Social Network Analysis and Mining (2022) 12:129
https://doi.org/10.1007/s13278-022-00951-3

REVIEW PAPER

Detection and moderation of detrimental content on social media platforms: current status and future directions

Vaishali U. Gongane[1] · **Mousami V. Munot**[1] · **Alwin D. Anuse**[2]

Received: 25 January 2022 / Revised: 6 August 2022 / Accepted: 8 August 2022

Abstract
Social Media has become a vital component of every individual's life in society opening a preferred spectrum of virtual communication which provides an individual with a freedom to express their views and thoughts. While virtual communication through social media platforms is highly desirable and has become an inevitable component, the dark side of social media is observed in form of detrimental/objectionable content. The reported detrimental contents are fake news, rumors, hate speech, aggressive, and cyberbullying which raise up as a major concern in the society. Such detrimental content is affecting person's mental health and also resulted in loss which cannot be always recovered. So, detecting and moderating such content is a prime need of time. All social media platforms including Facebook, Twitter, and YouTube have made huge investments and also framed policies to detect and moderate such detrimental content. It is of paramount importance in the first place to detect such content. After successful detection, it should be moderated. With an overflowing increase in detrimental content on social media platforms, the current manual method to identify such content will never be enough. Manual and semi-automated moderation methods have reported limited success. A fully automated detection and moderation is a need of time to come up with the alarming detrimental content on social media. Artificial Intelligence (AI) has reached across all sectors and provided solutions to almost all problems, social media content detection and moderation is not an exception. So, AI-based methods like Natural Language Processing (NLP) with Machine Learning (ML) algorithms and Deep Neural Networks is rigorously deployed for detection and moderation of detrimental content on social media platforms. While detection of such content has been receiving good attention in the research community, moderation has received less attention. This research study spans into three parts wherein the first part emphasizes on the methods to detect the detrimental components using NLP. The second section describes about methods to moderate such content. The third part summarizes all observations to provide identified research gaps, unreported problems and provide research directions.

Keywords Social media (SM) platforms · Detection and moderation · Natural language processing (NLP) · Artificial intelligence (AI)

1 Introduction

In recent years, internet has revolutionized the communication domain through social media networks where people from different communities, culture and organization across the globe interact virtually. Internet has brought a dramatic change from web-based search engines to social media websites and micro-blogging sites which is gaining more popularity. Social media are defined as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User-Generated Content" (Kaplan and Haenlein 2010). User Generated Content (UGC) describes the various forms of media content like text, video, and audio created by the end users with commercial marketing context in

✉ Vaishali U. Gongane
vug_entc@pvgcoet.ac.in

Mousami V. Munot
mvmunot@pict.edu

Alwin D. Anuse
alwin.anuse@mitwpu.edu.in

[1] E&TC Department, SCTR's Pune Institute of Computer Technology, SPPU, Pune 411046, India

[2] School of ECE, Dr Vishwanath Karad MIT_WPU, SPPU, Pune 411038, India

Published online: 05 September 2022

Table 1 Popular SM platforms

Name of SM platform	Category of SM	Year of launch	Characteristics of the platform
LinkedIn	Social networking site	2003	Professional networking website connect business people
			Used for used for professional networking and career development, and provide job opportunities
Facebook	Social networking site	2003	Enable users to stay connected with friends and relatives
			Users can chat, upload pictures, tell stories, share videos and links, post and read status
YouTube	Media sharing site	2005	Registered users can upload their content and share it with friends or provide it to the public
			Other users can rate and comment on the content
Twitter	Microblog	2006	Allows registered users to read and broadcast short messages called as tweets
			A tweet can be a text (140 characters), a photo or video content
WhatsApp	Messaging app	2009	Allows user to send text and voice messages, share images and videos
Instagram	Social networking site	2010	Designed to share photos and videos
			Users can create and share short videos with captions
Telegram	Messaging app	2013	Cloud based instant messaging and video calling service

mind. The UGC is published with either on publicly accessible website or social networking site accessible to certain group of people (Kaplan and Haenlein 2010). There are three aspects involved in definition of social media: Individuals create a public profile or a private profile. Second, individuals connect with friends, colleagues or relatives to form a network. Last, these individuals share their content and activities publicly in their network (Ellison 2007). All the three aspects are covered in various social networking sites like Facebook, Instagram, WhatsApp.

1.1 Various social media (SM) platforms

Before the invention of internet, SM began in the year 1844 with a series of electronic dots on a telegraph machine.[1] Bulletin Board Systems (BSS) was the first forms of SM that allowed users to log on and connect with each other. Usenet (USErNETwork) started by Tom Truscott and Jim Ellis in 1979 was a kind of discussion group where people can share views on topic of their interest and the article was available to all users in the group[1]. Six Degrees is considered to be the first social networking site similar to Facebook which had millions of registered users[1].

LiveJournal was a Weblog or blog publishing site that became popular in 1991. SM had various categories like blogs, forums, media sharing sites and social networking sites (Kaplan and Haenlein 2010). Table 1 shows the popular SM platforms[1] that have become an integral part of an individual's life. As shown in Table 1, the categories of SM provide the users to share the content in various formats. Figure 1 shows the statistics of monthly active users on SM platforms up to year 2022. Facebook is the most widely used platform. In the first quarter of year 2022, Facebook had roughly 2.93 billion monthly active users.[2] SM can also serve as apparatus that assists many external and internal organizational activities among peer groups, customers, business partners, and organizations which include knowledge sharing, marketing strategies, product management, collaborative learning and sharing (Ngai et al. 2015).

Statistics report 43% of users search for products online through SM networks[3] indicating a new platform for private organizations to promote their brand and reach out to customers across the globe. For example, LinkedIn provides

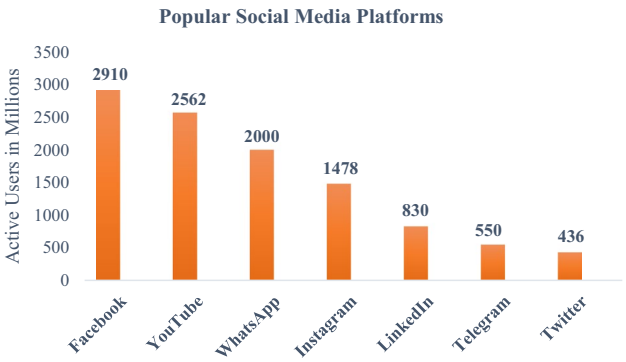


Fig.1 Statistics of monthly active users on various social media platforms (https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/)

[1] https://online.maryville.edu/blog/evolution-social-media/.

[2] https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide.

[3] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

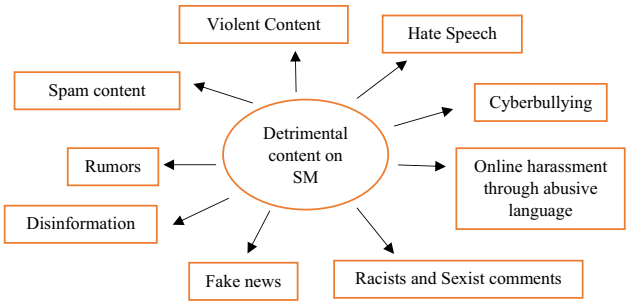


Fig. 2 Various forms of Detriment content published on SM

a platform for business to business and industry connectivity, career development activities, and job opportunities. There are also anonymous social networking mobile applications like Whisper where users post text messages and videos without revealing their identity.[4] The online social networking sites provide a platform for users to share their opinions on different aspects of social, political, economic, ethical, environmental issues in real time. The content called as User Generated Content (UGC) (Wyrwoll 2014) is shared on these platforms in form of text messages, images, videos, memes, audio. The terms like posts, tweets, comments, reviews, retweets are associated with UGC (Wyrwoll 2014). The content generated by the user is at times positive and at times is detrimental. The content on SM platforms is gaining importance through its use for screening students to provide placement opportunities and also used in a negative way that is affecting a person's mental health and also resulted in loss to economy. Recent years has shown an influential rise in the UGC on SM platforms which is creating a profound impact on the society.

1.2 The dark side of social media

Social media platforms like Twitter, Facebook, Reddit, and Instagram are the popular and widely used platforms that enable people to access and connect to a boundless world by forming a social network to express share and publish information (Nagi et al. 2015). Recent years have shown a substantial increase in the usage of SM platforms due to fast, easy access to information and a freedom to express through various formats (Wyrwoll 2014, Ruckenstein and Turunen 2020). This freedom of expression (Leerssen et al. 2020) is used in an improper way through the creation and publication of UGC that is provocative, inflammatory and threatening. In recent years the world is experiencing the negative aspect of SM through sharing of a detrimental content that is increasing at huge rate. A detrimental content on SM refers to sharing and publishing content with an intention to harm or distress a person or a community. Figure 2 depicts the detrimental form of UGC which includes hate speech content (Ayo et al. 2020), fake news, rumors (Shu et al. 2017), cyberbullying (Ofcom 2019), toxic content and child abuse material (Ofcom 2019) shown in Fig. 2. The definitions of the various forms of the detrimental/harmful content with an example content published on SM are depicted in Table 2. The term "fake news" on SM became prominent during US presidential election 2016. During the election period one of contenders made a speech: "The epidemic of malicious fake news and false propaganda that flooded social media over the past year. It's now clear that so-called fake news can have real-world consequences (Wendling 2018). As shown in Table 2, fake news, clickbait, rumors, satire news all come under misinformation (Islam et al. 2020) defined in context of two characteristics (Shu et al. 2017, Zhou et al. 2020):

(i) **Authenticity**: news that are non-factual or false which needs to be verified.
(ii) **Intent**: fake news is created with a wrong intention to mislead the users.

The authenticity characteristic cover disinformation, rumors, satire news and misinformation terms of fake news while the intent characteristic cover only disinformation and rumors. The current COVID-19 pandemic resulted in two million messages posted on Twitter with 7% of the total messages spreading conspiracy theories about the corona virus between 10 January 2020 and 20 February 2020 (Colomina et al. 2021).

Research studies have reported various definitions of hate speech like hate speech targets on a specific groups like ethnic origin, religion, or other, hate speech incite violence or hate toward minority, offensive and humorous content (Fortuna and Nunes 2018; Schmidt and Wiegand 2017). As shown in Table 2, hate speech content covers a broad spectrum of user created insulting words which are explored in various research works (Schmidt and Wiegand 2017). In many research articles, offensive content is also termed as abusive. Research articles have also reported the use of profane words in cyberbullying and hate speech content (Malmasi and Zampieri 2018). According to Pew Research survey 2018 conducted for teenagers, one in six teenagers have experienced one of the following forms of online abusive behavior as shown in Fig. 3.

The potential risk of SM has impacted the mental health of young generation in form of addiction, attention deficiency, aggressive behavior, depression, suicides (Ngai et al. 2015). According to National Crime Records Bureau (NCRB) data, cybercrimes are also increased on SM. In India, there are 578 cases of fake news on SM, 972 related

[4] https://en.wikipedia.org/wiki/Whisper_(app).

Table 2 Definition of various forms of inappropriate content published on SM

Name of the term	Definition	Example
Fake News	A news article that is intentionally false (Shu et al. 2017)	"*You see suicide rates are skyrocketing now*" (Patwa et al. 2021)
Deceptive News/ Disinformation	Deceptive news is articles with no correct facts, but articles are shared with authenticity (Shu et al. 2017, Zhou et al. 2020)	*Drinking hot water, cow urine, methanol or alcohol has been recommended as a proven cure for COVID-19* (Naeem et al. 2021)
Satire news	Satire news is form of fake news to attract the users with content written in a humorous and exaggerated way (Shu et al. 2017, Zhou et al. 2020)	*New UNL president a giant sea man* (Li et al. 2020)
Rumor	A piece of information that is shared on social media without being verified (Shu et al. 2017, Zhou et al. 2020)	*Saudi Arabia beheads first female robot citizen* (Islam et al. 2020, Ma et al. 2019)
Clickbait	Clickbait refers to attention grabbing form of news headlines on the social media (Shu et al. 2017, Zhou et al. 2020)	*This Rugby Fan's Super-Excited Reaction To Meeting Shane WilliamsWill Make You Grin Like A Fool* (Chakraborty et al. 2016)
Hate Speech:	The code of conduct as stated by European Union Commission: All public incitement to violence or hatred directed at a group of people or a member of that group based on race, color, religion, descent, nationality, or ethnicity (Fortuna and Nunes 2018)	*Refugees should face the figuring squad!* (Fortuna and Nunes 2018)
Cyberbullying	A form of harassment through electronic medium like mobile phone, computers conducted with an intention by a group or an individual against a person by sharing humiliating, wrong messages about him. This is a more general form of hate speech. (Fortuna and Nunes 2018)	*As long as fags don't bother me let them do what they want* (Dinakar et al. 2011)
Profanity	A sentence or a text with consists of offensive words or phrases. (Schmidt and Wiegand 2017)	*Holy shit, look at these ****** prices... damn!* (Malmasi and Zampieri 2018)
Toxic language	The toxic language is used in form of comments which include rude, disrespectful or unreasonable messages that can make other users to exit a discussion (Fortuna and Nunes 2018)	[a]*Ask Sityush to clean up his behavior than issue me nonsensical warnings*
Abusive language	A hurtful form of language that uses insulting or accusing words to someone but not targeted to a particular race, religion or ethnicity (Nobata et al. 2016)	*Add another JEW fined a billion for stealing like a lilmaggot. Hangthm all* (Nobata et al. 2016)
Sarcasm	Sarcasm is a form of ironic speech targeted to a particular victim to criticize him in a humorous way (Nobata et al. 2016)	*Most of them come north and are good at just mowing lawns* (Dinakar et al. 2011)

[a]https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

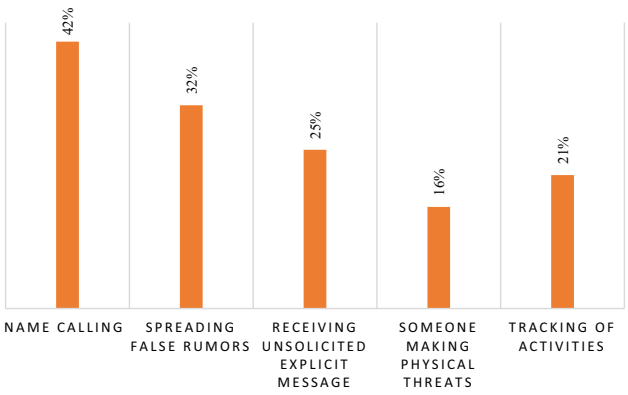


Fig. 3 Online abusive behavior experienced by teenagers

to cyber bullying of women and children, and 149 incidents of fake profile as reported in Times of India 2020. The recent example of COVID-19 pandemic has resulted in 80% of users reading fake news about the outbreak of the corona virus.[5] There were 7 million fake news stories, 9 million content encouraging extremist organization and 23 million hate speech content that were removed by SM companies during COVID-19 pandemic. This forced the European Commission (EU) to frame the policies to tackle the growing online threats and misinformation. The World Health Organization (WHO) reported that the citizens around the globe were victims of pandemic and "infodemic" that came up along with it 2020 (Colomina et al. 2021; Nascimento

[5] https://www.thenews.com/ Interesting statistics about fake news on social media/print/893091.

Table 3 Legal provisions to tackle the misuse of SM

Name of legal provision	Points covered in the provision	Remarks
The Information Technology (Amendment) Act, 2008 Sect. 66A (Mangalam and Kumar 2019)	Prohibits sending of offensive messages through an online medium	Declared 'unconstitutional' in 2015 due to lack of interpretation of terms like 'grossly offensive', 'insult', menacing' in the Sect. 66A
	Person sending information for the purpose of causing annoyance, obstruction, hatred or ill will through any digital platform is punishable with imprisonment for more than 3 years and with fine	Supreme Court found it violative of right to freedom of expression under Article 19 (1)
The Information Technology (Guidelines for Intermediaries and Digital Media Ethics Code) Rules, 2021[a]	Guidelines for SM intermediaries to identify the 'originator' of 'unlawful' messages	Differentiates between social media intermediary and significant social media intermediary
	Intermediaries shall remove or disable access within 24 h of receipt of complaints of sexual act morphed images etc	Challenged by SM companies as limit to freedom of expression

[a]G.S.R. 139(E): the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021.https://www.meity.gov.in/content/notification-dated-25th-february-2021-gsr-139e-information-technology- intermediary

et al., 2022). The definition of hate speech is subjective and varies with context in which the words are used in the content and is highly dependent on the geographic location.

1.3 Legal Provisions made by Government and SM companies to tackle the detrimental content

To curb with the increasing detrimental content on SM, Government has made legal provisions for example IT ACT 2000 law in India to deal with cybercrime and electronic commerce. The legal provisions defined by Government of India are summarized in Table 3.

As shown in Table 3, the compliance with the points defined in the legal provisions is challenging in view to safeguard the right of freedom of speech and expression of an individual on SM and a need to define what form of content is offensive or insulting. The Guidelines for Intermediaries and Digital Media Ethics Code) Rules show stringent guidelines for intermediaries in terms of taking down the 'unlawful' messages within a specific timeframe and providing information related to originator of such message and verification of identity to authorized agencies within 72 h.[6] This aspect though may help to control the spread of such messages but will be taken into account after such'unlawful' messages are flooded on SM and a damage is caused in the society. Government has also framed the legal rules and policies for SM companies that need to be implemented when an objectionable post published on online platforms. The rules are also defined for SM companies when an objectionable posts results in disturbing incidents and cause damage. The SM companies take counter actions by either removing or deleting the posts or by blocking the account of

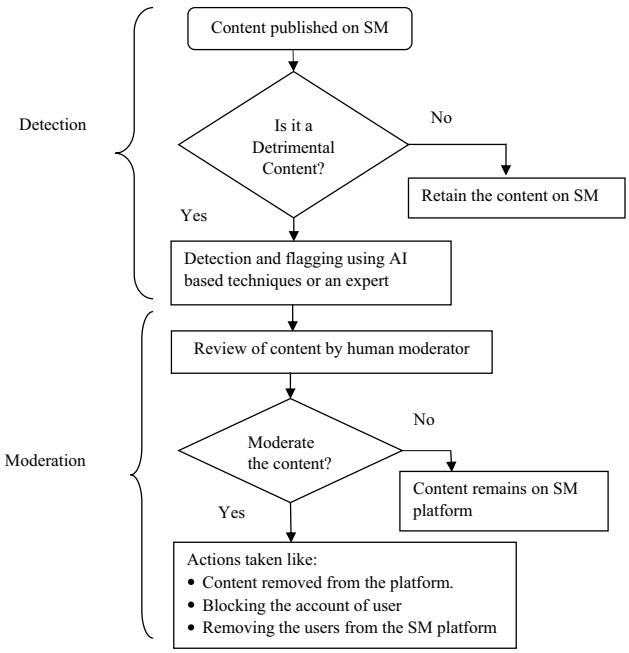


Fig. 4 Detection and moderation of UGC on SM platforms

the user who published the posts (Roberts 2017b). For example, Twitter platform received 1698 complaints pertaining to online abuse/harassment (1366), hateful conduct (111), misinformation and manipulated media (36), sensitive adult content (28), impersonation (25) in India via its local grievance mechanism between April 26, 2022 and May 25, 2022.[7] The action taken by Twitter is either in form of removing the accounts or banning the accounts that promote such activities.

