Advances
2023; 4(2): 49-59
http://www.sciencepublishinggroup.com/j/advances
doi: 10.11648/j.advances.20230402.13

Harmful Content on Social Media Detection Using by NLP

Iqra Naz, Rehhmat Illahi

Department of Computer Science and Information Technology, Ghazi University, Dera Ghazi Khan, Pakistan

Email address:
Iqran2756@gmail.com (Iqra Naz), Shahzadineelam842@gmail.com (Neelam Shahzadi)

To cite this article:
Iqra Naz, Rehhmat Illahi. Harmful Content on Social Media Detection Using by NLP. *Advances*. Vol. 4, No. 2, 2023, pp. 49-59.
doi: 10.11648/j.advances.20230402.13

Received: April 30, 2023; **Accepted**: June 12, 2023; **Published**: July 13, 2023

Abstract: Twitter, Facebook and Instagram are the popular social media platforms that allow people to access and connect to a world by a social network to express share and publish information. While online connection via media platforms is immensely desirable and come an unavoidable fact of daily life, the underbelly of social networks may be seen in the form of harmful/objectionable material. Fake news, rumors, hate speech, hostility, and bullying are examples of documented harmful material that are of major concern to society. Such damaging content hurts a negative impact on one's mental health and leads to financial losses that are rarely recoverable. Screening and filtering of such information is thus an urgent requirement. In this paper, we summarize some popular SM like Facebook WHATSAPP, LinkedIn etc. We use some notation like UGC, ML, and AI etc. In this review paper, focuses on methods for detecting harmful parts through natural language processing. The next phase looks at how to moderate this material.

Keywords: Social Media (SM) Platforms, Detection and Moderation, Natural Language Processing (NLP), Artificial Intelligence (AI). Hate Speech Detection

1 Introduction

Through social networking sites, which include placement connections between members of many groups, cultures, and organizations throughout the world, the web has recently changed the information sector. The internet has resulted in a significant shift away from browser browsers and toward social media and tweeting services, which are becoming more and more popular. "A series of World wide web apps that rely on the theoretical and technological underpinnings of Web 2.0, and that enable the production and exchange of User-Generated Content" [1] are how social media is defined. User-generated content (UGC) refers to the numerous types of media material, such as text, video, and graphics that is produced through end customers by a view toward sales promotion.

The UGC is posted on either a website that is available to the public or on an online community that is only available to a specific set of people [2].

1.1. Some Social Media Platforms

Before the web was created, social media started in the year 1844 through a telegraph machine's electrical dots. The first kinds of social media that enabled users to login and communicate with one another were bulletin board systems (BSS). Usenet (USErNETwork), founded in 1979 by (Tom Truscott) and (Jim Ellis), is a type of discussion forum where users may express their opinions on subjects that interest them. All group members have access to the item in question [4]. Six Degrees is regarded as the original social networking platform, comparable to Facebook, which had millions of people signed up. In 1991, Live Journal, a website for writing blogs or weblogs, sprang to prominence. Social networking sites, blogs, forums, and other types of SM were all available [5].

Table 1. *Most famous social media platforms.*

Name	Category	Year	Characteristics
LinkedIn	Social Site	2003	It is a Professional networking websites that connect business. People used it for professional networking and career development. It provide job opportunities.
Facebook	Social Site	2003	It allows users to stay connected with friends and family. Users can chat, upload pictures, share videos and links, post and read status, comment and reacts on it.
YouTube	Sharing Site	2005	It allows users to upload their content and share it with friends or provide it to the public.

Name	Category	Year	Characteristics
WhatsApp	Messaging App	2009	enables user to send text and voice messages, share images and videos
Instagram	Social Site	2010	It is used to share photos and videos.
Twitter	Blog Site	2006	It enables users to read and post short messages called as tweets. A tweet consists a text with limited character, a photo and video format.

Table 1 lists the well-known SM platforms that are now an essential part of every person's life. The categories of SM allow users to exchange information in a variety of ways, as illustrated in Table 1.

