Intelligence 48 (2018) 1218–1232.
[62] S. Xing, et al., Content-aware point-of-interest recommendation based on convolutional neural network, Applied Intelligence 49 (2019) 858–871.
[63] KM, A.K., et al., A multimodal approach to detect user's emotion. Procedia Computer Science, 2015. 70: p. 296-303.
[64] Y.-H. Kuo, et al., Integrated microblog sentiment analysis from users' social interaction patterns and textual opinions, Applied Intelligence 44 (2016) 399–413.
[65] F. Ali, E.K. Kim, Y.-G. Kim, Type-2 fuzzy ontology-based opinion mining and information extraction: A proposal to automate the hotel reservation system, Applied Intelligence 42 (2015) 481–500.
[66] G. Li, F. Liu, Sentiment analysis based on clustering: a framework in improving accuracy and recognizing neutral opinions, Applied intelligence 40 (2014) 441–452.
[67] Chung, Y.-L., et al., CONAN–COunter NArratives through Nichesourcing: a multilingual dataset of responses to fight online hate speech. arXiv preprint arXiv: 1910.03270, 2019.
[68] Sanguinetti, M., et al. An italian twitter corpus of hate speech against immigrants. in Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). 2018.
[69] E. Fu, J. Xiang, C. Xiong, Deep Learning Techniques for Sentiment Analysis, Highlights in Science, Engineering and Technology 16 (2022) 1–7.
[70] Xiang, B. and L. Zhou. Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. in Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers). 2014.
[71] Di Capua, M., E. Di Nardo, and A. Petrosino. Unsupervised cyber bullying detection in social networks. in 2016 23rd International conference on pattern recognition (ICPR). 2016. IEEE.
[72] S.S. Jacob, R. Vijayakumar, Sentimental analysis over twitter data using clustering based machine learning algorithm, Journal of Ambient Intelligence and Humanized Computing (2021) 1–12.
[73] Nurce, E., J. Keci, and L. Derczynski, Detecting abusive Albanian. arXiv preprint arXiv:2107.13592, 2021.
[74] Albadi, N., M. Kurdi, and S. Mishra. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. in 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). 2018. IEEE.
[75] Suryawanshi, S., et al. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. in Proceedings of the second workshop on trolling, aggression and cyberbullying. 2020.
[76] Davidson, T., et al. Automated hate speech detection and the problem of offensive language. in Proceedings of the international AAAI conference on web and social media. 2017.
[77] A. Jiang, et al., SWSR: A Chinese dataset and lexicon for online sexism detection, Online Social Networks and Media 27 (2022), 100182.
[78] Curry, A.C., G. Abercrombie, and V. Rieser, ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. arXiv preprint arXiv:2109.09483, 2021.
[79] Zampieri, M., et al., Predicting the type and target of offensive posts in social media. arXiv preprint arXiv:1902.09666, 2019.
[80] De Gibert, O., et al., Hate speech dataset from a white supremacy forum. arXiv preprint arXiv:1809.04444, 2018.
[81] Mollas, I., et al., Ethos: an online hate speech detection dataset. arXiv preprint arXiv:2006.08328, 2020.
[82] Wiegand, M., M. Siegel, and J. Ruppenhofer, Overview of the germeval 2018 shared task on the identification of offensive language. 2018.
[83] E. Fersini, P. Rosso, M. Anzovino, Overview of the task on automatic misogyny identification at IberEval 2018, Ibereval@ sepln 2150 (2018) 214–228.
[84] Pamungkas, E.W., V. Basile, and V. Patti. Do you really want to hurt me? predicting abusive swearing in social media. in Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020.
[85] Bhardwaj, M., et al., Hostility detection dataset in Hindi. arXiv preprint arXiv: 2011.03588, 2020.
[86] Mathur, P., et al. Did you offend me? classification of offensive tweets in hinglish language. in Proceedings of the 2nd workshop on abusive language online (ALW2). 2018.
[87] Bohra, A., et al. A dataset of Hindi-English code-mixed social media text for hate speech detection. in Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media. 2018.
[88] Ibrohim, M.O. and I. Budi. Multi-label hate speech and abusive language detection in Indonesian Twitter. in Proceedings of the third workshop on abusive language online. 2019.

[89] Ribeiro, M., et al. Characterizing and detecting hateful users on twitter. in Proceedings of the International AAAI Conference on Web and Social Media. 2018.
[90] Salminen, J., et al. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. in Proceedings of the International AAAI Conference on Web and Social Media. 2018.
[91] Romim, N., et al. Hate speech detection in the bengali language: A dataset and its baseline evaluation. in Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020. 2021. Springer.
[92] Basile, V., et al. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. in Proceedings of the 13th international workshop on semantic evaluation. 2019.