[6] G.S.R. 139(E): the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021.https://www.meity.gov.in/content/notification-dated-25th-february-2021-gsr-139e-information-technology- intermediary.

[7] https://www.business-standard.com/article/companies/twitter-says-it-has-banned-over-46-000-bad-accounts-in-india-in-may-122070300540_1.html.

1.4 Detection and moderation of detrimental content on SM

Considering the huge volume of UGC on various SM platforms, detection and moderation of detrimental content on SM is of paramount importance. When a content published on SM platforms, it is detected to identify or classify whether the published content is harmful or non-harmful. Figure 4 depicts the steps of UGC detection and moderation on SM platforms. Detection is a task of classifying UGC as a normal content or an inappropriate content. Detection method entails identifying: the slur or slang, abusive, profane words in the content and the fake news in the content, and checking whether the content is targeting to a particular community or an individual. Artificial Intelligence (AI) has emerged as an upcoming tool for automated detection of detrimental content on SM through Machine Learning (ML) algorithms and Natural Language Processing (NLP). The use of AI-based detection methods assists the human moderators in flagging the content. UGC moderation on SM platform is the systematic screening of User Generated Content (UGC) provided to websites, SM, and other online networks to determine the content's acceptability for a specific site, location, or jurisdiction (Roberts 2017a). Moderation is about making a decision about the checking and verifying the adequacy of the detected content according to the rules and policies as defined by a particular SM platform. So, moderation is with respect to a specific SM platform. For example, a dance video published on LinkedIn is unacceptable as it is a professional SM platform with emphasis on building a network of professionals from various industries across the globe. The same dance video is acceptable on Facebook as it promotes sharing of individual user content in various forms. So, content moderation is more dependent on SM platform.

1.5 Organization of the paper

The paper is organized as follows: Sect. 2 describes the Review methodology used for presenting the paper. Section 3 presents the datasets created by research community for UGC detection on SM platforms. Section 4 provides the UGC detection and next section presents UGC moderation. The article concludes with conclusion and directions for further research.

2 Review methodology

A systematic method of reviewing the available literature is adopted to explore the work done by researchers in the field of SM content moderation. The methodology of literature review is divided into following steps:

- Defining the research questions
- Collection of relevant topics from the scientific literature and recent articles.
- Mapping the information collected from the literature to the research questions.

Figure 3 shows the flow diagram of the selection process of research articles for the review. With an objective of understanding the SM content detection and moderation, an ordered process of search is utilized with research articles collected from various fields of social sciences, computational intelligence and technology. The literature survey for the study was restricted to the articles published during the year 2011–2021. With reference to the objective of the study, the first step consists of collecting the articles from IEEE, Springer, Elsevier and AAAI digital libraries and Google Scholar. Since Google Scholar consists of articles from all publishers, including Arxiv, duplicate articles were excluded. A total of 500 articles related to social media content were screened by reading the abstract of the article and the maximum number of citations received for the article. The process of collecting the articles by exploring the literature in domain of social sciences with keywords like "Content moderation on social media", "User generated content on social media", and "Need of content moderation" in the digital library database. This research paper focuses on detection and moderation of detrimental content on social media, so after giving this query on Google scholar resulted in articles related to detection of hate speech, fake news, rumors and cyberbullying content. On this basis, queries like "Detection of harmful/problematic social media content using Natural Language Processing", "Machine learning and Deep Learning algorithms for Hate Speech/Fake news/rumors", "NLP for Hate Speech/Fake news/rumors detection", "Hate Speech/Fake news/rumors detection using machine learning and deep learning techniques" were investigated on digital libraries. For query related to social media content moderation, majority of the articles were extracted from social science domain. Considering detection and moderation of SM content, a total of 125 articles were selected in this study (Fig. 5).

2.1 Research objectives

The study presents an exhaustive survey of research done in SM content detection and moderation techniques. The key research objectives of the study are to:

- Outline the various forms of detrimental contents like rumors, fake news, hate speech, abusive content which exemplify the inappropriate use of SM.

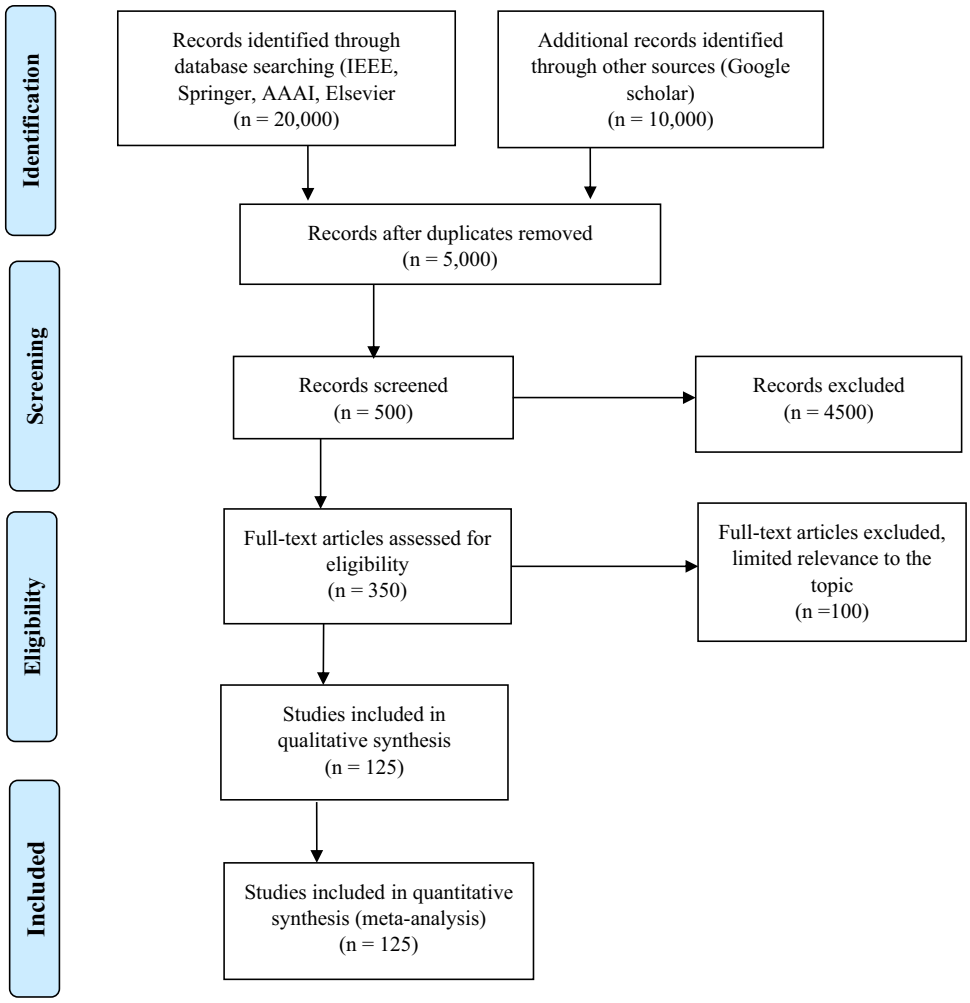


Fig. 5 Flow chart of selection of articles for review

- Review the datasets used for detection of detrimental content.
- Perform a comparative analysis of various Language Models (LM) and Machine Learning (ML) algorithms used for detection of detrimental content on SM platforms.
- Review the moderation techniques of detrimental content.
- Identify the challenges and research gaps of various reported techniques for UGC detection and moderation.

2.2 Research questions

The following research questions are framed to meet the research objectives.

Which datasets are used for detrimental content detection techniques?

What are the various methods to detect detrimental content on social media platforms?

What is content moderation and approaches to content moderation on social media platforms?

What are the challenges and research gaps in the reported techniques for content detection and moderation?

The paper is organized with the sections corresponding to meeting the defined objectives and answering the framed research questions.

2.3 Theoretical and practical implications of study (Cunha et al. 2021)

The literature review shows that there is a massive amount of research explored in the detection methods of various forms of detrimental content. From theoretical point of view, reported articles have focused more on the various aspects

Table 4 Popular datasets for fake news detection

Dataset	Features	Categories of labels assigned to articles	Skewness (Cunha et al. 2021)
LIAR (Wang 2017)	First dataset for deception detection	False (pants-fire)	Highly imbalanced
	12.8 K human labeled short statements evaluated PolitiFact.com	False	
		Barely true	
		Half true	
	Statement collected from news releases, TV/radio interviews, campaign speeches, TV ads, tweets, debates, Facebook posts, etc	Mostly true	
		True	
BUZZFEED NEWS[a]	2000 news samples published on Facebook during 2016 US Presidential elections	Mostly true	Highly imbalanced
		Not factual content,	
	Each post and linked articles were checked by 5 journalists	Mixture of true and false	
	Metadata information such as URL of the news post, published data, number of shares, reactions and comments	Mostly false	
CREDBANK (Mitra and Gilbert 2015)	60 million tweets from Twitter which covers 1049 real-world events	Real	Imbalanced
	Credibility verified by 30 annotators from Amazon Mechanical Turk	Fake	
FAKENEWSNET (Lee et al. 2018)	Highlights on dynamic context and social behavior of fake news	True	Balanced
	211 fake news and 211 true news	Fake	
	Data like publisher information, news content, and social engagements information gathered from fact checking websites BuzzFeed.com and PolitiFact.com		
Fake News Challenge-FNC-1[b]	Used for stance detection method with 50,000 stances, which targets on estimating the stance of a body text from a news article relative to a headline	Agrees	Highly imbalanced
		Disagrees	
		Discusses, or	
		Unrelated	
ISOT (Ahmad et al. 2017)	Real and fake news articles	Real-21417	Slightly imbalanced
	Real news gathered from crawling website Reuters.com	Fake-23481	
	Fake news from unreliable websites flagged by Politifact and Wikipedia		
	Include articles of political and World news topics		

[a] https://github.com/BuzzFeedNews/2016-10-facebookfact-check/tree/master/data

[b] https://github.com/FakeNewsChallenge/fnc-1

involved in terms of manual method of moderation and challenges that the AI based methods should address. There are less research articles that focus on fully automated moderation techniques of detrimental content on social media platforms. From practical point of view, more experimentations are done on language models, non-neural and neural network models for detection of detrimental content.

3 Datasets

Datasets form an important repository which contains information in form of a table. In context of detrimental form, the information in the datasets includes news articles, URLs, slang words, publisher information, social engagements, tweets gathered from social media platforms. Various ML algorithms are experimented on the available datasets for detection of fake news, hate speech and its related terms.

The datasets for fake news are prepared by extraction of online comments or posts from various social media platforms. The datasets are created with help of language experts and experts from field of journalism. The human experts analyze the posts and comments and assign labels to them as fake and real. Table 4 compares the list of features that can be extracted from the available datasets for fake news detection. As seen from Table 4, most of dataset's

target on the content features of the news, which might be not sufficient for an effective detection of fake news. Datasets like BuzzFeed News and FNC-1 and Fake News Net include metadata information and also the news content features which are explored in many research articles. The metadata information includes social network information, user's engagements in the news, users' profiles, etc. (Shu et al. 2017). LIAR dataset has considerably huge statements as compared to other datasets and also include meta-data information of each speaker (Wang 2017). The LIAR dataset also covers diverse subject topics like economy, healthcare, taxes, federal-budget, education, jobs, state budget, candidates-biography, elections, and immigration. Some datasets also assign labels to news articles to enable have a multi-level classification.

Table 5 summarizes the datasets for various forms of hate speech. The datasets of hate speech contain monolingual and multilingual content and also include score labels (Davidson et al. 2017) assigned to each characteristic of hate speech. The annotation of hate speech is done by different annotators; the evaluation of which is done by a metric called inter-annotator agreement (Nobata et al. 2016; Davidson et al. 2017; Kocoń et al. 2021). The inter annotator agreement defines the number of annotators that agree on a particular task of annotation (Kocoń et al. 2021). Fleiss's Kappa (κ) is a statistical metric that specifies the annotators rating for assigning a label to content (Davidson et al. 2017) and Krippendorff's alpha (Singhania et al. 2017) deals with annotations that are missed. These two measures are utilized for datasets with a high value of this measure signifies higher level of agreement. For example, Vigna et al 2017) reported a $\kappa = 0.26$ for 1687 comments annotated by 5 annotators for 2 classes of hate speech: weak hate and strong hate which shows the difficulty in annotation process. Nobata et al. (2016) reported a $\kappa = 0.26$ for 56,280 abusive comments annotated by 3 expert raters. Waseem et al. (2016) reported a $\kappa = 0.84$ with 85% disagreement for annotations of sexism. Due to highly subjective nature of hate speech, the inter-annotator agreement process becomes too challenging. Many research studies reported creation of datasets that assign labels as offensive, abusive, profanity, racism, sexism and general hate. As seen in Table 5, the skewness level (Cunha et al. 2021) of only few datasets is balanced. For example, the hatEval (Basile et al. 2019) dataset include 43% as hate content and 57% as non-hate content. Davidson et al. (2017) reported 5% as hate speech and 76% as offensive language. The label "relation" in (Bonet et al. 2018) indicates a hate speech sentence when it is combined with other sentences and label "skip" signifies a non-English sentence or a sentence with a hate or non-hate speech.

The inter-annotator agreement plays a vital role in creating the datasets for hate speech as it affects the performance of a ML algorithm (Kocoń et al. 2021). In context of fake news and hate speech, Twitter is the preferred social media platform for extracting information and preparing a dataset. The creation of datasets is dependent on the annotator's perspective of assigning a label and context information about the content. With a tendency of a user to write a post in multilingual form and code-mixed form (native language written in Roman script), research community have also created datasets in code-mixed language (Hindi + Engish) (Mathur et al. 2018) which are used for detection of hate speech using ML and neural network architectures. The annotation of such form of content is done with human annotators and inter annotator agreement is calculated. As shown in Table 5, there are too diverse variations in the datasets of hate speech, for example: "aggressive" word with labels like covertly and overtly aggressive and "hate-inducing" word. The multiple labels assigned to the text in the datasets, the size of datasets and skewness level affect the performance of ML algorithms and deep neural network models. Research articles have reported few questionable and doubtful cases (Mathur et al. 2018) of hate speech that were too challenging for human annotators to decide. Such cases were not considered in the dataset. Such uncertain cases need to be addressed in the dataset.

4 Detection of detrimental UGC on SM

Detection is a task of identifying the detrimental or objectionable content from the posts or text messages published by users on SM platforms. Detecting the detrimental content includes identifying the fake news, hate speech, abusive language content in an online post. Before moderating the content on SM platforms, it is first detected. Considering the amount of content published on SM platforms, (for example: An average 6000 tweets are posted every second on Twitter[8]), manual method of detection is not scalable. Artificial intelligence (AI) has emerged as an important tool for identifying and filtering out UGC that is offensive or harmful. Various techniques of AI in form of ML algorithms, DL, and Natural Language Processing (NLP) are deployed for detection of detrimental UGC (Ofcom 2019; Grimmelmann 2015). Research articles have reported that the AI based tools have achieved optimal accuracy and speed in detecting the detrimental content on SM platforms. This section describes the manual and AI-based methods of detecting detrimental content on SM platforms.

[8] https://www.internetlivestats.com/twitter-statistics/.

Table 5 Popular datasets for hate speech detection

Name of Dataset	Features	Categories of labels assigned to articles	Skewness (Cunha et al. 2021)
Davidson et al. (2017)	24,802 tweets from Hatebase	Hate speech-7%,	Highly imbalanced
	Contain large number of ethnicity content	Not offensive-,	
	Collection on offensive keywords	Offensive but not hate speech	
Stormfront (Bonet et al. 2018)	Textual hate speech annotated at sentence level	Hate,	Imbalanced
		No hate	
		Relation	
	10,568 sentences have been extracted from Stormfront	Skip	
ETHOS (Mollas et al. 2021)	Creation of two textual datasets using comments from Reddit and Youtube	Violence	Highly Imbalanced
		Directed_vs_generalized	
	First dataset includes 998 comments with two labels	Gender	
		Race	
	Second dataset includes 433 hate speech messages with 8 labels	National_origin	
		Disability	
		Sexual_orientation	
		Religion	
Hatebase[a]	Online repository of structured and usage-based hate speech	Archaic	Highly Imbalanced
		Class	
	Used to build a classifier for hate speech	Disability,	
		Ethnicity,	
		Gender,	
		Nationality,	
		Religion	
		Sexual orientation	
HASOC 2019 (Mandl et al. 2019)	Three datasets from Twitter and Facebook	Hate and offensive	Imbalanced
		Profane	
	German, English and Hindi language	Non- hate and offensive	
CONANN (Chung et al. 2019)	Expert based hate speech and counter narrative content	Hate speech	Imbalanced
	4078 pairs over the English, French and Italian language	Counter hate speech	
	Expert demographics, hate speech sub-topic and counter-narrative type		
Waseem and Hovy (2016)	136,052 tweets from Twitter Labeling of 16,192 tweets	Racist	Imbalanced
		Sexist	
		None	
Waseem and Hovy (2016b)	136,000 tweets from Twitter Annotations by experts (feminists and anti-racism activists) and crowdsource workers	Racist	Highly Imbalanced
		Sexist	
		None	
		Both	
TRAC (Ojha et al. 2018)	15,000 instances from Twitter and Facebook	Overtly aggressive	Highly imbalanced
		Covertly aggressive	
	4 different datasets for English and Hindi language	Non-aggressive	

Table 5 (continued)

Name of Dataset	Features	Categories of labels assigned to articles	Skewness (Cunha et al. 2021)
GERMEVAL (Ross et al. 2016)	5009 tweets from Twitter in German	Offensive	Imbalanced
	Shared task with binary classification and fine-grained classification	Profanity	
		Abuse	
		Other	
SemEval 2019 hatEval (Basile et al. 2019)	13,240 tweets from Twitter	Hateful	Imbalanced
	Hateful content against immigrants and women in English and Spanish	Aggressive	
KAGGLE[b]	8832 social media comments	Insulting	Imbalanced
		Non-insulting	
Golbeck et al. (2017)	20,362 tweets from Twitter	Positive Harassment	Imbalanced
		Negative Harassment	
HOET (Mathur et al. 2018)	3679 tweets in Hindi-English code-switched language	Not Offensive	Highly Imbalanced
		Abusive	
Hindi-english offensive tweet		Hate-Inducing	

[a] www.hatebase.org

[b] https://kaggle.com/c/detecting-insults-insocial-commentary

4.1 Manual method of fake news detection

Fact checking is a detection method that decides of whether the published content is real or fake (Barrett 2020). Fact checking does not evaluate an objectionable content, but classifying whether the content is true or false (Barrett 2020).

Table 6 depicts the fact-checking websites. Fact checking websites make use of human experts in the journalism domain that check the veracity of the news content.

The experts are called as fact checkers that follow a methodology to evaluate a content. The methodology utilized by fact-checking websites includes:

(i) By skimming through news items, political commercials and speeches, campaign websites, social media, and press releases, TVs, and interviews, a topic or a claim to be examined is chosen.
(ii) Fact checkers most typically employ fundamental methodologies and types of sources while conducting research on assertions, as well as official regulations and editorial norms that govern their approaches.
(iii) Claim assessments, which are systems and processes used by fact-checkers to determine the validity of a claim.[9]

The fact checking website like Politifact[9] has developed datasets and made it publicly available for automatic detection of fake news content. These websites provide an expert analysis for checked news as which news articles are fake and reason for why it is fake (Zhou and Zafarani 2020). SM platforms like Facebook sends flagged content to more than 60 fact-checking organizations worldwide, but each organization typically assigns only a handful of reporters to investigate Facebook posts (Barrett 2020). The manual method of checking the facts for detecting the fake news is a complex task. Factors like time needed to check the veracity of the news and the knowledge of the context around the fake news need to be considered in the detection task.