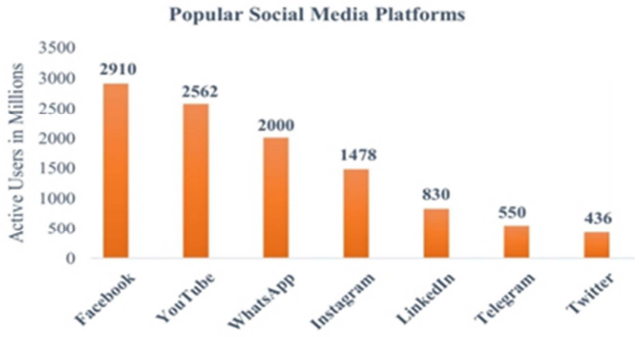

Figure 1. *Popular social media platforms.*

Figure 1 displays data on the SM platform's daily active users through the year 2022. The network with the most users is Facebook. Facebook has around 2.93 billion monthly active users in the first quarter of 2022. SM may also be a tool for a variety of internal and external organizational tasks, including knowledge sharing, corporate development, marketing campaigns, and cooperative shared knowledge, with peers, clients, and other businesses [6]. A platform for local companies to market their brands and connect with users around the world is shown by the 43% of users who use social media platforms to conduct product searches online. Linked career advancement activities, job prospects, and businesses-to-businesses and industry linkages, for instance. Additionally, users may submit texts and videos on anonymous online social mobile applications including Whisper without disclosing their identities. Users can express their ideas in real-time on various areas of social, political, financial, ethical, and environmental concerns using online social media platforms. These systems allow users to post content known as User Generated Content UGC) [7] in the form of text messages, Photographs, videos, jokes, and audio. UGC is referred to by terminology like "posting," "tweets," "posts," "reviews," and "retweets"[8]. The user-generated material may be both beneficial and harmful at different times. A person's overall health may be negatively impacted by the material on social media sites, which is also causing economic losses. These platforms are used to screen students for placement prospects. A significant increase in user-generated content (UGC) on social media platforms over the past several years has had a significant influence on social environment.

1.2. Black Side of Social Media

Twitter, YouTube, and Facebook are a few of the well-known and extensively used social networking sites that allow users to express themselves, exchange information, and connect to a limitless world [9]. Due to quick, simple accessing information and the opportunity to express oneself in a variety of ways, SM platforms have seen a significant surge in usage in recent years [10]. Through to the development and dissemination of UGC that is provocative, provocative, and frightening, this freedom of expression [11] is misused. Through the rapid growth in the spreading of harmful information, social media has recently become a problem for the entire globe. On social media, uploading and releasing information with the aim of hurting or upsetting a person or a group is referred to as harmful content.

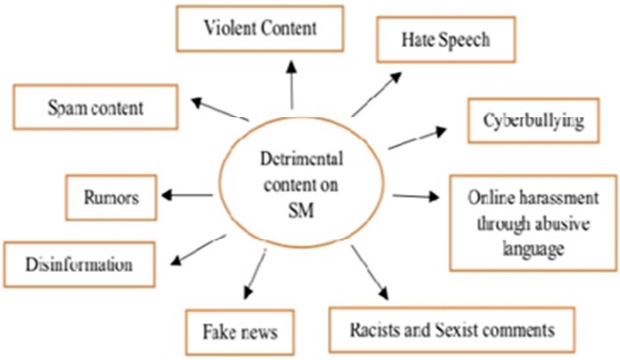

Figure 2. *Different forms of Detriment content published on social media.*

Figure 2 shows the harmful types of user-generated content, such as bold speech [7], fake news [12], harassment [3], disinformation [13].

Table 2. *List of notations used in this paper.*

Notations	Description
DL	Deep Learning
UGC	User Generated Content
NLP	Natural language processing
SM	Social Media
AI	Artificial Intelligence
LM	Language Model
GRU	Gated Recurrent Unit
API	Application Programming Interface
ANN	Artificial Neural Network

2. Review Methodology

To examine the work conducted by scholars on the topic of SM content moderation, a methodical approach is used to evaluate the pertinent literature. The basic process makes up the literature methodology:

1) Outlining the study's inquiries.
2) A compendium of pertinent issues drawn from current publications and scientific research.
3) Connecting the research issues to the data gathered

from the literature.