[93] Fersini, E., D. Nozza, and P. Rosso, AMI@ EVALITA2020: Automatic misogyny identification, in Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020). 2020, (seleziona...).
[94] Pitsilis, G.K., H. Ramampiaro, and H. Langseth, Detecting offensive language in tweets using deep learning. arXiv preprint arXiv:1801.04433, 2018.
[95] Qian, J., et al., A benchmark dataset for learning to intervene in online hate speech. arXiv preprint arXiv:1909.04251, 2019.
[96] Ousidhoum, N., et al., Multilingual and multi-aspect hate speech analysis. arXiv preprint arXiv:1908.11049, 2019.
[97] Zampieri, M., et al., Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). arXiv preprint arXiv:1903.08983, 2019.
[98] Mandl, T., et al. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. in Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation. 2019.
[99] A. Vaswani, et al., Attention is all you need, Advances in neural information processing systems 30 (2017).
[100] Kennedy, C.J., et al., Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. arXiv preprint arXiv:2009.10277, 2020.
[101] Grimminger, L. and R. Klinger, Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. arXiv preprint arXiv:2103.01664, 2021.
[102] Caselli, T., et al. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. in Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020.
[103] Pavlopoulos, J., et al. SemEval-2021 task 5: Toxic spans detection. in Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021). 2021.
[104] Mathew, B., et al. Hatexplain: A benchmark dataset for explainable hate speech detection. in Proceedings of the AAAI conference on artificial intelligence. 2021.
[105] L. Khan, et al., Deep sentiment analysis using CNN-LSTM architecture of English and Roman Urdu text shared in social media, Applied Sciences 12 (5) (2022) 2694.
[106] De la Peña Sarracén, G.L. and P. Rosso. Unsupervised embeddings with graph auto-encoders for multi-domain and multilingual hate speech detection. in Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022.
[107] G.L. De la Peña Sarracén, P. Rosso, Convolutional graph neural networks for hate speech detection in data-poor settings. International Conference on Applications of Natural Language to Information Systems, Springer, 2022.
[108] Sarracén, G.L.D.l.P. and P. Rosso, Offensive keyword extraction based on the attention mechanism of BERT and the eigenvector centrality using a graph representation. Personal and Ubiquitous Computing, 2023. 27(1): p. 45-57.
[109] Kirk, H.R., et al., Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. arXiv preprint arXiv:2108.05921, 2021.
[110] Röttger, P., et al., HateCheck: Functional tests for hate speech detection models. arXiv preprint arXiv:2012.15606, 2020.
[111] Fanton, M., et al., Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. arXiv preprint arXiv:2107.08720, 2021.
[112] Waseem, Z. and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. in Proceedings of the NAACL student research workshop. 2016.
[113] B. Kennedy, et al., The gab hate corpus: A collection of 27k posts annotated for hate speech, PsyArXiv. (July, 2018.) 18.
[114] Assenmacher, D., et al. $\texttt {RP-Mod}\\&\\texttt {RP-Crowd:} $ Moderator- and Crowd-Annotated German News Comment Datasets. in Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). 2021.
[115] M.O. Ibrohim, I. Budi, A dataset and preliminaries study for abusive language detection in Indonesian social media, Procedia Computer Science 135 (2018) 222–229.
[116] Moon, J., W.I. Cho, and J. Lee, BEEP! Korean corpus of online news comments for toxic speech detection. arXiv preprint arXiv:2005.12503, 2020.
[117] De Pelle, R.P. and V.P. Moreira. Offensive comments in the brazilian web: a dataset and baseline results. in Anais do VI Brazilian Workshop on Social Network Analysis and Mining. 2017. SBC.
[118] Zueva, N., M. Kabirova, and P. Kalaidin, Reducing unintended identity bias in Russian hate speech detection. arXiv preprint arXiv:2010.11666, 2020.
[119] Álvarez-Carmona, M.Á., et al. Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. in Notebook papers of 3rd sepln workshop on evaluation of human language technologies for iberian languages (ibereval), seville, spain. 2018.
[120] S. Frenda, V. Patti, P. Rosso, Killing me softly: Creative and cognitive aspects of implicitness in abusive language online, Natural Language Engineering (2022) 1–22.
[121] J. Sánchez-Junquera, et al., How do you speak about immigrants? taxonomy and stereoimmigrants dataset for identifying stereotypes about immigrants, Applied Sciences 11 (8) (2021) 3610.
[122] S. Frenda, et al., The unbearable hurtfulness of sarcasm, Expert Systems with Applications 193 (2022), 116398.
[123] L.I. Merlo, et al., When humour hurts: linguistic features to foster explainability, 2023.