The detection of other forms of detrimental content like hate speech and abusive language is done by the user community to express their concern about the content posted on SM platforms (Gillespie 2018, Crawford and Gillespie 2016). There is risk of bias getting introduced by the user in the detection of such content. With overflowing increase in detrimental content, the manual method of detection will not be adequate.

4.2 Detection of detrimental UGC using natural language processing (NLP)

The manual approach of fake news detection has many challenges in terms of the volume, veracity and speed of content to be analyzed, the cultural, historical and geographical context around the content. Many companies and governments are proposing automated processes to assist in detection and analysis of problematic content, including disinformation, hate speech, and terrorist propaganda (Leerssen et al.

[9] www.politifact.com.

Table 6 Fact-checking websites

Fact Checking website	Description
Politifact.com[a]	US-based website for fact checking of political news and information
	Provide labels to the claims and statements as True, Mostly true, Half true, Mostly false, False, and Pants on fire
	Experts verify the creditability of the statement and claims by examining a specific word and the context of the claim
Snopes.com[b]	First online fact checking website considered by many journalists and researchers for verifying internet rumors and misinformation
	Website covers various subjects like medicine, science, history, crime, frauds, etc
	Websites provide a comprehensive evaluation of various types of printed resources and assign truth ratings to them based on the knowledge from professional individuals and organizations
FactCheck.org[c]	Website evaluates the truthfulness of the facts and claims made during the election years by U.S. political players on various platforms like television, social media, speeches and interviews
	With help of experts, each piece of information is checked and analyzed in a systematic manner
Hoax-slayer.net[d]	Website has debunked email and internet hoaxes, thwarted scammers, educated web users about security issues and combated spam based on meticulous research on the information gathered from news articles, press release, government publications, reputed websites
	Websites include articles that include hyperlinks and reference list that allow the reader to check the information for themselves

[a] www.politifact.com

[b] www.snopes.com

[c] https://www.factcheck.org

[d] www.hoax-slayer.net

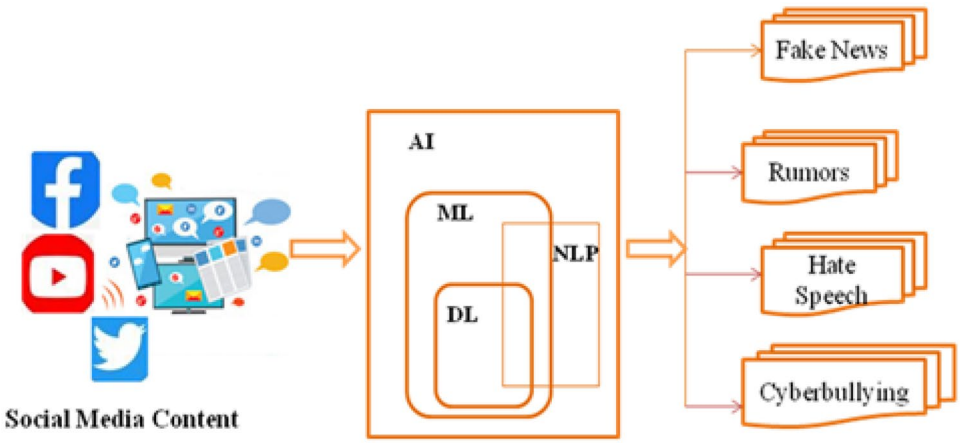


Fig. 6 AI-based techniques for detection of detrimental content on SM platforms

2020). Past decade has shown significant developments in AI through the advances in the algorithms, computational power and data (Ofcom 2019). Deep Learning (DL) is a subfield of ML that makes use of Artificial Neural Networks (ANN) to process huge amount of data. Natural Language Processing (NLP) is a subfield of AI that uses techniques to parse the text using computers (Hirschberg and Manning 2015). Natural Language Processing (NLP) is a computational linguistic field which makes use of computational techniques to learn and understand human language (Hirschberg and Manning 2015).

ML, ANN and NLP are the key components that have contributed to automated detection of detrimental form of SM content. Figure 6 shows the AI based approach of detection of detrimental content on SM platforms. A large volume of research has been explored on use of AI based techniques for detection of fake news, rumors, abusive/offensive language, and hate speech on SM platforms. The task of automated detection of UGC using NLP, ML and DL algorithms consists classifying the online comments/posts as detrimental (which include hate speech, abusive, toxic, rumors, cyberbullying) or a normal content. NLP has opened new spectrum of automating the linguistic structure of language

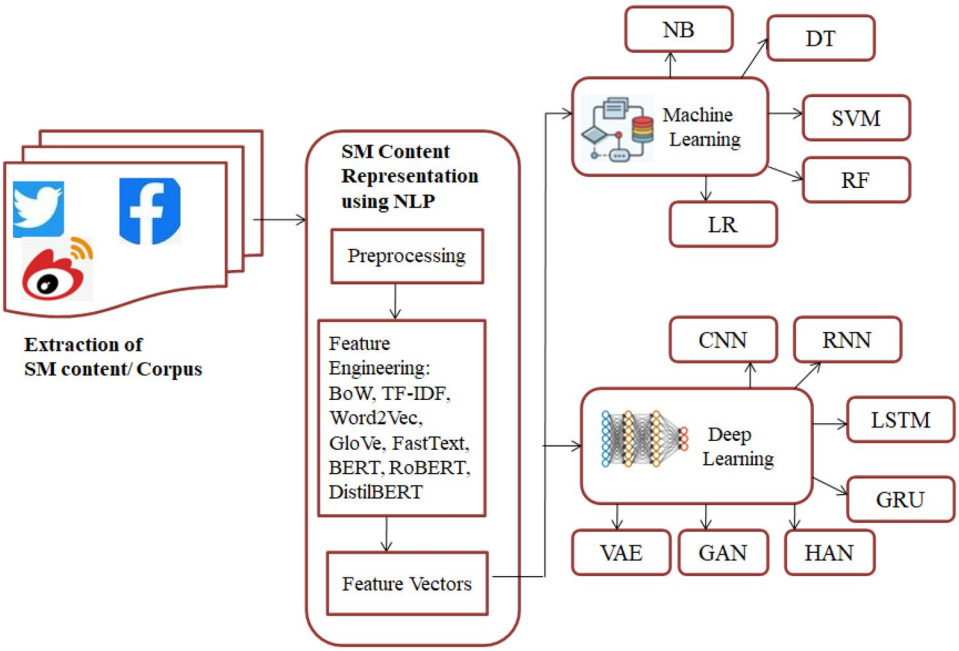


Fig. 7 A generic block diagram of automated SM content detection using NLP, ML and DL

in creation of speech-to-speech translation engines, mining SM for information about health or finance, and identifying sentiment and emotion toward products and services (Hirschberg and Manning 2015), filtering offensive content, and improving spam detection (Duarte et al. 2017), creation of chatbots for customer service (Ofcom 2019).

The noteworthy advancements in NLP have played a major role in detection of detrimental content on SM platforms. NLP tools are widely used to process the text-based online comments on SM (Ofcom 2019). In context of content moderation, NLP techniques are used to process the online text, extract the features from text which are used to detect the harmful forms content like fake news, hate speech, cyberbullying.

Recent years have shown advancements in NLP tools working as text classifiers that use neural networks and ML to analyze the features of text and classify the text into one of the categories of detrimental content and normal content (Duarte et al. 2017). Considering the amount of SM content, analysis of content using NLP includes quantitative and qualitative analysis. Quantitative analysis makes use of statistical measures like counting the frequency of words in content. Qualitative analysis investigates the meaning and semantic relationship of words and phrases in the content. Figure 7 depicts a generalize block diagram of UGC detection. NLP tools are deployed to process the online content published on the SM platforms. As shown in Fig. 7, the extraction of SM content comprises of acquisition of online comments and posts through Application Programming Interface (API) and crawling methods provided by SM platforms. For example, Twitter provides two tools namely the Search and Streaming API to collect the data (Ayo et al. 2020). A corpus is created that covers all diverse forms of SM content in monolingual and multi-lingual configuration with metadata information like geographical location, user profiles and followers (Schmidt and Wiegand 2017; Duarte et al. 2017).

This corpus is created with help of experts and crowdsource workers that assign labels to content as a normal one or harmful one (Roberts 2017a). The corpus thus created is called as dataset and researchers have made a significant contribution in creation of dataset that covers all terminologies of detrimental content like fake news, rumors, hate speech and cyberbullying content. The comment features are extracted from the corpus using NLP tools. The features can be words, phrases, characters, unique words (Schmidt and Wiegand 2017; Ahmed et al. 2017) that differ depending on the form of content to be processed. Many feature representation techniques like Bag of Words (BoW), Term Frequency-Inverse Document Frequency, n-grams (Schmidt and Wiegand 2017; Ahmed et al. 2017), Word2Vec (Mikolov et al. 2013), GloVe (Pennington et al. 2015), Bidirectional Encoder Representation from Transformer (BERT) (Vaswani et al. 2017; Devlin et al. 2019) map the text features from the content to vectors of real numbers known as feature vectors.

The feature vectors obtained after processing the SM content using NLP tools are applied to a classifier model which can be either a non-neural model or a neural model (Cunha et al. 2021). Classifier models are used to detect detrimental content based on features extracted from SM content. Research literature reports the use of supervised ML algorithms like Support Vector Machine (SVM), Logistic

Table 7 Pre-processing Steps (Vijayarani et al. 2015, Ahmed et al. 2017, Robinson et al. 2018, Elhadad et al. 2020)

Pre-processing step	Description
Conversion of the text to lower case	All words in the text are written in lower case to eliminate the difference between the letters of the same words
	The words "GOOD" and "gOoD" convey the same meaning but written in mixed casing style
	Lower casing of the entire text makes analysis easier and leads to reduction in dimensionality of data
Removal of Stop Words	Stop words are articles ('a', 'an', 'the') and prepositions ('for', 'on') in the text and convey no useful meaning
	Stop words are not considered as keywords in text analysis so they are removed
Stemming	Rule-based method of reducing inflectional form of a word to its base or root form
	Common stemming algorithms include Porter's Stemming, Dawson Stemmer, N-gram Stemmer
	Removal of suffixes from a word which reduces the dimensions of data in terms of space and memory
	For example, words like playing, played are related to a word "play" with same meaning
Lemmatization	Lemmatization identifies the base form of the word through morphological analysis
	Lemmatization always provides a dictionary meaning of a word
Tokenization	Tokenization is a method of breaking the raw text into tokens or words and is language dependent
	A Python tokenization package named "Twikenizer" is capable of dealing with special characters attached to the words

Regression (LR), Naïve Bayes (NB), and Random Forest (RF) and deep neural networks like non-Sequential neural network models: Convolutional Neural Networks (CNN) and sequential neural network models: Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), Transformer models, Variational Autoencoder (VAE) models, and Graph based neural networks for the detection and classification of detrimental SM content which predominantly include fake news and hate speech. The non-neural and neural network models (Cunha et al. 2021) are trained on various features extracted from the labeled datasets using various feature representation techniques. The trained network is applied on the test data for detection or classification. The classification can be a multi-class classification (e.g., classifying a content into offensive, hate and non-hate, Davidson et al. 2017) or a binary classification (e.g., classification of real and fake news, Ahmed et al. 2017). DL algorithms which work with huge amount of data offers a significant advantage of automatically discovering the features for classification which an ML algorithm does with human intervention (Ayo et al. 2020). Considering the amount of content published on SM, neural networks have proven to be an effective tool for automatic detection of SM content.

4.2.1 Role of NLP for detection of detrimental content om SM

The manual approach of parsing the vast volume of SM text is challenging in terms of time required to understand the unstructured and noisy text, training to the moderators to parse such text which is costly. Natural Language Processing is an automated tool to parse the text using computers (Hirschberg and Manning 2015; Duarte et al. 2017). NLP has made incredible advancements in text feature representation techniques through pre-trained generalized language models. The process of converting the raw text features into numerical feature vectors is achieved using various feature representation techniques which include frequency-based techniques and neural network-based word embeddings. Scientific research articles have reported the use of these techniques in detection of detrimental content on SM. An NLP pipeline in detection of detrimental content on SM consists of Pre-processing phase and Feature Engineering phase which are detailed as follows:

Pre-processing of the content Processing and analysis of the SM data comes under the field of data and text mining. Text mining is a process of extracting knowledge and information from an unstructured and noisy data (Vijayarani et al. 2015). Processing SM content is challenging task due to the unstructured form of UGC. The UGC on social media is often noisy and written in an informal way (Ahmed et al. 2017; Robinson et al. 2018) with sentences or texts lack in punctuations, use of more abbreviations, emoticons (e.g.,::-), special characters (e.g.,"@Sush","U9","#happy") and use of repeated characters (for example "cooooll", "haaa") in the text. This ambiguous form content makes text interpretation too challenging. So pre-processing forms a crucial step to transform such free form of content into a structured form in order to have an effective analysis of the UGC. The important pre-processing steps are detailed in Table 7. As shown in Table 7 stemming and lemmatization are similar, but lemmatization is preferred over stemming as it converts each word to its base form. Stemming and lemmatization are together called as normalization (Vijayarani et al. 2015).

The pre-processing steps summarized in Table 7 vary depending on the form of the content analyzed. For example, for fake news detection, the URLs, hyperlinks are important, however for hate speech detection they may not be of much significance. The pre-processing is performed using Python NLTK library. The profane words make use special characters like "g@y", "f**c" makes tokenization challenging (Robinson et al. 2018). The pre-processing of raw text facilitates selection of features and improves the performance of ML classifiers by reducing the dimensions of input words (vocabulary words) in the text thereby reducing the processing requirements and also selecting the features that are essential for classification.

Feature engineering Feature selection and representation together called as Feature Engineering form a noteworthy element contribute to the success of NLP text classifiers (Duarte et al. 2017). The features can be words, phrases, characters, unique words (Schmidt and Wiegand 2017; Ahmed et al. 2017) that differ depending on the form of content to be processed. The lexical, syntactic and semantic elements of text contribute to selection of features for SM content. The lexical elements are expressed at word-level lexicons in subjective, objective, formal or informal form (Verma and Srinivasan 2019). The syntactic elements refer to arrangement of words and phrases that define a sentence (Verma and Srinivasan 2019). The semantic elements include of identifying the attributes to extract the meaning of the sentence (Verma and Srinivasan 2019). The sentiments conveyed by the text can be analyzed through semantic elements. The additional features are also selected based on the meta-information accompanying the text. These include multimedia data and information about the users and its followers, geographical location which defines the environment about the content (Shu et al. 2017; Zhou et al. 2020; Fortuna and Nunes 2018; Schmidt and Wiegand 2017). In context of fake news and rumors, the lexical, semantic and syntactic features can be extracted from the news headline and main text of the news article. The images features can be extracted from image/video attribute (Shu et al. 2017; Zou and Zafarani 2020a). For hate speech content, the linguistic characteristics of the text define the features. A hate speech text is characterized by negative words (Schmidt and Wiegand 2017). An online hate message will consist of short length text, use of distinctive words that differentiate from a normal message, text with special characters, punctuation marks, user mentions etc. all from which the lexicon, syntax and semantic features can be extracted (Schmidt and Wiegand 2017; Watanabe et al. 2018; Robinson et al. 2018). The lexical, syntactic and surface features for fake news and hate speech content are similar in terms of use of words, typed dependency and use of special characters like hashtags (#), user handles (@), punctuation marks, etc. (Schmidt and Wiegand 2017; Zhang and Ghorbani 2020). For hate speech, word level features and sentiment features are considered to be important and is explored by many researchers. The use of emojis is widely used in hate speech content while the news headline forms an important feature for fake news. The feature selection method of NLP is extremely dependent on the type of SM content. The creator of news for fake news detection is used to determine the legitimate users and suspicious users (Shu et al. 2017; Zhang and Ghorbani 2020). The user profile features, user credibility features and user behavior features are deployed to determine the suspicious users which aid in detection of fake news (Zhang and Ghorbani 2020). Research has reported that the meta-information features are more important and is also exploited in detection of fake news whereas these features are considered not of much importance in hate speech detection. However, meta-information can be one of the important features for detection of certain ambiguous word content.

Feature representation is a technique of representing textual features which include words, phrases, and characters in a numerical form as shown in Fig. 7. The feature representation techniques assign a numerical value which indicates a frequency of word or a binary value which indicates the presence or absence of word in a text (Burnap et al. 2019). The numerical value represents a vector which is applied as input to ML algorithm for detection of words that are harmful. The character n-gram feature representation has shown improved performance as compared to word n- grams for noisy words like use of special characters in between the word (e.g., yrslef, a$$hole) (Schmidt and Wiegand 2017). Since Bag of Words (BoW) fail to understand polysemy word, it has shown high false positives for hate speech detection as reported in literature (Davidson et al. 2017). In some literature Parts of Speech (PoS) is considered as a pre-processing stage. BoW, n-grams and Term Frequency-Inverse Document Frequency (TF-IDF) generate the feature vectors based on frequency of words in text, there can be sparse representation of vectors due to short posts on social media which increase the memory and computational requirements. Table 8 shows the definition of the techniques. Frequency based feature representation techniques with supervised ML algorithms are used for detection of fake news, offensive content, profanity, clickbait on SM platforms. The sparse representation of feature vectors is addressed by using word embeddings. Word embeddings are pre-trained neural network based unsupervised word distribution models in which words in a huge corpus of unlabeled text are represented as numerical vectors (Schmidt and Wiegand 2017) resulting in high dimensional vector space.