Only studies published between 2011 and 2021 were included in the study's literature review. Gathering the publications from Google Scholar, IEEE, Springer, Elsevier, and AAAI digital libraries is the first stage of the study's goal. Duplicate articles weren't included because Google Scholar includes content from all publishers. The highest number of reports issued for the article and the abstract of each of the 500 publications connected to social video content were read. Searching the digital books database for publications using keywords such as "Material moderation on social media," "User-produced content on social media," and "Need of content moderation" will help you find them. Since the focus of this research work is on the identification and moderation of harmful social media information, a Google Scholar search for literature on this topic turned up articles on the identification of hate or bold speech, fake news, and cyberbullying data. This led researchers to look into topics like "Detection of dangerous social media content using Natural Language Processing," "Machine learning and Learning Techniques for Hate Speech/Fake news/rumors," "NLP for Hate Speech/Fake news/rumors detection," and "Hate Speech/Fake news/rumors sensing using machine learning and deep learning techniques" on libraries. Most of the articles for a query on social content moderation were taken from the social science field.

2.1. Research Objectives

In-depth research on SM content identification and moderating approaches is included in the report. The study's primary research goals are to: The study's main research goals are to:

1) Examine the databases used to find harmful content.
2) Conduct a comparison of different Language Models and algorithms used to find damaging information on social media platforms.
3) Review the methods for avoiding harmful conversation.
4) Recognize the challenges and knowledge gaps associated with the various described strategies for UGC identification and moderating.

2.2. Research Questions

To achieve the study goals, the following research questions have been developed.

1) What are the numerous techniques for spotting harmful deception on social media sites?
2) What are the definitions of content moderation and methods used on social media platforms?
3) What are the described strategies for content identification and restraint's problems?

2.3. Theoretical and Practical Implications of Study

The assessment of the literature has revealed that there has been a substantial number of studies done regarding how to identify different types of harmful information. Theoretically speaking, published publications have concentrated more on the many components of human moderating and the difficulties that AI-based solutions must overcome. Less research has been conducted on completely automated strategies for removing harmful data from social media platforms.

3. Datasets

Datasets are a crucial source of information in a format of a table. The data in the datasets comprise articles, URLs, phrases, publisher data, social interactions, and retweets acquired via social media sites in the setting of damaging form. On the given datasets, a variety of ML algorithms are being tested for the identification of false news, hateful speech, and phrases associated with it.

The comment boards or postings from various social media platforms are extracted to create datasets for false news. The datasets were developed with the assistance of linguists and media professionals. The postings and comments are examined by human professionals, who classify them as true or false.

Table 3. *Some famous datasets for fake news detection.*

Datasets	Features	Categories	Skewness
LIAR	First dataset for deception detection 12.8 K user labeled short statements evaluated PolitiFact.com Data gathered from news releases, TV, Facebook posts, etc.	False (pants-fire) False Barely true Half true Mostly true True	Highly imbalanced
BUZZFEED NEWS	2000 news samples published on Facebook during 2016 US Presidential elections Each post and linked articles were checked by 5 journalists Metadata information such as URL of the news post.	Mostly true Not factual content, Mixture of true and false Mostly false	Highly imbalanced
FAKENEWSNET	Highlights on social behavior of fake news 212 fake news and 212 true news Data like publisher information, news content gathered from fact checking websites BuzzFeed.com and PolitiFact.com	True Fake	Balanced

The list of characteristics that may be retrieved from the data sets for the identification of false news is compared in Table 3. According to table 3, the majority of data focuses on the news's editorial content, which may not be enough to effectively identify false news. Sets of data like Daily Post and fake news contain both the metadata and the news data elements that have been the subject of several academic studies. The metadata contains user profiles, social data, and other data on how people interact with the news [15]. In comparison to previous datasets, the LIAR dataset has significantly larger comments, and it also includes speaker meta-data [14]. The LIAR datasets also include a wide range of subject areas, including the federal budget, the economic, health, taxes, school, jobs, and state budgets, as well as candidate biographies, elections, and migration. To facilitate multi-level categorization, some databases also label news stories.