The BoW technique that failed to extract the semantically similar words are addressed by word embeddings by creating the vector with values of semantically similar words placed close to each other (Mikolov et al. 2013). Research literature

Table 8 Feature representation techniques in NLP

Feature Representation	Description	Findings
Bag-of-Word (BoW) (Davidson et al. 2017, Robinson et al. 2018)	Each word is used as a feature with a numerical value representing the frequency of occurrence of a word in text	Fail to understand the polysemy words and fail to convey whether the word is a noun, verb or adjective
	Represent lexical features in the text	
N-grams (Davidson et al. 2017; Ahmed et al. 2017; Horne and Adali 2017)	Sequence of adjacent words or characters of length 'n' extracted from text	Fail to extract the correlations between the words that are a distance apart which would affect the performance of machine learning classifier
	The value of "n" can be 1 referred as unigrams, 2 referred as bigrams or 3 referred as trigrams	
	Character n-gram and word n-grams	
Term Frequency (TF)-Inverse Document Frequency (IDF) (Ahmed et al. 2017)	Statistical method to transform the words in a text into numeric representation	Semantic similarities between the words are not considered
	Term Frequency measures the frequency of words	
	Inverse Document Frequency indicates how important a word in a document is	
Parts-of-Speech (PoS) Tagging (Zhang and Luo 2019, Burnap et al. 2019)	Represent syntactic features that extract type dependencies by exploiting the grammatical relationships between the words in the text	Performance of machine learning classifier does not show improvement when n-gram features are combined with PoS information
	The long-distance words can be well capture with POS tagging	

Table 9 Word embeddings in NLP

Pre-trained Word Embeddings	Description	Number of words for pre-training	Dimension of vector	Findings
Word2vec by Google (Mikolov et al. 2013)	Feed forward neural network with one hidden layer	100 billion words from Google News	300	Ignore the morphological features of text and no vector representations for Out Of Vocabulary (OOV words)
	Creation of close vectors for semantically similar words			
CBOW by Google (Mikolov et al. 2013)	A continuous distributed representation that outputs a single word based on the context of neighboring words		300	Poorly utilize the statistics of the corpus and are not trained on global corpus
Skip-grams by Google (Mikolov et al. 2013)	A neural network that takes a single word as input and outputs multiple words based on the context of single word		300	
GloVe by Stanford (Pennington et al. 2015)	A log-bilinear regression model for the unsupervised learning of word representations	6B token corpus (Wikipedia 2014 + Gigaword 5)	300	Ignore the morphological features of text and no vector representations for Out Of Vocabulary (OOV words)
	Capture the global corpus statistics using co-occurrence matrix			
FastText by FaceBook AI team (Bojanowski et al. 2017)	Exploits a character level n-gram to represent a word	157 languages, trained on Common Crawl and Wikipedia	300	Long-term dependencies in the text cannot be learnt
	The generated vector of a word is the sum of its character n-grams with n ranging from 1 to 6			
EMLo by (Peter et al. 2018; Naseem et al. 2019)	Use of CNN to extract the raw characters	1 billion words with 800 M tokens of news crawl data from WMT 2011	1024	Used as augmentation with word2vec and GloVe
	Use of BiLSTM to represent context level word representations			Computationally intensive
BERT Google AI (Devlin et al. 2019, Vaswani et al. 2017)	Unsupervised language representation pre-trained on unlabeled text that considers the context of word from both sides of text	BooksCorpus (800 M words) and English Wikipedia (2500 M words)	768	Fails to work for generalized negation words
	Multi-layer bidirectional Transformer model			Computationally intensive during inference
RoBERT (Liu et al. 2019)	Similar to BERT	160 GB Text (16 GB of BERT + CommonCrawl News dataset (63 million articles, 76 GB), Web text corpus (38 GB) & Stories from Common Crawl (31 GB))	–	More training time than BERT
	Large batch training size			
ALBERT (Lan et al. 2019)	18 × fewer number of parameters as compared to BERT	English Wikipedia (2,500 M words) + Stories from Common Crawl (31 GB))	128	Reduced accuracy
	Reduced size of Embedding layer			Computationally intensive
	Parameter sharing achieved through one encoder layer applied each time on input			

Table 9 (continued)

Pre-trained Word Embeddings	Description	Number of words for pre-training	Dimension of vector	Findings
DistilBERT (Sanh et al. 2020)	40% less parameters than BERT	16 GB of BERT + 3.3 billion words	–	Suffer from biased predictions
	Use of knowledge distillation to red			
XLNET (Yang et al. 2020)	Auto-regressive language model	130 GB of textual data (33 billion words)	–	More Computationally and resource intensive
	Permutation Language modeling with all tokens predicted in random order			Longer training time than BERT
GPT-3 (Brown et al. 2020)	175 billion parameters	Common Crawl (410 billion tokens), webtexts (19 billion tokens), books (27 billion tokens), and Wikipedia (3 billion tokens)	12,888	Complex and costly inference
				Text generated by the model can introduce bias in language

CBOW Continuous Bag Of Words, *GloVe* Global Vectors for word representation, *BERT* Bidirectional Encoder Representations from Transformers, *EMLo* Embeddings from Language Models. *RoBERTa* A Robustly Optimized BERT Pretraining Approach. *ALBERT* A Lite BERT Self-supervised Learning of Language Representations, *GPT* Generative Pre-trained Transformer

reports the use of word embeddings have shown a significant performance improvement in the detection of SM content using ML algorithms.

Table 9 shows the widely used word embeddings in NLP. Pre-trained word embeddings preserve the syntactic and semantic information in the text (Pennington et al. 2014). Word embedding models are trained on huge corpus with various dimensions of word vectors. In word2vec models, the pre-calculation of vectors for words serves as a limitation for words that are non-grammatical. The contextual meaning of word within the sentence is not considered in word2vec model. This contextual understanding is considered in BERT and EMLo in which the vectors are calculated depending upon the context of word in the sentence. The real time calculation of vector representations has shown significant results in terms of accuracy in detection of SM content as reported in literature. BERT and EMLo are deep bidirectional language models that work on transfer learning (Pan and Yang 2010) concept are pre-trained on a corpus and are fine-tuned for a new corpus (Devlin et al. 2019). Both CBOW and skip-gram exhibit low computational complexity and can be trained on a large dataset; however, BERT and EMLo are computationally intensive indicating more response time. The feature vectors are applied as an input to a ML algorithm or a DL algorithm. As shown in Table 9, word embeddings are self-supervised pre-trained language models that are trained on large unlabeled dataset. Considering the amount of data (from 100 billion words to 130 GB of text data) the language models are trained on implies an increased number of hyperparameters (3000 of Word2Vec to 175 billion of GPT-3). This also signifies an increased training time to train the model and the number of computational resources required for training. For example, XLNet requires 512 TPUs and 2.5 days for training (Yang et al. 2020). Pre-trained language models are experimented for detection of fake news and hate speech on SM. Table 10 shows the use of language models for detection of detrimental content. As shown in Table 10, pre-trained language models perform better for fake news detection task and have reported low F1-score for hate speech detection task. However, BERT pre-trained on COVID-19 fake news dataset extracted from Twitter has reported highest F1-score.

This indicates that there is a need to create pre-trained language model that will consists of words and phrases that target on inflammatory or abusive words. The skewed nature of datasets also affects the performance of pre-trained language models. Malik et al. 2022 have experimented transformer models like small BERT (trained on less amount of dataset), BERT, ALBERT on three different datasets of hate speech and offensive language and compared the performance of these language models in terms of training time the model takes per epoch. The study reported that the training time of ALBERT language model is highest as compared to

Table 10 Language models for detection of detrimental on SM

Type of detrimental content	Dataset	Language Model	Performance metric	Findings
Fake News Glazkova et al	COVID-19 Healthcare Misinformation Dataset (Cui and Lee 2020)	BERT	F1- score: 96.75%	Misclassification of true posts about corona vaccine predicted false by the model
		RoBERT	F1- score: 97.62%	
		COVID-Twitter-BERT (CT-BERT)	F1- score: 98.37%	
Hate Speech (Zhou et al. 2020)	SemEval 2019 Task 5	EMLo	F1- score: 63.6%	Fusion method resulted in better F1-score
		$BERT_{base}$	F1- score: 62.3%	
		CNN	F1- score: 69.8%	
		Mean Fusion method of EMLo + BERT + CNN	F1- score: 70.4%	
Hate Speech and Offensive Language (Mutanga et al. 2020)	(Davidson et al. 2017)	BERT	F1- score: 73%	DistilBERT outperformed other transformer models
		RoBERT	F1- score: 69%	
		LSTM with attention	F1- score: 66%	
		XLNET	F1- score: 72%	
		DistilBERT	F1- score: 75%	
Satire news (Li et al. 2020)	4000 satirical and 6000 regular news articles	Vision & Language BERT (ViLBERT) (Lu et al. 2019)	F1- score: 92.16%	Model fails to extract relationship between text and image which results into misclassification

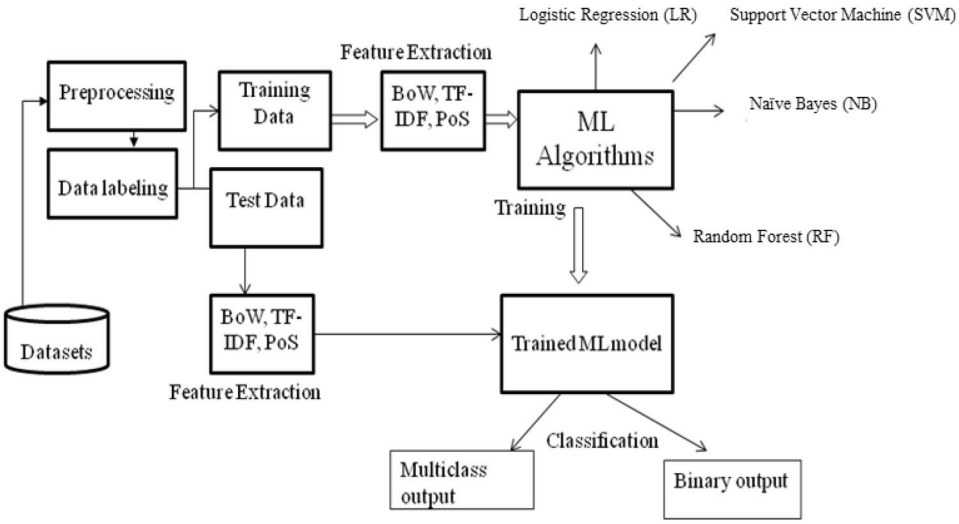


Fig. 8 Process of detection and classification of a SM content using ML algorithms

BERT and small BERT but ALBERT performed better in terms of F1-score (90%) than other models. The computational efficiency of language model in terms of training time is a crucial factor that needs to be considered for detection of detrimental content on SM.

4.2.2 ML and DL algorithms for detection of detrimental content on SM platforms

ML is a vital and largest subfield of AI that includes techniques to provide systems the ability to automatically learn and improve from experience without being explicitly programmed. Many subfields of AI are addressed with ML methods (Ofcom 209). Figure 8 shows the process of detection and classification of a SM content using ML algorithms. Research literature have reported the use of supervised ML algorithms like SVM, LR, NB, and RF for the detection and classification of SM content which predominantly include fake news and hate speech. The ML algorithms are trained on various features extracted from the labeled datasets using BoW, TF-IDF, n-grams feature representation techniques. The trained ML algorithm is applied on the test data for classification as shown in Fig. 8. The classification can be a multiclass classification for example classifying a content into offensive, hate and non-hate (Davidson et al. 2017) or a binary classification

Table 11 Neural network model for SM content detection and classification

Deep Neural Network model	Description
CNN (Ayo et al. 2017, Islam et al. 2020, Gambäck and Sikdar 2017)	Trained with word vectors using a fixed kernel size and number of filters
	1-Dimensional CNN (1D-CNN) is used to extract the local features from the text
	Extract the image features in multimodal approach of SM content detection
RNN (Nasir et al. 2021)	Sequential neural network with internal memory to process the short text
	For fake news detection, RNN is used to capture the temporal features of posts over time
LSTM (Ayo et al. 2017, Ruchansky et al. 2017)	Special type of RNN that learns the typed dependencies in the long text
	The memory unit consists of cell that use gates and a carry
	Carry responsible retaining the information during the sequential process
GRU (Ayo et al. 2017, Amrutha and Bindu 2019)	Addresses the short-term memory problem of RNN with gates
	Gates that decide the amount of information to be passed in the network and amount of information to be neglected
HAN (Singhania et al. 2017, Singh et al. 2020)	Makes use of an attention layer between the encoder LSTM and decoder LSTM
	Attention is given to each encoder input at each time step
GAN (Ma et al. 2019, Sahu et al. 2019, Islam et al. 2020)	An unsupervised learning method that generates new data
	Discriminative model used for classification
VAE (Khattar et al. 2019, Islam et al. 2020)	A generative autoencoder model that learns the latent state distribution of input data in a probabilistic manner
	Consists of encoder, decoder and a loss function
Capsule networks (Goldani et al. 2020)	Capsule indicates a group of neurons
	Text features extracted through n-gram convolutional layer
	Features are then processed in primary capsule layer, convolutional capsule layer and feed forward capsule layer
Transfer Learning (Pan and Yang 2010)	A knowledge transfer ML model in which learned features trained on one task are reused for learning another task through language models like BERT, RoBERT

for example classification of real and fake news (Ahmed et al. 2017). The performance of ML algorithm is evaluated on the datasets which contain a huge data extracted from popular SM platforms like Facebook, Twitter, Instagram, and Reddit. ML algorithms are considered as traditional algorithms for detection and of SM content. The handcrafted features used by ML algorithms are time consuming, incomplete, and labor intensive with the performance of a ML algorithm is dependent on the features selected for classification. Deep Learning (DL) a sub-field of ML has attracted the industry and academia for various applications. DL is basically a neural network with an input layer, one or more hidden layers and an output layer (Ayo et al. 2020).

A neural network with more hidden layers is a deep neural network. DL algorithms make use of deep neural networks to train on a data and predict the output or do classification. DL which works with huge amount of data offers a significant advantage of automatically discovering the features for classification which an ML algorithm does with human intervention (Ayo et al. 2020). Considering the amount of content published on SM, neural networks have been an effective tool for automatic detection of SM content. Table 11 depicts the various neural network models deployed for detection of SM content. Considering the various characteristics of SM content, different neural network models are deployed. For example, discriminative models that consider SM content and context features are CNN and RNN (Islam et al. 2020).

Generative models like Generative Adversarial Network (GAN) and (VAE) that generate new data are explored for rumor detection (Ma et al. 2019; Sahu et al. 2019; Khattar et al. 2019). Hybrid models like CNN-RNN, RNN-GRU, CNN-LSTM, GAN-RNN (Shu et al. 2019; Badjatiya et al. 2017; Zhang et al. 2018) are explored for multimodal approach of SM content detection with visual features and textual features from two neural networks concatenated together for a classification task.

The performance of a machine learning and neural network applied to a particular task is evaluated on the

Table 12 Performance metrics of ML algorithm and DL

Performance Metric	Table/Formula
Confusion Matrix: The complete performance of a ML algorithm is measured with the confusion matrix with 2×2 dimensions i.e., "Actual" and "Predicted" giving an output with 4 values: "True Positive (TP)", "False Negative (FP)", "False Positive (FP)", and "True Negative (TN)"	*Actual* *Predicted* *True Positive(TP)* *False Positive (FP)* *False Negative (FN)* *True Negative (TN)*
Accuracy (A): It is a common evaluation metric that measures the correct prediction made by an algorithm	$Accuracy = \frac{TP+TN}{TP+FN+FP+FN}$
Precision (P): It is the number of correct positive results divided by the number of positive results predicted by the algorithm	$Precision = \frac{TP}{TP+FP}$
Recall or Sensitivity (R): It refers to the true positive rate and summarizes how well the positive class was predicted by an algorithm	$Recall = \frac{TP}{TP+FN}$
Specificity: It refers to the true negative rate and summarizes how well the negative class was predicted by an algorithm	$Specificity = \frac{TN}{FP+TN}$
F1-Score(F1): It is a harmonic mean between precision and recall and measures the robustness of an algorithm	$F1-score = 2 \times \frac{Precision \times Recall}{Precision+Recall}$

performance metrics detailed in Table 12. The detection of SM content using ML algorithms and DL is evaluated using accuracy, precision recall and F1-score. The performance of ML algorithm is tested for the number of false positives and false negatives which implies the misclassification rate of a specific content. For example, non-hate speech content misclassified as a hate content which indicates a high false negative. It is desirable that the ML algorithm should achieve a low false negative rate. Automated Techniques for fake news detection rely on AI based techniques with NLP tools combined with traditional ML algorithms and DL techniques. Various research articles have reported detection and classification of fake news and its types by exploring its content, user and social network characteristics (Shu et al. 2017; Zhou and Zafarani 2020). Various supervised ML algorithms like NB, SVM, KNN, LR, DT, and RF are experimented on various datasets to classify fake news as a binary classification task or a multi-class classification task.

Table 13 depicts the various supervised ML algorithms for detection of different forms of fake news like satire news, rumors, clickbait. Table 13 shows the diversity in features and also the datasets for detection of various forms of fake news. In context of fake news, detection involves classifying a piece of information as real or false which can be considered as two class classification problem. Most of the research literature shows the use of supervised ML algorithms that work on the available datasets for detection. There is a need to exploit unsupervised ML algorithms for detection. The lexical features, semantic and syntactic features are common feature selection methods for all forms of fake news. The writing style-based features vary for rumors, satire and clickbait detection. Many researchers have considered accuracy as a performance metric for evaluating the ML algorithm, however precision and recall are also important metrics that provide the percentage of fake news detected. The labor intensive and time-consuming task of developing handcrafted features for ML algorithms is considered by deploying DL neural networks which process huge amount of data without human intervention for extracting the features from such data.

Table 14 shows the use of supervised ML algorithms and ensemble ML algorithms (Malmasi and Zampieri 2018) for hate speech detection with SVM and LR reported better performance. In ensemble classifiers, individual classifiers are combined using various methods like Borda Count, Mean Probability Rule, and Median Probability Rule which help in improving the accuracy of classification task (Malmasi and Zampieri 2018). The ML algorithms are experimented on datasets and these datasets include diverse and fine-grained form of hate speech content. Twitter is widely used platforms for accessing the hate speech forms and creating a dataset. Figure 9 shows the statistics of ML algorithms deployed for detection and classification of SM content. As shown in Fig. 9 SVM is the most widely used algorithm for SM content detection with average accuracy of around 75% to 80%. SVM algorithm has shown increased accuracy for fake news detection as compared to hate speech. This is due to subjectiveness and variations in the hate speech words whereas fake news is objective in nature.