Table 4. *Well-Known datasets for hate speech finding.*

Datasets	Features	Categories	Skewness
Davidson	24,802 tweets from Hate base Contain large number of ethnicity content Collection on offensive keywords	Hate speech-7%, Not offensive Offensive but not hate speech	Highly Imbalanced
Storm front	Textual bold speech annotated at sentence level 10,568 sentences have been extracted from	Hate Not hate Relation Skip	Imbalanced
KAGGLE	8832 social media comments 20,362	Insulting Non insulting	Imbalanced
Williams and Matthew	136,000 tweets from Twitter Annotations by experts (feminists and anti-racism activists) and crowd- source workers	Racist Bold None Both	Imbalanced

The datasets for different categories of hate or bold speech are compiled in Table 4. The databases of hate speech contain material in both mono and multilingual formats, together with score labels [4] for each aspect of hate speech. A statistic known as inter-annotator agreement [17] is used to measure how well various annotators have categorized hate speech. The number of annotations that agree on a certain job of annotation is defined by the inter-annotator disagreement. Fleiss's Kappa () is statistical indicator that describes how well annotators label content [4] and Krippendorff's alpha addresses missing annotations.

Two metrics are used for datasets where a higher value of the metric denotes a greater level of agreement. For instance, [1] reported a = 0.26 for 1687 comments tagged by 5 annotators for 2 kinds of bold speech: mild hatred and high hate which demonstrates the difficulties in the assignment. [17] Reported a = 0.26 for 56,280 abusive remarks that were analyzed by 3 professional raters [4] Reported a = 0.84 with 85% disagreement for racism annotations. The process of inter-annotator consensus becomes too difficult since hate speech is so incredibly subjective. Numerous studies have shown the creation of databases that categorize content as offensive, abusive, profane, racist, or just hateful. For instance, Davidson [4] claimed that 76% of the language was objectionable and 5% was hate speech. When used in conjunction with other sentences, the label "relation" in [5] denotes a hate speech sentence, whereas the label "skip" denotes a phrase containing bold or non-bold speech.

The inter-annotator agreements are crucial in building the datasets for bold speech since it influences how well an ML system performs. Twitter is the chosen media site for data extraction and data preparation in the area of false news and hates speech. The viewpoint of the annotator, who labels material and provides context information, determines how datasets are created. Due to a user's propensity for writing posts in many languages and using mixed coding. Due to the user's propensity for posting in international and code-mixed forms (native language printed in Roman), researchers have also developed datasets in mixed languages (Urdu + English) [4] that are utilized for the machine learning architectures-based identification of hate speech. Human annotators are used to annotate this type of information, and the inter-annotator agreement is computed. The size of the data, the amount of accuracy, and the number of labels allocated to the text in the datasets all have an impact on how well deep neural network models function. Few dubious and ambiguous instances of bold speech [4] that were too difficult for user annotators to judge were recorded in research articles. Such situations weren't taken into account when creating the dataset.

4. Detection of Detrimental UGC on SM

The job of identification is to locate harmful or undesirable information in posts or texts that users have posted on social media platforms. Finding harmful content online involves spotting false information, hate speech, and verbal abuse. The material on SM platforms is initially detected before being moderated. The manual method of identification is not scale able given the volume of information released on social media (for instance, average of 6000 tweets are sent on Twitter8 per second). The ability to recognize and remove objectionable or damaging UGC using machine learning (AI) has become crucial. For the purpose of identifying harmful UGC, a variety of AI approaches, including machine learning algorithms, and Natural Language Processing, is used [1].

According to research publications, AI-based methods have detected harmful information on SM platforms with the greatest speed and accuracy. The manually and artificial intelligence based ways of identifying harmful data on social media platforms are discussed in this section.

4.1. Manual Method of Fake News Detection

A process of determining whether posted stuff is true or false is called fact-checking. Fact-checking only categorizes information as true or untrue rather than evaluating it as objectionable [1].

Fact-checking sites utilize real people who are professionals in media to examine the accuracy of the news. Experts who use a method to evaluate information are referred to as fact-checkers. Some methods used by fact-checking websites are:

1) A topic or a statement to be investigated is selected through articles, political advertisements and speeches, campaign websites, social media, TV, and interviews.
2) When researching on statements, fact-checkers often use basic methodology, various sorts of sources, as well as formal rules that guide their methods.
3) Claim assessments are techniques that fact checkers use to assess the availability of claims.

For automated detection of false news information, fact-checking websites like Politifact9 have developed databases and made them available to the public. These sites offer a professional analysis of verified news, including a list of items that are fraudulent and an explanation of why [1]. More than 60 fact-checking groups are sent false information by social media sites like Facebook, yet most of them only devote a small number of researchers to look into Facebook posts [1]. The laborious process of fact-checking to find fake news is a difficult effort. In the detection phase, factors like the amount of time required to verify the news and the understanding of the context around the fake news must be taken into account.

The user base conducts the identification of various types of harmful content, including bold speech and abusive language, to express their concerns about the content posted on SM platforms [4]. The identification of such content runs the danger of the user introducing bias. The method of detection won't be sufficient given the exponential growth of harmful content.

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 [17].

Because of advancement in algorithms, computing power, and information, the recent decade has witnessed remarkable gains in AI [2]. Deep Learning is a machine learning field that uses Artificial Neural Systems to handle huge amounts of data. Natural language processing (NLP) is a branch of AI that uses computers to parse text [18]. Natural Language Processing (NLP) is a branch of computer linguistics that use computational methods to learn and interpret human language [18].

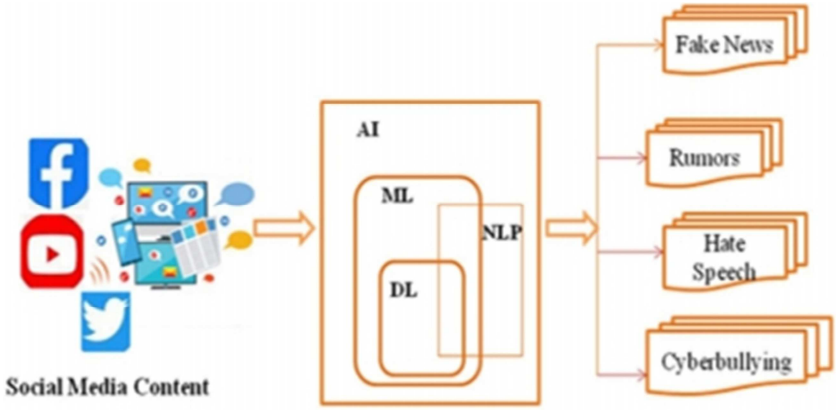

Figure 3. *Artificial intelligence techniques for detection of detrimental data on social media.*

Figure 3 depicts an AI-based technique for detecting harmful information on social media. A substantial amount of research has been conducted on the application of AI-based algorithms for detecting fake news, disinformation, verbal abuse, and hateful words on social media. The objective of automated identification of UGC utilizing NLP, ML, and DL algorithms is to categorise online comment threads as detrimental (including hateful speech, violent, toxic, rumors, and cyberbullying) or acceptable material. NLP has opened up a new range of possibilities for automating the linguistic language structure in the formation of speaking transcription engines, mining social media for health-related data or finance, and recognizing behavior and emotion forward into product lines and services [11], sifting offensive material and improving spam filters [19], and creating chatbots for customer support [5].

Advance features of NLP have played main role in finding of detrimental data or information on social media. NLP artificial intelligence are mainly used to process the text-type online comments on social media [20]. In context of content

moderation, NLP tools are used to process the online text, extract the features from text which are used to find the harmful forms data like fake news, bold speech, and disinformation.