The performance metrics of an ML algorithm are more dependent on the datasets on which it is experimented. The ability to process huge data with automatic extraction of features from the data is unique characteristic of DL neural network models. This characteristic is explored in form of extracting various features like news content features, user responses to news, temporal characteristics using social graph which aid in fake news detection by neural network

Table 13 ML algorithms for various forms of fake news detection

Type of fake news	Dataset	Feature selection	ML algorithm	Performance Metric (in %)				Findings
				A	P	R	F1	
Fake news (Granik and Mesyura 2017)	BuzzFeed	Text features	NB	74	–	–	–	Low value of recall
Satire and humor News (Rubin et al. 2016)	360 news articles	Absurdity, Humor, Grammar, Negative Affect, Punctuation TF-IDF	SVM		90	84	87	Punctuation marks features are important for satire detection
Fake news (Ahmed et al. 2017)	Kaggle	Text features N-gram with TF-IDF	SVM	86	–	–	–	Best accuracy obtained for unigram features with decreased accuracy for n=4
			LSVM	92	–	–	–	
			KNN	83	–	–	–	
			SGD	89	–	–	–	
			DT	89	–	–	–	
			LR	89	–	–	–	
	BuzzFeed	N-gram with TF-IDF	LSVM	87	–	–	–	
Fake news (Horne and Adali 2017)	BuzzFeed	number of nouns, lexical redundancy (TTR), word count, number of quotes	LSVM	71	–	–	–	High accuracy for fake and satire news from real news
Satire news (Horne and Adali 2017)	Burfoot and Baldwin 2009		LSVM	67	–	–	–	
Real, fake and satire (Horne and Adali 2017)	Political news dataset		LSVM	91	–	–	–	
Fake news (Kleinberg et al. 2017)	Fake news AMT	Lexical, syntactic, semantic, readability features	LSVM	60	–	–	57	Comparative analysis of manual and automatic fake news detection
	Celebrity Fake news			61	–	–	57	
Rumors (Zhang et al. 2015)	Sina Weibo	Shallow text features, implicit user and content features	SVM	–	72	59	65	Exploration of rumor characteristics for detection
Rumor (Yang et al. 2015)	Twitter	Content based features, Twitter based features and Network	LR	–	98	95	96	Rumor identification using hot topic detection
			NB	–	98	90	94	
			RF	–	98	99	98	
Clickbait (Chakraborty et al. 2015)	7623 articles from BuzzFeed, Viralnova, Scoopwhoop, Viral stories	Sentence structure, word patters, n-grams, clickbait language	SVM	93	95	90	93	Work on English headlines
			DT	90	91	89	90	Approaches to block clickbaits
			RF	92	94	91	92	

Table 14 ML algorithms for various forms of hate speech detection

Type of Hate Speech	Data Source	Feature extraction	ML classifier	Performance Metrics (%)			Findings
				P	R	F1	
Hateful and antagonistic content (Burnap and Williams 2015)	Twitter	n-gram	BLR	89	69	77	Syntactic features reduced false negatives by 7%
		BOW	RFDT	89	68	77	
			SVM	89	68	77	
Hateful Offensive (Davidson et al. 2017)	Twitter	Bigram	LR	91	90	90	Multi-class classification
		Unigram					40% of hate speech misclassified
		Trigram					
		Each weighted by its TF-IDF					
Hate and Offensive (Watanabe et al. 2018)	Twitter	Sentiment based	RF	60	59	59	Binary classification for clean and Offensive text
		Semantic					
		Pattern	SVM	64	57	60	
		Unigram	J48graft-DT	79	78	78	Ternary classification for hate speech
Abusive language (Nobata et al. 2016)	Yahoo	Token n-grams	LR	77	79	78	Context of comment not taken into account
		Characters n-grams					
		Word2vec					
Aggression (Modha et al. 2020)	TRAC-Facebook	Unigrams with tf-idf	LR	68	57	60	Multiclass classification of aggression as overtly aggressive, covertly aggressive, non-aggressive
		Char 5-g					
		Length of the post	SVM	68	57	60	
	TRAC-Twitter		LR	52	52	49	Better performance for Facebook dataset
			SVM	49	49	49	
Racist (Kwok and Wang 2013)	Twitter	Unigrams	NB	**A**: 76			Considered only hate speech against blacks
							Reduced accuracy outside the context
Cyberbullying text (Dinaker et al. 2011)	Youtube comments	TF-IDF	NB	**A**: 63			Clustering of messages relevant to cyberbullying
		POS	J48 DT	**A**: 61			Detection of profanity and negativity from the clusters
			SVM	**A**: 66			
Hate, Aggressive, Profanity (Sharma et al. 2018)	Twitter	TF-IDF	NB	**A**:73.42			Harmful Speech categorized into 3 classes
		BOW	RF	**A**:76.42			
		TF-IDF	SVM	**A**:71.71			
Hate, Offensive (Malmasi and Zampieri 2018)	(Davidson et al. 2017)	Surface n-grams	Ensemble Classifier	**A**: 77			Misclassification of hate class as offensive
		Word skip grams					
		Brown clusters	LSVM	**A**: 78			
			RBF-SVM	**A**: 80			

models. The ML algorithms perform best for small datasets with TF-IDF representation technique (Cunha et al. 2021).

As shown in Table 15, pre-trained word embeddings are the most common feature representation techniques for classification. CNN and GAN architectures have shown significant performance in NLP tasks like text classification, sentiment analysis (Goldani et al. 2020). Transformer models have reported better classification accuracy for large datasets but at the cost of increased computational time and resources (Cunha et al. 2021).

State-of-art hybrid architectures like CNN-RNN, Attention- LSTM have also reported promising results in terms of accuracy and F1-score with few architectures implementing early detection of rumors and fake news. However, the time indication of early detection is not addressed in the literature.

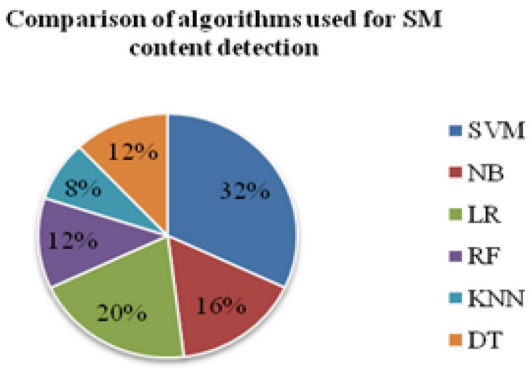


Fig. 9 Statistics of ML algorithms for SM content Detection

The social context for fake news detection task is also considered by neural network models like in CSI architecture (Ruchansky et al. 2017) and FANG (Nguyen et al. 2020) architecture. Nguyen et al. (2020) reported a Factual News Graph (FANG) framework that constructs a social context graph of list of news articles, news sources, social users and social interactions using Graph Neural Networks. The FANG framework showed AUC of 0.7518 on limited training data. However, the depending on the event for which fake news and rumors are disseminated, the social network graph features will change indicating the importance context that needs to be taken care for real time detection of fake news and rumors. The DL algorithms based on neural network architectures have outperformed traditional ML algorithms for detection of hate speech task. Various DL techniques like CNN, RNN, LSTM, Capsule networks, and Transformer models have shown good performance in terms of accuracy and F1-score.

Hybrid architecture like VAE + CNN (Qian et al. 2018) are experimented to generate user responses to news articles and extract semantic text features from posts assist in early detection of fake news.

Table 16 summarizes the DL techniques for detection of hate speech as reported in research. As shown in Table 16, various state-of-art DL techniques with hybrid neural networks are deployed for hate speech detection. CNN architectures are able to extract contextual features which are exploited in form of character CNN and word CNN (Park and Fung 2017) for hate speech detection and sentence level features with margin loss for fake news detection (Goldani et al. 2020). Like traditional ML algorithms, DL techniques also fail to detect and classify fine grained hate speech content like abusive content, offensive content and aggressive content. The error analysis for detection and classification of such content is missing and needs to be considered in research.

4.2.3 Multimodal approach of detecting detrimental content on SM

The multimedia forms an important attribute and modality that can assist in the moderation of SM content. The multimedia content includes images, videos, GIFs (Graphics Interchange Format). The development in multimedia technology has shifted the paradigm of text-based news articles to news articles that include a images and videos accompanied with text which attracts a greater number of readers (Qi et al. 2019). For example, a post or a tweet with images gets 89% more likes and a number of reposts for a tweet or posts with images is 11 times larger than a post without image (Cao et al. 2020). Recent years has also observed a rise in fake images attached to news article. As reported by Qi et al. (2019), the visual content which are false can be in form of tampered images, misleading images and images with wrong claim as shown in Fig. 10 (Qi et al. 2019, 2020). For detection of fake news from visual content includes exploring the diverse characteristics of the fake image (Cao et al. 2020) as these characteristics differ from a real image. These characteristics form the features which include forensic features, time-context features and statistical features (Cao et al. 2020) that are extracted to determine correctness of image. Qi et al. (Dinakar et al. 2011) experimented with forensic features using DCT to transform the image from pixel domain to frequency domain and the multiple semantic features of the image were captured using CNN with a bidirectional GRU (Bi-GRU) network to model the sequential dependencies between these features. These two features were concatenated together to detect the fake news achieving a accuracy of 84.6%. Boididou et al. (2015) experimented with forensic features and extracted descriptive statistics to detect fake news. The forensic features were combined with content-based features and user-based features which showed recall of 0.749, precision of 0.994 and F1- score of 0.854. The capabilities of DL neural networks are extended by combining the content and visual features together for detection of fake news and have shown promising results in terms of early detection of fake news and event discriminators (Wang et al. 2018). The user on a social media network publishes the content using different modalities like text, image and video. This form of modality is also observed in fake news and hate speech content sharing on social media. Majority of research literature has focused on exploring the textual content of news article for fake news detection. A textual content accompanied with visual content conveys more information that will assists in detection process.

Combining the textual and visual features together is challenging in terms of different characteristics like complex and noisy pattern of news article. Research studies have

Table 15 DL neural network models for fake news detection

Refs	Feature extraction method	Features of DL neural network used for detection	Dataset	Performance metric (in %)
Nasir et al. (2021)	Word2vec	Hybrid CNN-RNN architecture	ISOT	A, P, R, F1: 99
		Training on ISOT dataset and testing on FA-KES dataset	FAKES	A, P, R, F1: 60
Singhania et al. (2017)	GloVe	Representation of news article with news vectors	20,372 articles from 16 sites labeled fake	A: 96.24 for 3HAN
		3 level HAN for word, sentence and headline of input article		
		Bidirectional GRU for word, sentence and headline encoder	20,932 articles from 9 websites labeled genuine	A: 96.77 for Pre-trained 3HAN
		Attention weights to word, sentence and headline through heatmap		
Goldani et al. (2020)	n-grams	Pre-trained static, non-static and multi-channel word embeddings	ISOT	A: 99.8
		Model includes n-gram convolutional layer, the primary capsule layer convolutional capsule layer, and a feed-forward capsule layer for different length news statement	LIAR	
Paka et al. (2020)	BERT word embeddings	Creation of COVID-19 misinformation dataset	CTF (**C**OVID-19 **T**witter **F**ake News)	F1: 95
		Cross-Sean a semi-supervised attention neural model on unlabeled tweet texts		
		Encoding textual data using bidirectional LSTM		
		Cross-stitch unit for encoding user and tweet features		
		Chrome-SEAN, a chrome extension to flag COVID-19 fake news on Twitter		
Momtazi et al. (2020)	Static word embedding	Embedding layer with pre-trained word vectors	ISOT	ISOT: A: 99.1
	Non-Static word embedding	CNN layer with 3-g features of sentence	LIAR	
	Multi-channel embedding	Fully connected layer for classification using margin loss and softmax function		

Table 15 (continued)

Refs	Feature extraction method	Features of DL neural network used for detection	Dataset	Performance metric (in %)
Shu et al. (2019)	News sentences and user comments	Weighted sum of attention vectors for news sentence features and user comments features	GossipCop Politifact	GossipCop: A:80.8 P: 72.9 R: 72.2 F1-: 75.5
		Explainable fake news detection		
	Word and sentence level encoding with bidirectional RNN with GRU	Use of metric MAP@k(Mean Average Precision)(k = 5 or 10) to evaluate the model		Politifact: A:90.4 P: 90.2 R: 95.6 F1: 92.8
Qian et al. (2018)	Sentence and word representation	Early detection of fake news Two Level CNN to capture semantic information from long text Conditional Variational Autoencoder (CVAE) to generate user responses conditioned to a given news article	Weibo Twitter	Weibo: A: 89.84 Twitter: A:88.83
Dong et al. (2019)	Hybrid feature learning unit based on RNN	Credibility inference model from heterogeneous information fusion within the social networks A gated diffusive unit model that exploits the relationship among news articles, creators and subjects Hybrid feature learning unit for textual content and explicit and latent features	Politifact	A: 63
Hamdi et al. (2020)	Node2vec	Hybrid approach that exploits the credibility of information sources Graph embeddings to extract features from Twitter social graph User features combined with user graph features	Graph Dataset: ego-Twitter CREDBANK	A: 98.21 P: 91.3 R:99 F1: 98.2
Ruchansky et al. (2017)	Doc2vec	RNN to capture the temporal patterns of text LSTM model for temporal response of users to the article Users source characteristics using weighted user graph Prediction using the combined temporal and textual features and source score of users	Twitter Weibo	Twitter: A: 89.2 F1: 89.4 Weibo: A: 95.3 F1: 95.4

Table 15 (continued)

Refs	Feature extraction method	Features of DL neural network used for detection	Dataset	Performance metric (in %)
Ma et al. (2019)	Textual features using BoW	Rumor detection using generator and discriminative model (GAN)	Twitter PHEME	Twitter: A: 86, P: 88 R: 89 F1: 86 PHEME: A: 78, P: 77 R: 79 F1:78
		Generator model trained on a claim to generate uncertain and conflicting issues		
		Discriminative classifier learns to distinguish whether an instance is from real world using discriminative and low frequency patterns		
Singh et al. (2020)	13 linguistic and user features using GloVe	Hybrid feature extraction using CNN and LSTM	PHEME	P: 82, R: 81 F1: 82
		Selection of optimal feature set using Particle Swarm Optimization (PSO) algorithm		
		Attention LSTM for rumor veracity detection		

reported state-of-the-art multimodal architectures that are detailed next.

Figures 11 and 12 illustrate the multimodal approach of fake news detection. The two architectures concatenated the textual and visual representation features for detection of fake news. SpotFake (Singhal et al. 2019) architecture experimented with visual and textual features to classify fake news and EANN architecture explored the multimodal features for fake news detection and capturing event invariant features. EANN architecture (Wang et al. 2018) extracted the textual and visual features using CNN while BERT language model is used to extract the text features from the news articles in SpotFake architecture. Table 17 presents the multimodal architectures for fake news and rumor detection. The visual content extraction is done using VGG-19 (convolutional network pre-trained on ImageNet dataset) by most of the architectures. The dimensional vectors of visual and text modalities are made similar and combined together to form a feature vector which is then applied to a fully connected neural network with hidden layers and a classification layer. There is need to utilize these architectures for real time detection of fake news.

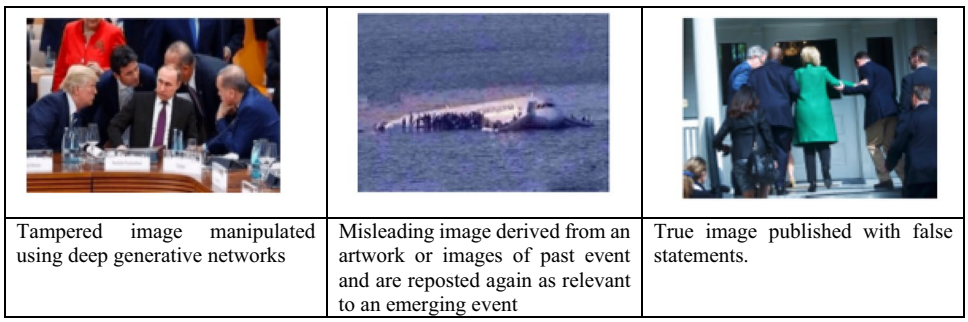
As reported in many research articles a single modality feature is not sufficient to identify a hate speech or abusive content. Many ML algorithms have reported false positive rates as certain words are either misclassified as hate speech words. A user on a social media can use various modalities like text, video, image and audio to share. The image accompanied with text will assist in detection of hate speech. Research studies have reported the use of image and text modalities for detection of hate speech. Kumar et al. (2021) presented a multi-modal neural network-based model that combined the text and image features to classify asocial media post into Racist, Sexist, Homophobic, Religion-based hate, other hate and No hate. Figure 13 shows the neural network model architecture for hate speech classification as reported in Kumar et al. (2021). The image content features are extracted using pre-trained CNN based VGG-16 network. The text features are extracted using text CNN architecture with GloVe word embeddings. The text and image features are concatenated and then applied to softmax layer for classification into 6 classes of hate speech. The model achieved weighted precision of 82%, weighted recall of 83%, and weighted F1- scores of 81% tested on the Dataset MMHS150K. The proposed model achieves high true positives for non-hate class with high false positive for homophobe and religion class. Kumari et al. (2021) reported a multimodal approach for a multiclass classification of

Table 16 Deep learning techniques for hate speech detection

Type of hate speech	Data source	Feature extraction	ML classifier	Performance metrics (%)				Findings
				P	**R**	**F1**	**A**	Random Embeddings (RE) detect hatred words better than GloVe embedding
Sexist, Racist (Badjatiya et al. 2017)	Twitter	Glove word embedding	LSTM + RE + GBDT	93	93	93	–	
			CNN + RE + GBDT	86	86	86	–	
			FastText + RE + GBDT	88	88	88	–	
Aggression (Modha et al. 2020)	TRAC-Twitter	FastText	CNN	70	60	64	–	Covertly aggressive content misclassified as non-aggressive
			BiLSTM + attention	71	51	55		
			BERT	72	58	62		Real time visualization of aggressive comments through web-browser plugin
	TRAC-Facebook		CNN	57	59	55	–	
			BiLSTM + attention	56	58	58		
			BERT	58	58	57		
Hate words (Amrutha and Bindu 2019)	Twitter	Word Embeddings	GRU	–	–	65	95	Models evaluated with 2 performance metrics
			CNN	–	–	64	94	
			ULMFiT			97	97	
								Context around the hate word not considered
Cyber hate: Religion (Liu et al. 2019)	Twitter	BOW	Fuzzy based: 4 fuzzy forms + KNN	84	40	5267	-	Fusion of multiple fuzzy classifiers and instance-based reasoning to detect the ambiguous instances in different types of hate speech
		Doc2vec						
Race				93	50	46	–	
Disability				96	60	74	–	
Sexual orientation				69	35	46	–	
Abusive language (Park and Fung 2017)	Twitter	Word2vec	CharCNN	74	67	70	-	Multi class classification of abusive and non-abusive content
			WordCNN	73	72	73		
			HybridCNN	72	75	73		
Sexist and racist comments		Word2vec	CharCNN + LR	94	94	94	–	Sexist and racist classification using CNN with LR
			Word CNN + LR	95	95	95	–	
			Hybrid CNN + LR	95	95	95	–	
Racism, sexism (Gambäck and Sikdar 2017)	Twitter	Random vector	CNN	87	67	75	–	Multi-level classification based on feature embedding
		Word2vec		86	72	78	–	
		Character n-gram		85	70	76	–	High false positive rate for non-hate text
		Word2vec + character n-gram		86	70	77	–	Not able to identify combined racist and sexist content
Sexist and Racist Comments (Pitsilis et al. 2018)	Twitter	Word embedding	RNN-LSTM	93	93	93	–	Multi class classification of sexist, racist and neutral content
								User behavior to content considered as feature
Hate speech (Roy et al. 2020)	Twitter	GloVe	C-LSTM	75	43	55	–	tenfold cross validation gave best recall prediction for hate and non-hate classification
			LSTM	64	53	58	–	
			DCNN	97	88	92	–	

Table 16 (continued)

Type of hate speech	Data source	Feature extraction	ML classifier	Performance metrics (%)				Findings
				P	**R**	**F1**	**A**	Random Embeddings (RE) detect hatred words better than GloVe embedding
Hate Inducing Abusive (Mathur et al. 2018)	HOET	GloVe	Ternary Trans-CNN	80	69	71	84	Three class classification for code switched hate speech Transfer Learning neural network model



Tampered image manipulated using deep generative networks

Misleading image derived from an artwork or images of past event and are reposted again as relevant to an emerging event

True image published with false statements.

Fig. 10 Example images in fake news articles

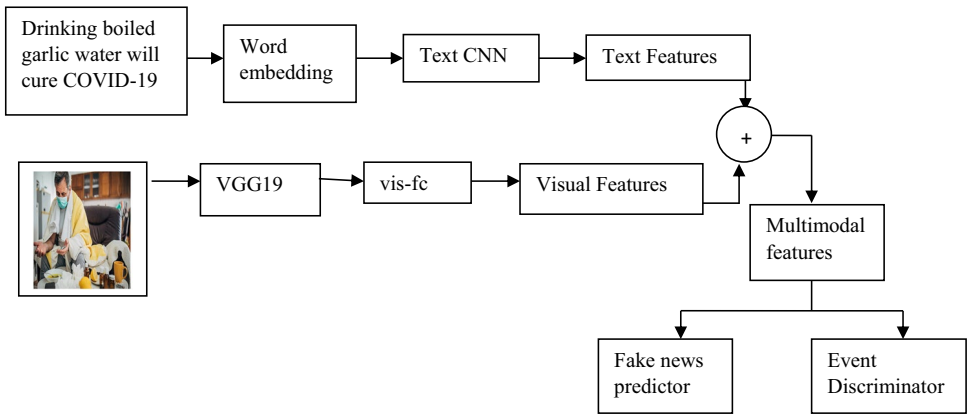


Fig. 11 Architecture of EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection (Wang et al. 2018)

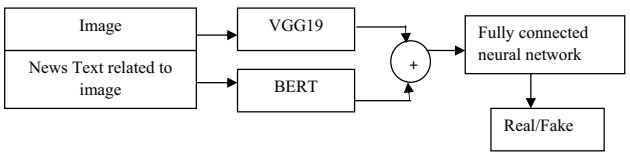


Fig. 12 Architecture of SpotFake (Singhal et al. 2019)

cyber-aggression on social media for posts which consists of symbolic image together with text.

The symbolic image features were extracted using VGG-16 network and textual features using CNN with three layers. The concatenated image and textual features were optimized using Binary Particle Swarm Optimization (BPSO) algorithm.

Using BPSO algorithm, the redundant features were eliminated, and the new hybrid features were applied to Random Forest ML classifier to classify the social media posts into non-aggressive, medium- aggressive and high aggressive. The dimensions of concatenated features were reduced from 1024 to 507 using BPSO algorithm. The proposed system achieved weighted precision of 74%, weighted recall of 75%, and weighted F1-scores of 75% on a created dataset of 3600 images together with text acquired from Facebook, Twitter and Instagram. The study reported to have a performance

Table 17 Multimodal architectures for fake news detection

Refs	Text representation	Visual content representation	Performance metric				Dataset	Key findings/Features
			A	P	R	F1		
(Wang et al. 2018)	Text-CNN	Pre-trained VGG-19	72	82	64	72	Twitter	Multimodal features to learn the discriminative
			83	85	81	83	Weibo	Representations for fake news identification learn the event invariant representations by removing the event-specific features
(Singhal et al. 2019)	BERT	Pre-trained VGG-19	77	75	90	82	Twitter for fake news	Explored only content and visual features for detection. Empirical analysis of fake news through public surveys
			89	90	96	93	Weibo for fake news	
Khattar et al. (2019)	Bi-LSTM	Pre-trained VGG-19	75	80	72	76	Twitter for fake news	Variational Autoencoder to discover the correlations of text and visual features for fake news detection
			82	85	77	81	Weibo for fake news	
Zafarani et al. (2020b)	Text-CNN	Image2sentence model	87	88	90	89	Politifact	Recognizing the falsity of news by detecting the mismatch between image and text
			83	85	93	89	GossipCop	
Cui et al. (2019)	Glove for content	VGG	Micro F1 for 80% training ratio: 77				Politifact	Latent sentiments in the users' comments for fake news detection
	One- hot encoding for profile and user comments		Macro F1 for 80% training ratio:76				GossipCop	Adversarial mechanism to conserve semantic relevance and the representation consistency across different modalities
			Micro F1 for 80% training ratio: 80					
			Macro F1 for 80% training ratio:81					

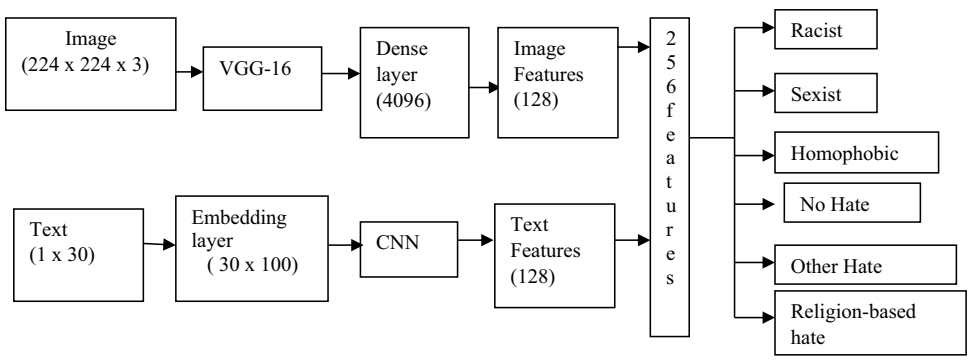


Fig. 13 Neural network model architecture for multimodal hate speech classification Kumar et al. (2021)

improvement of 3% when optimized features are used for classification.

Cheng et al. (2019) reported a collaborative multimodal approach of cyberbullying detection based on heterogeneous network representation learning. The study used 5 modalities from Instagram like image, user profile (number of followers, the total number of comments, and the total number of likes received), timestamp of posting an image, description of the image and comments, and dependencies between social media sessions through relations among users. The system reported a Macro F1-score of 96% and Micro F1-score of 98%.

Sahu et al. (2021) experimented with GAN-fusion model that combined different adversarial models for text, caption and image achieving a precision of 61%, recall of 51% and F1-score of 56%. The experimentation was done on a MMHS150K dataset which includes image, its caption and text.

The multimodal approach of various forms of hate speech detection includes extracting features from different modalities using deep neural networks. For multi-modal detection, context is important feature which is missing in reported systems and needs to be considered in research. The method of concatenating the features from different modalities is less detailed in the literature.

5 Moderation of detrimental content on SM platforms

The exploitation of SM for a wrong purpose is increasing substantially every year and is imposing challenges to various sectors like private organizations, government and civil society (Ganesh and Jonathan 2020). Inspite of legal measures enforced by the government to control the devastating detrimental content on SM, the dissemination of such content has not stopped. So, content detection and moderation on SM platforms is of primary importance. Content moderation on online platforms has drawn attention in academia with many research articles published in scientific journals. Traditional publishing platforms detect and moderate the content by verifying the content with known facts (Wyrwoll 2014). Content moderation involves decisions about decreasing the presence of extremist contents or suspending exponents of extremist viewpoints on a platform (Ganesh and Jonathan 2020), elimination of offensive or insulting material, the deletion or removal of posts, the banning of users (by username or IP address), making use of text filters to disallow posting of specific types of words or content, and other explicit moderation actions (Ganesh and Jonathan 2020). Content moderation involves law enforcement organizations set by government and civil society (Ganesh and Jonathan 2020). Commercial content moderation is a method of

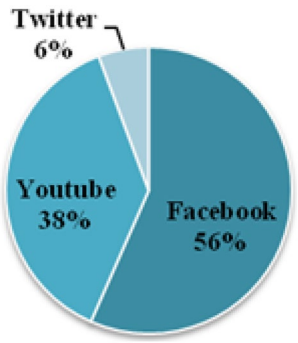


Fig. 14 Manual Content moderation on SM Platforms

screening the UGC on SM platforms like Facebook, Twitter, Youtube, Instagram with help of large-scale human moderators that make decisions about the appropriateness of UGC (text, image, video) posted on SM (Roberts 2017b). Content moderation is implemented by SM companies in three discrete phases namely (Common 2020).

- Creation: Creation describes the development of the rules (the terms and conditions) that platforms use to govern the user's conduct.
- Enforcement: Enforcement includes flagging problematic content, making decision on whether the content violates the rules set in creation stage and accordingly the action to be taken for the problematic content.
- Response: Response describes the internal appeals process used by platforms and the methods of collective action activists might use to change the platform from the outside. For example, controversies that arose over the live streaming of murder and sexual assaults were considered by social media companies in form of response as announcing hiring more moderators to have better control over such events. (Gibbs 2017).

This section describes the manual, semi-automated and fully-automated methods of moderation.

5.1 Manual approach of moderating detrimental content on SM platforms

Content moderation as defined by Grimmelmann (2015), is the use of administrators or moderators with authority to remove content or prohibit users and making the design decisions that organize how the members of a community engage with one another. Content moderation is considered as indispensable component for SM platforms (Barrett 2020; Roberts 2016). Content moderators are important stakeholders that ensure safety of SM platforms (Roberts 2017; Gillespie 2018; Barrett 2020). The content moderators

decide which content is appropriate to be kept on SM and which content should be removed (Barrett 2020).

Commercial content moderation is particularly meant for moderating the objectionable content on SM platforms with help of human moderators that adjudicate such content (Roberts 2016).

The origin of content moderation started with an intention to protect the users of SM platforms from pornography and offensive content (Barrett 2020). Content moderation initially was done by in-house team of people who review the content based on the set of moderation rules defined by social media company and instructions about the removal of certain content (Barrett 2020, Crawford and Gillespie 2016). With the increase in the usage of users and the content shared by them, it became challenging for in-house team to moderate the content. Figure 14 shows the statistics of moderators hired by popular SM platforms (Barrett 2020). As shown in Fig. 14, Facebook holds the highest number of moderators around 15,000 followed by YouTube with 10,000 moderators and Twitter having around 1500 moderators (Barrett 2020). The figures quantify the amount of content shared on these platforms and the number of moderators who do the task of screening the content. To scale up with the increasing content, social media companies have marginalized the people and have outsourced the task of moderation to third-party vendors who work at different geographical locations which include U.S., Philippines, India, Ireland, Portugal, Spain, Germany, Latvia, and Kenya (Barrett 2020). The task of moderation is also done using online websites like Amazon Mechanical Turk (Roberts 2016).

Flagging is a detection mechanism used by the user community to report an offensive content, violent graphic content to the SM platforms (Gillespie 2018; Roberts 2016; Crawford and Gillespie 2016). To scale with the content published on SM, AI based methods are deployed to detect the detrimental content (Barrett 2020; Crawford and Gillespie 2016). The flagging mechanism is widely observed in the SM platforms that allow the users to express their concern about the content posted on these platforms (Gillespie 2018; Crawford and Gillespie 2016). The flagged content is then by reviewed by the content moderators who checks whether the content violates the Community guideline policies of the platform (Gillespie 2018). Many SM platforms consider the content flagged by the user as important, as it helps in maintaining their brand (Gillespie 2018). The flagging mechanism also reduces the load of content moderators as they need to review only the flagged content instead of reviewing all the posts.

Human content moderators analyze the online comments and posts shared by the users using the Community Guidelines defined by the SM platforms (Roberts 2016). The Community Guidelines are framed by all social media platforms that define the rules and policies about the types of content to be kept and content to be removed from on the platform. For example, Youtube's Community Guidelines include excluding shocking and disgusting content and content featuring dangerous and illegal acts of violence against children (Roberts 2016). Facebook defines Community Standards that include policies on hate speech, targeted violence, bullying, and porn, as well as rules against spam, "false news," and copyright infringement with policy rules made by lawyers, public relations professionals, ex-public policy wonks, and crisis management (Koebler and Cox 2018).

The process of content moderation starts with training of the volunteers about the policies set by the platforms and making them observe the moderation work done by the experts. The volunteers are given the information through the database regarding what constitutes hate speech, violent graphic content (Koebler and Cox 2018) and also includes on-boarding, hands-on practice, and ongoing support and training (Barrett 2020). The moderators are given the task to moderate any specific form of objectionable content. The moderators then decide whether the content is according to the policy standards as defined by the platforms (Barrett 2020). Each moderator is given a handling time to process the content and then make a decision, which is approximately 10–30 s per content Common 2020; Barrett 2020). The moderators after screening the content, remove it, retain it or mark it as disturbing Common 2020; Barrett 2020). SM platforms expect 100% accuracy from content moderators[10] but as Mark Zuckerberg admitted in a white paper that moderators "make the wrong call in more than one out of every 10 cases,"[10].

Moderators also review the content in a different language by using the social media company's proprietary translation software (Barrett 2020). Many times, the moderators had to remove the same content multiple times which have led to many health problems (Barrett 2020; Roberts 2016). Over exposure to disturbing videos and images of sexual assault and violent graphics, the moderators experienced insomnia and nightmares, unwanted memories of troubling images, anxiety, depression, and emotional detachment and suffered from post-traumatic stress disorder (PTSD) (Ofcom 2019; Barrett 2020).

Human experts are involved in pre-moderation phase (moderate the content before it is published) and post-moderation phase (moderate the content after it is published) (Ofcom 2019). The manual approach of moderation requires that the expert must be aware of the context in terms of geographical location and its laws from where the content is shared and published, the SM platform and must be well versed with the language of the content to understand

[10] https://www.forbes.com/ Facebook Makes 300,000 Content Moderation Mistakes Every Day.

the meaning and the relevance (Roberts 2017a). All these aspects demand a special training for moderators to screen the online content.

5.2 Semi-automated technique of moderating detrimental content on SM platforms

The manual approach of content moderation has many challenges in terms of the volume, veracity and speed of problematic content to be analyzed, the cultural, historical and geographical context around the content. Many companies and governments are proposing automated processes to assist in detection and analysis of problematic content, including disinformation, hate speech, and terrorist propaganda (Leerssen et al. 2020).

Semi-automated moderation techniques include use of AI tools to automatically flag the text, image, video content and review of the flagged content done by the human moderators. The automated flagging mechanism will reduce the workload of human reviewers. The AI based tools like hash matching in which a fingerprint of an image is compared with a database of known harmful images, and 'keyword filtering' in which words that indicate potentially harmful content are used to flag content (Ofcom 2019) facilitate the review process of human moderation. The Azure content moderator by Microsoft is AI based content moderation tool that scans text, image, and videos and applies content flags automatically. The web-based Review tool stores and display content for human moderators to assess the content.[11] The tool includes moderation Application Programming Interface (API) that checks the objectionable content like offensive content, sexually explicit or suggestive content, and profanity, checks the images and videos that contain adult or racy content. The review tool assigns or escalates content reviews to multiple review teams, organized by content category or experience level[11].

Andersen et al. (2021) presented a real time moderation of online forums with a Human-In-the-Loop (HiL) to increase the moderation accuracy by exploiting human moderation of uncertain instances in test data. Each comment is classified as valid or blocked using a ML algorithm with an additional comment marked as uncertain which is evaluated and labeled by human moderators. The human labeled instances are added to the training data and then the ML model is re-trained. With moderating 25% of test dataset, the detection of valid comments is increased to 92.30% with help of manual intervention.

The performance of semi-automated techniques of content moderation is more dependent on the accuracy of AI tools used to flag a content and image. The AI tools should also detect the degree of diversity used in the social media UGC which is challenging and demands more attention in the research. The automatic flagging mechanism needs to be experimented in real time and monitor how these tools assist the human moderation process. AI based flagging tools should be exploited more to detect a harmful text or image and give an indication in form of a flag that signifies a terrifying or dreadful content to be screened by a human moderator.

5.3 Automated technique of moderating detrimental content on SM platforms

The psychological trauma experienced by the human moderators (Roberts 2016) and the challenge of handling the significant rise in the UGC on SM platforms demands for a use of automated technologies in the form of AI. With the increasing pressures of government on SM companies to grapple with the disturbing content, both government organization and SM companies are suggesting the use of technical solutions for moderating the SM content (Gorwa et al. 2020). AI and automated systems can assist manual moderation by reducing the amount of content to be reviewed thus increasing the productivity of moderation and also help in restricting the exposure to disturbing content by manual moderators (Ofcom 2019). History reports the use of automated systems like "Automated Retroactive Minimal Moderation" systems to filter the growing spam content on USENET using automated filters (Gorwa et al. 2020).

Systems like automated 'bot' moderators fought vandalism and moderated the articles on Wikipedia (Gorwa et al. 2020). Automated content moderation also referred as algorithmic moderation or algorithmic commercial content moderation are systems that identify, match, predict or classify the UGC which takes the form of text, audio, video or image based on the exact properties and general features of UGC with a decision and governance outcome in form of deletion, blocking the user or removal of account of user (Ofcom 2019; Grimmelmann 2015).Artificial intelligence (AI) is often proposed as an important tool for identifying and filtering out UGC that is offensive or detrimental.

Automated tools are used by the SM platforms to monitor the UGC which covers terrorism content, graphic violence, toxic speech like hate speech and cyberbullying, sexual content, child abuse and spam/fake account detection (Grimmelmann 2015). The Global Internet Forum to Counter Terrorism (GIFCT) is founded by SM platforms like Facebook, Twitter, Microsoft and Youtube to remove the extremists and terrorism content from SM (Ganesh and Jonathan 2020; Grimmelmann 2015). The SM platforms under GIFCT have created a secret database of digital fingerprints (called as 'hash') of terrorist content (images, text, audio, video) called as Shared Industry Hash Database (SIHD) which contain

[11] https://docs.microsoft.com/azure/ Content Moderator Overview.

Table 18 Automated Tools to moderate UGC

SM platform	Automated tool	Type of content moderated	Methodology
Google Jigsaw's (Hosseini et al. 2017)	Perspective API	Toxic comments	ML model to score the toxicity of input comments in real time
Microsoft[a]	PhotoDNA	Child exploitation images	Unique digital signature ('hash') for an illegal image compared against a database of another digital signature
Youtube[b]	Content ID	Audio and video	Music and video files uploaded on Youtube are scanned against a database of files
Twitter[c]	Quality Filter	Harassment text	Tool to hide the low-quality notifications from bots and spammers
Counter Extremism Project[d]	eGLYPH	extremist content	Like PhotoDNA, a hash for extremist content
Facebook[e]	RoBERT, RIO, LinFormer	Hate speech content	NLP models in different languages using Transfer learning

RIO Reinforced Integrity Optimizer, *LinFomer* Linear Transformer

[a]https://www.microsoft.com/ Photodna

[b]https://support.google.com/youtube/ YouTube Operations Guide Using Content ID

[c]https://techcrunch.com/2016/08/18/ Twitter is introducing a quality filter to clean up your notifications tab

[d]https://www.counterextremism.com/video/how-ceps-eglyph-technology-works

[e]https://ai.facebook.com/blog/how-ai-is-getting-better-at-detecting-hate-speech, Nov 2020

40,000 image and video hashes (Singh2019) and developed automated systems to detect the terrorist content (Gorwa et al. 2020). The database is updated by adding content through trusted platforms (Grimmelmann 2015). The image or video content uploaded by social media platform users are hashed and checked against the SIHD and if the content matched with hash in database, it is blocked (Gorwa et al. 2020).

Many SM platforms relied on automated techniques of content moderation during COVID 19 pandemic as many human moderators were sent home to limit the exposure to virus (Barrett 2020). Table 18 depicts the automated tools used by SM platforms to moderate the detrimental UGC. The automated tools used by SM platforms make use of ML algorithms that are applied to diverse categories of UGC like text, image video and audio formats. As shown in Table 18, automated tools developed by Facebook like RoBERT architecture detect hate speech in multiple languages across Facebook and Instagram.[12]Facebook reported that AI tools like RIO were able to detect 94.7% of hate speech and was removed from Facebook[12]. Tools like PhotoDNA[13] and ContentID[14] work by generating a digital fingerprint called as 'hash' for each of illegal image file or audio and video file. These signatures are stored in a database which is used to compare with other signatures. Signatures identical to stored ones are automatically flagged.[15] As reported by Microsoft[13], PhotoDNA is not face recognition software and hash is not reversible so the tool cannot be used to recreate an image. ML algorithms are used by automated tools for matching the content against the stored database of content which worked best for detection of illegal image.

However, the automated tool named eGLPHY[16] to detect the extremist content raised major concerns about what constitutes an extremist content to be included in the hash database and each platform framed its own policies and definitions of extremist content (Gorwa et al. 2020). This implies a biased decision making as extremist content is subjective and more dependent on geographical location.

6 Discussion and conclusion

SM has brought a big revolution in the society exploring new dimensions of communication through connectivity with people across the globe and providing ample opportunities in professional domain through social media marketing. While SM is proving to be a boom and a kind of blessing to entire society, it is actually a blessing in disguise due to

[12] https://ai.facebook.com/blog/how-ai-is-getting-better-at-detecting-hate-speech, Nov 2020.

[13] https://www.microsoft.com/ Photodna.

[14] https://support.google.com/youtube/ YouTube Operations Guide Using Content ID.

[15] www.snopes.com.

[16] https://www.counterextremism.com/video/how-ceps-eglyph-technology-works.

its negative impact which is up surging now with millions of posts on hate speech, online abusive and cyberbullying content, and hundreds of fake news generated by users. Such incidences have led to many fatal deaths, psychological disorders, and depression. This catastrophic negative impact of social media on the society necessitates the dire need of detrimental content detection and moderation. Content detection and moderation is now an inevitable component of SM platforms that is flourishing in real time. This research presents an exhaustive survey with pointers, findings and research gaps involved in detrimental content detection and moderation on social media platforms.

With a phenomenal increase in detrimental content on social media platforms, an accurate detection of such content is important at its first place. Manual detection methods cannot scale up with the increasing detrimental content. The recent advancements in AI through state-of-art algorithms, computational power and the ability to handle huge data (Ofcom 2019) have opened doors to automate the detection process of online content. NLP techniques have shown significant results in parsing the specific form of social media content. Feature engineering techniques like BoW, n-grams, TF-IDF, and PoS tagging are vital components of NLP that extract the character and word level features from the content and create numerical feature vectors. These frequency-based features representation methods suffer from higher dimensionality and sparse feature vectors which are addressed by word embeddings feature representation techniques. The NLP based ML algorithms perform best when they are trained on a dataset that consists of particular type of content like hate words, abusive words or rumor statements achieving accuracy of around 80% for a specific dataset. In case of hate speech content, there is a spectrum of variation in such content that is dependent on demographic locations, cultures, age, gender, religion. Research have reported the use of ML algorithms for detection of a particular type of hate speech. These algorithms show high rate of false positives when applied to different type of hate speech content. These classifiers lack in ability to capture the nuances in the language used by social media users which needs to be considered in research.

Non-contextual word embeddings like word2vec, GloVe, and contextual word embeddings like FastText, BERT, GPT-3 XLNET, DistilBERT are neural network based pre-trained language models that consider the semantic, syntactic, multilingual, morphological features and Out Of Vocabulary (OOV) words in the text. The maximum accuracy achieved with pre-trained word embeddings alone is around 80%-85%. When using pre-trained models, the number of hyper parameters raise up to billions. Automated systems deploying pre-trained models with these huge parameters leads to increased training time of the neural network model, more compute intensive work which in turn will affect the speed of system. Also, the pre-trained language models trained on huge, uncurated static datasets collected from the Web encode hegemonic views that are harmful to marginalized populations.The pre-trained language models also reveal various kind of bias with more negative sentiments toward specific groups and overrepresentation for words like extremists, toxic and violent content (Bender et al. 2021) This kind of bias in the training characteristics of language models can show a potential risk of wrong judgment when deployed for practical implementation of detection of detrimental content on SM. Transfer learning techniques that make use of pre-trained models is preferred method for detection of English-only content. Automated Systems deploying transfer learning approach have reported low recall due to the variability in definition of a particular content. For example, offensive words misclassified as non-hate words. Although the contextual pre-trained models consider the context of the word in the sentence, the context in social media posts is not considered by these models and the cases of false positives and false negatives is not taken into account by the systems deploying pre-trained models. Exhaustive experimentation and validation are needed before these models are practically deployed.

The present automated systems are dependent on datasets which are created by annotators which has a potential risk of biased decision by the annotator in assigning a label to content. One should also consider the process of automating the annotation which will actually add true essence to the complete automation process. If the current systems focus on manual annotation in developing of automating systems will still end up in designing of semi-automated systems. From the perspective of a fully automated system, automation of this process is also important which is not considered in the present available system Exhaustive research for annotation considering the labeling of data from the angle of the context needs to be operated in these systems.

Traditional ML algorithms need human intervention to extract the important features for detection of inappropriate content. Hand-crafted features are often either over-specified or incomplete. Considering the size of the SM data, developing hand-crafted features for such task is costly and complex job. A bias introduced in developing the features may cause harm in making an incorrect decision by a ML algorithm which restricts its practical deployability in real time. Automatic extraction of features and ability to process huge data by DL techniques through various language models has shown significant results in the task of content moderation. However, DL techniques have difficulty in finding the optimal hyper parameters for a particular dataset (Nasir et al.2021) which increases the training time and inference time during testing. DL techniques also rely on language models which are trained on billions of hyper parameters [for example T-NLG by Microsoft trained on 17 billion

parameters (Bender et al. 2021)]. DL techniques along with language models perform sophisticated task but at a cost of increased computational resources and cost, increased training time, and inference time. These aspects restrict their practical implementation in real time. Optimization of DL techniques and fine tuning the language models with optimal parameters is of supreme importance that needs further research.

The present automated system that deploys ML and deep neural networks for detection and classification of detrimental content have considered accuracy, precision and recall as performance metrics. None of the systems to the best of our knowledge have reported the time taken by an algorithm to detect an objectionable content. NLP and neural network models show increased accuracy when they are trained to detect a particular type of detrimental content like abusive speech. These models show decreased accuracy when they applied across different detrimental content format, language and context. Considering practical deployment of these algorithms in real time, time is an inevitable parameter in automated systems. Further research with rigorous experimentation on the time required will be an important contribution in this domain and therefore needs to be considered.

Content moderation is a process of making an indispensable decision about which form of UGC should be kept online and which form should be removed from the SM platforms. The task of moderation done by SM platforms involves use of human experts that analyze violent, sexually explicit, child abuse content, toxic, illegal hate speech and offensive content in text, image and video format. The experts then flag the content and remove it from the platform if it violates the community guidelines as defined by social media companies. According to statista.com, SM companies spend around $1,440 to $ 28,800 annually on these moderators to review billions of posts every day. With an extensive training of three to four weeks, the moderators evaluate each content within an average time frame of 30 s to 60 s; covering almost 700 posts in an eight-hour shift (Barrett 2020) with accuracy of moderation ranging from 80 to 90%. Considering the time taken by moderators to evaluate content and the accuracy achieved within this stringent time frame is uncertain. The time at which content are moderated is nowhere comparable to the frequency at which posts are published. There are multiple factors involved in manual moderation like training time, the noisy form of content, the mental state of moderators after checking the huge volume of content, understanding the dynamic and reactive community guidelines set by SM platforms before moderation, the amount paid to the moderators, and accuracy of moderation is also not comparable. All these aspects of manual moderation can be bridged by an automated system. Manual moderation also includes reviewing a content that is flagged by a user community on social media. This helps the human moderator but there are more chances of bias getting introduced in the decision of flagged content made by the user community. The context around a content is vital and a crucial aspect which is completely missing in the present manual moderation process.

Semi-automated moderation systems try to deal with the trade-off of volume of content versus the time taken to analyze a content using manual moderation technique. Semi-automated systems are deployed by SM companies to curb with the accelerating increase in the problematic or objectionable content. These systems make use of AI tools to automatically flag a content which is then reviewed by a human moderator. Such systems facilitate the review process of human moderation in evaluating only the objectionable content. However, a particular content flagged by AI tool might not be objectionable from moderator's perspective. This additionally entails a bias and discrepancy in the decision made by the AI tool and the moderator. Transparency in the decision made by AI tool to assist manual moderation is missing and demands immediate attention and further research.

In some typical alarming situations government raises red flags and demands urgent content moderation from social media companies. In such situations, it is challenging for SM companies to appoint manual experts for flagging the flooding content. More ever it has to be done in stipulated time on urgent basis and needs to be done accurately. In such scenarios, automated systems will play a vital role and will obviously be preferred over any semi-automated and manual system. So further research in developing an automated system is a dire need considering such real time situations.

The scalability problem of social media content and psychological trauma experienced by human moderators can be addressed by an automated approach of moderation. The automated approach of content moderation fueled by AI and ML is deployed by many SM platforms in form of automated tools like PhotoDNA by Microsoft, ContentID by YouTube, Quality Filter by Twitter, and RoBERT by Facebook. These tools organize, filter and curate extremists and violent content, child abuse material, hate speech content, and copyright violations in text, image audio and video format. These tools work by creating a common database of illegal images and text content and this database is used by companies to moderate the content. The database is updated with new text and image content. However, each social media platform has their own definition of illegal or harmful text which is stored in the database. This leads to discrepancy in moderating a specific form of content with the possibility of automated tool making an incorrect decision. The definition of an extremist content or hate speech is dependent on the demographic location which is not considered in the current systems. Considering these variations and subjectiveness, it is very important to design and develop an automated system which can be globally deployed across any demographic

location and still give encouraging results for content moderation. This is an aspect of paramount importance but has received major attention in the current systems. Therefore, further research is necessary to design globally deployable systems with objectified decision making.

The current trend shows that the user on social media has shown an inclination toward audio and video clips, emojis, smiley's, GIFs formats for expressing their views. The current social media is acutely inclined toward use of these formats. This makes the task of manual moderation too challenging in terms of interpretation, time to evaluate and making a decision about flagging and removing such form of content which can lead to error and affect the accuracy of moderation. An automated approach can assist in moderating multimedia content. However, the current designed automated systems are all focused on words and driven by the content in terms of the text. These systems need exhaustive research to imbibe smileys, emojis, and gifs format so as to make it full proof. To the best of the knowledge, this aspect is completely ignored in the present automated system. Designing a system that will take into account this multimedia and give the decision is of dire need considering the present scenario. Advanced ML algorithms will be needed to design such systems which are yet not explored. Heavy experimentation and designing the datasets which will include all characteristics of a content making it publicly available, keeping it open for the research community to float their ideas and developing a system that will be universally acceptable is an important aspect that needs to be covered in research.

Google has launched a text translator for 109 languages. Considering a typical case like India, users write regional language in English. Another characteristic of present social media is lack of restricting to one particular language or preference of using combination of languages when expressing and sharing views, (For example writing Marathi in English) called as code mixed language. The liberty of using Hinglish language (Hindi + English) or Reglish (Regional + English) language is another dimension of content moderation that has received little attention in research. Research community has reported creation of datasets in code mixed language for hate speech and abusive content. Multilingual BERT (mBERT) pre-trained models developed by Google which include more than 100 languages are trained on certain code-mixed content (Hinglish) but often lack in detection of fine-grained definition of hate speech content.

Even though the deep neural networks-based NLP models have currently shown promising performance in machine translation, named entity recognition, sentiment analysis, but have underperformed for automated analysis of social media content. It is very important to develop these models that will capture the subtleties of language across different context which needs to be explored in research.

Fairness and trust in decision making by an AI based systems is an important aspect for realization of real time applications. NLP techniques are considered as white-box models that are inherently explainable (Bender et al. 2021). However, due to word embeddings, the present NLP models are based on deep neural network are considered as black box which lack in interpretability. Explainable AI (XAI) (Danilevsky et al. 2020) is a new emerging field of AI aimed at developing a model more explainable and interpretable in terms of making a user understand of how a model arrived at a result. Research literatures have reported various forms of explanation in NLP through feature importance, surrogate model, example driven, provenance, declarative induction (Danilevsky et al. 2020). The explainable aspect is explored for fake news detection (Shu et al. 2019) through attention-based models. XAI though not a fully developed field needs to be explored for developing a transparent automated SM content moderation system with more exploitation on the features extracted from the user's posts on SM.

Further research is needed so as to make context driven decision making about the content is of paramount importance considering manual approach. Content moderation is subjective, and perspective of objectionable language varies according to user, geographic location, culture and history. This all necessitates a exhaustive research and a thorough understanding of social media content while designing a fully automated content moderation system.

The detrimental content posted on the social media has already caused the damage to the society. Present systems focus on moderating it or removing it after the damage has already done. But to the best interest of mankind and humanity, researchers need to think beyond moderating the content and going step further to prevent it wherein there is some flagging assigned to a user and after a threshold is decided on number of inappropriate posts, like ATM cards the user for 24 h is banned from SM. So, the researchers need to think beyond the obvious of only moderating or only restricting their research to moderation of content, but prevention of such cases will actually serve as a boom to social media. Designing a system that will monitor the user's history of posting detrimental content, setting a threshold on the number of objectionable posts and then raising a flag when the threshold has crossed will ensure a safe environment on social media.

References

Ahmed H, Traore I, Saad S (2017) Detection of online fake news using N-Gram analysis and machine learning techniques. In: Traore I, Woungang I, Awad A (eds) Intelligent, secure, and dependable systems in distributed and cloud environments. ISDDC. Lecture Notes in Computer Science, Vol. 10618, pp 127–138. https://doi.org/10.1007/978-3-319-69155-8_9.

Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. J Econ Perspect 31(2):211–236

Amrutha BR, Bindu KR (2019) Detecting hate speech in tweets using different deep neural network architectures. In: Proceedings of the international conference on intelligent computing and control systems (ICICCS 2019) IEEE, pp 923–926. https://doi.org/10.1109/ICCS45141.2019.9065763.

Andersen JS, Zukunft O, Maalej W (2021) REM: efficient semi-automated real-time moderation of online forums. In: Proceedings of the joint conference of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing: system demonstrations. pp 142–149.

Ayo FE, Folorunso O, Ibharalu FT, Osinuga IA (2020) Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. Comput Sci Rev Elsevier. https://doi.org/10.1016/j.cosrev.2020.100311

Badjatiya P, Gupta S, Gupta M, Varma V (2017) Deep learning for hate speech detection in tweets. In: 26th international conference on world wide web companion, Perth, Australia, pp 759–760. https://doi.org/10.1145/3041021.3054223.

Barrett PM (2020) Who moderates the social media giants? A call to end outsourcing. report: NYU Stern Center Centre for Business and Human Rights.

Basile V, Bosco C, Fersini E, Nozza D, Patti V, Pardo FMR, Rosso P, Sanguinetti M (2019) 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, pp 54–63. https://doi.org/10.18653/v1/S19-2007.

Bender EM, Gebru T, Shmitchell S, McMillan A (2021) On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp 610–623. https://doi.org/10.1145/3442188.3445922.

Boididou C, Papadopoulos S, Nguyen DT, Boato G, Kompatsiaris Y(2015) The certh-unitn participation@ verifying multimedia use 2015. Verifying multimedia use at MediaEval 2015. In: MediaEval benchmarking initiative for multimedia evaluation.

Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguistics 5:135–146. https://doi.org/10.1162/tacl_a_00051

Bonet OG, Miguel NP, Garcia-Pablos A, Cuadros M (2018) Hate speech dataset from a white supremacy forum. In: 2nd workshop on abusive language online @ EMNLP. https://doi.org/10.18653/v1/W18-5102.

Brown TB et al (2020). Language models are few-shot learners. arXiv:2005.14165v4 [cs.CL]

Burfoot C, BaldwinT (2009) Automatic satire detection: are you having a laugh? In: Proceedings of the ACL-IJCNLP 2009 conference short papers, pp 161–164.

Burnap P, Williams ML (2015) Cyber hate speech on twitter: an application of machine classification and statistical modeling for policy and decision making. Policy Internet 7(2):223–42. https://doi.org/10.1002/poi3.85

Cao J, Qi P, Sheng Q, Yang T, Guo J, Li J (2020) Exploring the role of visual content in fake news detection. arXiv:2003.05096v1 [cs.MM].

Chakraborty A, Paranjape B, Kakarla S, Ganguly N (2016) Stop clickbait: Detecting and preventing clickbaits in online news media. In: IEEE/ACM international conference on advances in social networks analysis and mining, pp 9–16. https://doi.org/10.1109/ASONAM.2016.7752207.

Cheng L, Li J, Silva Y, Hall D, Liu H (2019) Xbully: cyberbullying detection within a multi-modal context. In: Proceedings of the twelfth ACM international conference on web search and data mining, pp 339–347. https://doi.org/10.1145/3289600.3291037.

Chung YL, Kuzmenko E, Tekiroglu SS, Guerini M (2019) CONAN–COunter NArratives through Nichesourcing: a multilingual dataset of responses to fight online hate speech. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 2819–2829. https://doi.org/10.18653/v1/P19-1271.

Colomina C, Margalef HS, Youngs R (2021) The impact of disinformation on democratic processes and human rights in the world. Policy Depart Director-General External Policies. https://doi.org/10.2861/677679

Common MF (2020) Fear the Reaper: how content moderation rules are enforced on social media. Int Rev Law Comput Technol. https://doi.org/10.1080/13600869.2020.1733762

Crawford K, Gillespie T (2016) What is a flag for? social media reporting tools and the vocabulary of complaint. New Media Soc 18(3):410–428. https://doi.org/10.1177/1461444814543163

Cui L, Lee D (2020) CoAID: COVID-19 healthcare misinformation dataset. arXiv:2006.00885

Cui L, Wang S, Lee D (2019) SAME: sentiment-aware multi-modal embedding for detecting fake news. IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), pp 41–48. https://doi.org/10.1145/3341161.3342894.

Cunha et al (2021) On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. Inf Process Manag 58(3):102481

Danilevsky M, Qian K, Aharonov R, Katasis Y, Kawas B, Sen P (2020) A Survey of the state of explainable AI for natural language processing. arXiv: 2010.00711v1 [cs.CL].

Davidson T, Warmsley D, Macy M, Weber I (2017) Automated hate speech detection and the problem of offensive language. In: Proceedings of the 11th international AAAI social media, ICWSM ', vol 17, 512–515.

Devlin J, Chang M, Lee K, Toutanova K (2019). BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

Dinakar K, Reichart R, Lieberman H (2011) Modeling the detection of textual cyberbullying. In: Fifth international AAAI conference on weblogs and social media, pp 11–17.

Duarte N, Llanso E, Loup A (2017) Mixed Messages? The limits of automated social media content analysis.

Elhadad MK, Li KF, Gebali F (2020) A novel approach for selecting hybrid features from online news textual metadata for fake news detection. 3PGCIC 2019. LNNS 96:914–925. https://doi.org/10.1007/978-3-030-33509-0_86

Ellison NB (2007) Social network sites: Definition, history, and scholarship. J Computer-Mediated Commun 13(1):210–230

Fortuna P, Nunes S (2018) A survey on automatic detection of hate speech in text. ACM Comput Surv 51(4):1–30. https://doi.org/10.1145/3232676

Gambäck B, Sikdar UK (2017) Using convolutional neural networks to classify hate-speech. In: Proceedings of the first workshop on abusive language online, pp 85–90. https://doi.org/10.18653/v1/W17-3013.

Ganesh B, Jonathan B (2020) Countering extremists on social media: challenges for strategic communication and content moderation. Policy Internet 2(1):6–19. https://doi.org/10.1002/poi3.236

Gibbs S (2017) Facebook live: Zuckerberg adds 3000 moderators in wake of murders. https://www.theguardian.com/technology/2017/may/03/facebook-live-zuckerberg-adds-3000-moderators-murders. Accessed 17 October 2021.

Gillespie T (2018) Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media. Yale University Press, New Haven

Gitari ND, Zuping Z, Damien H, Long J (2015) A Lexicon-based Approach for Hate Speech Detection. Int J Multimed Ubiquitous Eng 10(4):215–230. https://doi.org/10.14257/ijmue.2015.10.4.21

Glazkova, A., Glazkov, M., Trifonov, T. (2021). g2tmn at Constraint@AAAI2021: Exploiting CT-BERT and Ensembling Learning for COVID-19 Fake News Detection. In: Chakraborty, T., Shu, K., Bernard, H.R., Liu, H., Akhtar, M.S. (eds) combating online hostile posts in regional languages during emergency situation. CONSTRAINT 2021. Communications in Computer and Information Science, vol 1402. Springer, Cham. doi: https://doi.org/10.1007/978-3-030-73696-5_12

Golbeck et al (2017) A large, labeled corpus for online harassment research. In: WebSci '17: proceedings of the 2017 ACM on web science conference, 229–233. https://doi.org/10.1145/3091478.3091509.

Goldani MH, Momtazi S, Safabakhsh R (2020a) Detecting fake news with capsule neural networks. Appl Soft Comput J. https://doi.org/10.1016/j.asoc.2020.106991

Goldani MH, Safabakhsh R, Momtazi S (2020b) Convolutional neural network with margin loss for fake news detection. Inf Process Manage 58:1–12. https://doi.org/10.1016/j.ipm.2020.102418

Gorwa R, Binns R, Katzenbach C (2020) Algorithmic content moderation: technical and political challenges in the automation of platform governance. Big Data Soc. https://doi.org/10.1177/2053951719897945

Granik M, Mesyura V (2017) Fake news detection using naive Bayes classifier. In: IEEE first Ukraine conference on electrical and computer engineering (UKRCON), pp 900–903. https://doi.org/10.1109/UKRCON.2017.8100379.

Grimmelmann J (2015) The virtues of moderation. Yale J Law Technol. https://doi.org/10.31228/osf.io/qwxf5

Hakak S, Khan AM, S, Gadekallu TR, Maddikunta PKR, Khan WZ, (2021) An ensemble machine learning approach through effective feature extraction to classify fake news. Futur Gener Comput Syst 117:47–58. https://doi.org/10.1016/j.future.2020.11.022

Hamdi T, Slimi H, Bounhas I, Slimani Y (2020) A hybrid approach for fake news detection in twitter based on user features and graph embedding. ICDCIT 2020. LNCS 11969:266–280. https://doi.org/10.1007/978-3-030-36987-3_17

Hirschberg J, Manning HD (2015) Advances in natural language processing. Science 349(6245):261–266. https://doi.org/10.1126/science.aaa8685

Horne BD, Adali S (2017) This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In: The workshops of the eleventh international AAAI conference on web and social media AAAI technical report WS-17, News and Public Opinion, pp 759–766.

Hosseini H, Kannan S, Zhang B,Poovendran R (2017) Deceiving Google's perspective API built for detecting toxic comments. arXiv:1702.08138v1 [cs.LG].

Islam MdR, Liu S, Wang X, Xu G (2020) Deep learning for misinformation detection on online social networks: a survey and new perspectives. Social Netw Anal Mining. https://doi.org/10.1007/s13278-020-00696-x

Kaplan AM, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons 53(1):59–68

Khattar D, Goud JS, Gupta M, Varma V(2019) MVAE: multimodal variational autoencoder for fake news detection. In: The World Wide Web conference (WWW '19). Association for computing machinery, New York, NY, USA, pp 2915–2921. https://doi.org/10.1145/3308558.3313552

Kocoń J, Figas A, Gruza M, Puchalska D, Kajdanowicz T, Kazienko P (2021) Offensive, aggressive, and hate speech analysis: from data-centric to human-centered approach. Inf Process Manag. https://doi.org/10.1016/j.ipm.2021.102643

Koebler J, Cox J (2018) The impossible job: inside Facebook's struggle to moderate two billion people. https://www.vice.com/en/article/how-facebook-content-moderation-works. Accessed on 25 October 2021.

Kumar G, Singh JP, Kumar A (2021) A Deep Multi-modal neural network for the identification of hate speech from social media. IFIP Int Feder Inf Process LNCS 12896:670–680. https://doi.org/10.1007/978-3-030-85447-8_55

Kumar R, Ojha AK, Malmasi S, Zampieri M (2018) Benchmarking aggression identification in social media. In: Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), pp 1–11

Kumari K, Singh JP, Dwivedi YK, Rana NP (2021) Multi-modal aggression identification using convolutional neural network and binary particle swarm optimization. Future Gener Comput Syst 118:187–197. https://doi.org/10.1016/j.future.2021.01.014

Kwok I, Wang Y (2013) Locate the hate: detecting tweets against blacks. In: Proceedings of the twenty-seventh AAAI conference on artificial intelligence, AAAI'2013, pp 1621–1622.

Lan Z et al (2020). ALBERT: A LITE BERT for self-supervised learning of language representations. arXiv:1909.11942v6 [cs.CL]

Leerssen P, Hoboken J V, Harambon J, Lanso E (2020) Artificial Intelligence, Content Moderation, and Freedom of Expression, Transatlantic Working Group.

Li L, Levi O, Hosseini P, Broniatowski D (2020). A multi-modal method for satire detection using textual and visual cues: In: Proceedings of the 3rd NLP4IF workshop on NLP for internet freedom: censorship, disinformation, and propaganda, barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL), pp 33–38

Li L, Levi O, Hosseini P, Broniatowski DA (2021). A multi-modal method for satire detection using textual and visual cues. arXiv:2010.06671v1 [cs.CL]

Liu H, Burnap P, Alorainy M, Williams ML (2019) A fuzzy approach to text classification with two-stage training for ambiguous instances. IEEE Trans Comput Soc Syst 6(2):227–240. https://doi.org/10.1109/TCSS.2019.2892037

Liu Y et al (2019). RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692v1 [cs.CL]

Ma J, Gao W, Wong KF (2019) Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In: Proceedings of the 28th international conference on World Wide Web, ACM: 3049–3055. doi:https://doi.org/10.1145/3308558.3313741.

Malmasi S, Zampieri M (2018) Challenges in discriminating profanity from hate speech. J Exp Theor Artif Intell 30:187–202. https://doi.org/10.1080/0952813X.2017.1409284

Mandl T, Modha S, Majumder P, Patel D, Dave M, Mandlia C, Patel A (2019) Overview of the HASOC track at FIRE 2019: hate speech and offensive content identification in Indo-European languages. In: Proceedings of the 11th forum for information retrieval evaluation (FIRE '19). Association for Computing Machinery, New York, USA, pp 14–17. https://doi.org/10.1145/3368567.3368584.

Mangalam K, Kumar A (2019) Section 66A: an unending saga of misuse and harassment. https://lawschoolpolicyreview.com/2019/06/04/

Mathur P, Shah R, Sawhney R, Mahata D (2018) Detecting offensive tweets in Hindi-English code-switched language. In: Proceedings of the sixth international workshop on natural language processing for social media, pp 18–26. https://doi.org/10.18653/v1/W18-3504.

Mikolov T, Le QV, Sutskever I (2013) Exploiting similarities among languages for machine translation. arXiv:1309.4168v1 [cs.CL].
Mitra T, Gilbert E (2015) Credbank: A largescale social media corpus with associated credibility annotations. Proc Int AAAI Conf Web Social Media 9(1):258–267
Modha S, Majumder P, Mandl T, Mandalia C (2020) Detecting and visualizing hate speech in social media: A cyber Watchdog for surveillance. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2020.113725
Mollas I, Chrysopoulou Z, Karlos S, Tsoumakas G (2021). ETHOS: an online hate speech detection dataset. arXiv:2006.08328v2 [cs.CL].
Mutanga RT, Naicker N, Olugbara OO (2020). Hate speech detection in twitter using transformer methods, pp 614–620. (IJACSA) Int J Adv Comput Sci Appl, 11(9)
Naeem SB, Bhatti R, and Khan A (2021). An exploration of how fake news is taking over social media and putting public health at risk. Health Info Libr J 38(2):143–149. https://doi.org/10.1111/hir.12320. Epub 2020 Jul 12. PMID: 32657000; PMCID: PMC7404621.
Nascimento et al (2022) An overview of systematic reviews of the current state of the art of infodemics and health misinformation and its repercussions in public health: recommendations, challenges, and available research opportunities. Bulletin of the World Health Organization. May 2022.
Naseem U, Razzak I, Hameed IA (2019) Deep context-aware embedding for abusive and hate speech detection on twitter. Australian J Intell Inf Process Syst 15(4):69–76
Nasir JA, Khan OS, Varlamis I (2021) Fake news detection: A hybrid CNN-RNN based deep learning approach. Int J Inf Manag Data Insights 1(1):1–13. https://doi.org/10.1016/j.jjimei.2020.100007
Ngai EWT, Tao SSC, Moon KKL (2015) Social media research: Theories, constructs, and conceptual frameworks. Int J Inf Manage 35(1):33–44. https://doi.org/10.1016/j.ijinfomgt.2014.09.004
Nguyen VH, Sugiyama K, Nakov P, Kan MY(2020) FANG: leveraging social context for fake news detection using graph representation. arXiv:2008.07939v2 [cs.SI].
Nobata C, Tetreault JR, Thomas A, Mehdad Y, Chang Y (2016) Abusive language detection in online user content. In: Proceedings of the 25th international conference on World Wide Web, pp 145–153. https://doi.org/10.1145/2872427.2883062 .
P´erez-Rosas V, Kleinberg B, Lefevre A, Mihalcea R (2017) Automatic detection of fake news. arXiv: 1708.07104.
Paka WS, Bansal R, Kaushik A, Sengupta S, Chakraborty T (2020) Cross-SEAN: A cross-stitch semi-supervised neural attention model for COVID-19 fake news detection. Appl Soft Comput. https://doi.org/10.1016/j.asoc.2021.107393
Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359. https://doi.org/10.1109/TKDE.2009.191
Papakyriakopoulos O, Medina Serrano JC, Hegelich S (2020) The spread of COVID-19 conspiracy theories on social media and the effect of content moderation. The Harvard Kennedy School (HKS) Misinform Rev. https://doi.org/10.37016/mr-2020-034
Park JH, Fung P (2017) One-step and two-step classification for abusive language detection on twitter. ALW@ACL: 41–45. https://doi.org/10.18653/v1/w17-3006.
Patwa P et al. (2021). Fighting an infodemic: COVID-19 fake news dataset. combating online hostile posts in regional languages during emergency situation.In: CONSTRAINT 2021. Communications in Computer and Information Science, vol 1402. Springer, Cham. https://doi.org/10.1007/978-3-030-73696-5_3.
Pennington G, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543. doi:https://doi.org/10.3115/v1/D14-1162.
Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, vol 1, pp 2227–2237. https://doi.org/10.18653/v1/N18-1202.
Pitsilis GK, Ramampiaro H, Langseth H (2018) Effective hate-speech detection in Twitter data using recurrent neural networks. Appl Intell 48(12):4730–4742. https://doi.org/10.1007/s10489-018-1242-y
Qi P, Cao J, Yang T, Guo J, Li J (2019) Exploiting multi-domain visual information for fake news detection. In: 2019 IEEE international conference on data mining (ICDM), pp 517–527. https://doi.org/10.1109/ICDM.2019.00062.
Qian F, Gong C, Sharma K, Liu Y (2018) Neural user response generator: fake news detection with collective user intelligence. In: Proceedings of the twenty-seventh international joint conference on artificial intelligence (IJCAI-18), pp 3834–3840. https://doi.org/10.24963/ijcai.2018/533.
Ofcom Report (2019) Use of AI in online content moderation.
Roberts ST (2016) Commercial content moderation: digital laborers' dirty work. In: Media Studies Publications. 12. https://ir.lib.uwo.ca/commpub/12
Roberts ST (2017a) Content moderation. UCLA Previously Published Works, pp 1–6
Roberts ST (2017b) Social media's silent filter. https://www.theatlantic.com/technology/archive/2017b/03/commercial-content-moderation/518796/ Accessed 17 October 2021.
Robinson D, Zhang Z, Tepper J (2018) Hate speech detection on Twitter: feature engineering v.s. feature selection. In: Proceedings of the 15th extended semantic web conference, pp 46–49, 2018. doi: https://doi.org/10.1007/978-3-319-98192-5_9.
Ross B, Rist M, Carbonell G, Cabrera B, Kurowsky N, Wojatzki M(2016) Measuring the reliability of hate speech annotations: the case of the European Refugee Crisis. In: Proceedings of NLP4CMCIII:3rd workshop on natural language processing for computer-mediated communication (Bochum), vol. 17, pp 6–9. https://doi.org/10.17185/duepublico/ 42132.
Roy PK, Tripathy AK, Das TK, Gao XZ (2020) A framework for hate speech detection using deep convolutional neural network. IEEE Access 8:204951–204962. https://doi.org/10.1109/ACCESS.2020.3037073
Rubin VL, Conroy N, Chen Y, Cornwell S (2016) Fake news or truth? Using satirical cues to detect potentially misleading news. In: Proceedings of the second workshop on computational approaches to deception detection, pp 7–17. https://doi.org/10.18653/v1/W16-0802.
Ruchansky N, Seo S, Liu Y(2017) CSI: a hybrid deep model for fake news detection. In: Proceedings of the 2017 ACM on conference on information and knowledge management, ACM, pp 797–806. https://doi.org/10.1145/3132847.3132877.
Ruckenstein M, Turunen LL (2020) Re-humanizing the platform: content moderators and the logic of care. New Media Soc 22(6):1026–1042
Sahu G, Cohen R, Vechtomova O (2021) Towards a multi-agent system for online hate speech detection. arXiv:2105.01129v1 [cs.AI].
Sanh V, Debut L, Chaumond J, and Wolf T (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108v4 [cs.CL]
Schmidt A, Wiegand M (2017) A survey on hate speech detection using natural language processing. In: Proceedings of the fifth international workshop on natural language processing for social media, pp 1–10. https://doi.org/10.18653/v1/W17-1101.
Sharma S, Agrawal S, Shrivastava M (2018) Degree based classification of harmful speech using twitter data. arXiv:1806.04197.

Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: a data mining perspective. ACM SIGKDD Explorations Newsl 19(1):22–36. https://doi.org/10.1145/3137597.3137600

Shu K, Mahudeswaran D, Wang S, Lee D, Liu H (2018) Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. arXiv:1809.01286v3 [cs.SI].

Shu K, Cui L, Wang S, Lee D, Liu H (2019) DEFEND: explainable fake news detection. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery data mining, pp 395–405. https://doi.org/10.1145/3292500.3330935.

Singh S (2019) Everything in Moderation. https:// newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificialintelligence-moderate-user-generated-content.

Singh JP, Kumar A, Rana N, Dwivedi Y (2020) Attention-based LSTM network for rumor veracity estimation of tweets. Inf Syst Front. https://doi.org/10.1007/s10796-020-10040-5

Singhal S, Shah RR, Chakraborty T, Kumaraguru P, Satoh S (2019) SpotFake: a multi-modal framework for fake news detection. IEEE fifth international conference on multimedia big data (BigMM), pp 39–47. https://doi.org/10.1109/BigMM.2019.00-44.

Singhania S, Fernandez N, Rao S (2017) 3HAN: a deep neural network for fake news detection. ICONIP 2017. Part II, LNCS 10635:1–10. https://doi.org/10.1007/978-3-319-70096-0_59

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30:5998–6008

Verma G, Srinivasan BV (2019) A lexical, syntactic, and semantic perspective for understanding style in text. arXiv:1909.08349 v1 [cs.CL].

Vigna FD, Cimino A, Dell'Orletta F, Petrocchi M, Tesconi M (2017). Hate me, hate me not: hate speech detection on facebook. In: Proceedings of the first Italian conference on cybersecurity (ITASEC17), 86–95.

Vijayarani S, Ilamathi J, Nithya S (2015) Preprocessing techniques for text mining - an overview. Int J Comput Sci Commun Netw 5(1):7–16

Wang WY (2017) Liar, liar pants on fire: a new benchmark dataset for fake news detection. arXiv:1705.00648v1 [cs.CL] .

Wang B, Ding H (2019). YNU NLP at SemEval-2019 task 5: attention and capsule ensemble for identifying hate speech. In: Proceedings of the 13th international workshop on semantic evaluation (SemEval-2019), pp 529–534

Wang Y, Ma F, Jin Z, Yuan Y, Xun G, Jha K, Su L, Gao J (2018) EANN: event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp 849–857. https://doi.org/10.1145/3219819.3219903.

Waseem Z (2016) Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In: Proceedings of the first workshop on NLP and computational social science, pp 138–142. https://doi.org/10.18653/v1/W16-5618.

Waseem Z, Hovy D (2016) Hateful symbols or hateful people? Predictive features for hate speech detection on twitter, In: Proceedings of NAACL-HLT, pp 88–93.

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. https://doi.org/10.1109/ACCESS.2018.2806394

Wendling M (2018) The (almost) complete history of 'fake news'. https://www.bbc.com/news/blogs-trending-42724320. Accessed 12 October 2021.

Wyrwoll C (2014) User-generated content. In: Social media, pp 11–45. https://doi.org/10.1007/978-3-658-06984-1_2.

Yang Z, Wang C, Zhang F, Zhang Y, Zhang H (2015) Emerging rumor identification for social media with hot topic detection. In: 12th web information system and application conference (WISA), pp 53–58. https://doi.org/10.1109/WISA.2015.19.

Yang Z et al (2020). XLNet: generalized autoregressive pretraining for language understanding. arXiv:1906.08237v2 [cs.CL]

Zhang X, Ghorbani AA (2020) An overview of online fake news: characterization, detection, and discussion. Inf Process Manag. https://doi.org/10.1016/j.ipm.2019.03.004

Zhang Z, Luo L (2019) Hate speech detection: a solved problem? The challenging case of long tail on twitter. Semantic Web 1:925–945

Zhang Z, Robinson D, Tepper J (2018) Detecting hate speech on twitter using a convolution-GRU based deep neural network. ESWC 2018 LNCS 10843:745–760. https://doi.org/10.1007/978-3-319-93417-4_48

Zhang Q, Zhang S, Dong J, Xiong J, Cheng X (2015) Automatic detection of rumor on social network. LNAI 9362, NLPCC, pp 113–122. https://doi.org/10.1007/978-3-319-25207-0_10.

Zhang J, Dong B, Yu PS (2019) FAKEDETECTOR: effective fake news detection with deep diffusive neural network. arXiv:1805.08751v2 [cs.SI].

Zhong H, Li H, Squicciarini AC, Rajtmajer SM, Griffin C, Miller DJ Caragea C (2016) Content-driven detection of cyberbullying on the instagram social network. In: IJCAI, proceedings of the twenty-fifth international joint conference on artificial intelligence, pp 3952– 3958

Zhou Y, Yang Y, Liu H, Liu X, and Savage N (2020) Deep learning based fusion approach for hate speech detection, pp 128923–128929.IEEE Access, vol. 8, https://doi.org/10.1109/ACCESS.2020.3009244.

Zhou X, Wu J, Zafarani R (2020b) SAFE: similarity-aware multi-modal fake news detection. In: The 24th pacific-asia conference on knowledge discovery and data mining, LNAI 12085: 354–367, 2020b. https://doi.org/10.1007/978-3-030-47436-2_27.

Zhou X, Zafarani R (2020) A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput Surv 53(5):1–30. https://doi.org/10.1145/3395046