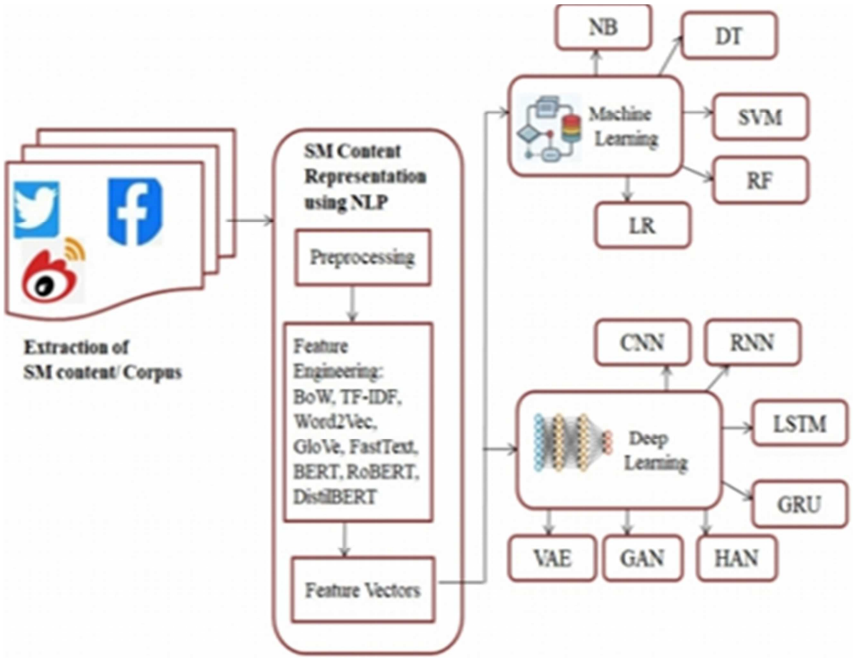

Figure 4. *Diagram of automated social media content detection using NLP, ML and DL.*

Figure 4 shows a simplified block structure of UGC detection. To analyze the online information posted on SM platforms, NLP technologies are used. As seen in Figure 4, SM content extraction entails acquiring online comments and postings using an Application Programming Interface or crawler ways given by social media. To obtain data, twitter provides two tools [10]. A corpus is constructed that includes all types of SM material in monolingual and multilingual formats, as well as metadata like as location, user profiles [11, 12]. This corpus was generated with the assistance of experts and crowd-sourced employees who labeled information as normal or harmful [2]. Researchers have made major contributions to the establishment of a dataset that encompasses all categories of harmful information such as false information, rumors, hate speech, and cyberbullying content. NLP technologies are used to extract the comment characteristics from the corpus. Words, phrases, characters, and unique phrases [21] are examples of qualities that vary based on the type of material to be processed. Many feature representation approaches, such as the Bag of Words (BoW), Term Frequency-Inverse Articles, n-grams [21], Word2Vec [17], GloVe [22], and Deep Bidirectional Representation from Transformer (BERT) [23], translate text characteristics from the content to vectors of real values. The feature vectors acquired after utilizing NLP tools to analyze the SM content are fed into a classifier model, which can be a non-neural classifier or a cognitive prototype [19]. Based on attributes collected from SM material, classify models are utilized to find harmful data. The use of supervised machine leaning algorithms such as Support Vector Machines, Logistic Regression, Nave Bayes, and Random Forest and deep networks such as CNN architectures: Convolutional Neural Networks, and sequenced neural network models: Recurrent Neural Networks, Long Short-Term Memory, Gated Recurrent Unit, Converter models, Classifier Auto-encoder. The non-neural and neural network models [19] are trained using various feature representation approaches on various characteristics collected from labeled datasets. For detection or classification, the trained network is used to test data. A multiclass classifier or a classification can be used [21]. DL algorithms that operate with large amounts of data have the benefit of automatically identifying the characteristics for categorization that a machine learning algorithm performs through human intervention [24]. Given the volume of SM information, neural networks have shown an excellent method for automatic social media information identification.

4.2.1. Role of NLP for Detection of Detrimental Content on SM

The manual method of parsing the incredible amount of SM material is difficult in terms of the period needed to comprehend the chaotic and loud text, as well as the pricey training of the moderators to analyze such text. An automatic text-parsing technology is called (natural language processing). Through the use of expanded language models that have already been trained, NLP has achieved significant strides in text feature representation techniques. Various feature presentation approaches, including frequency-based methods and embedding's based on neural networks, are used to turn the raw text characteristics into numeric feature vectors.

According to scientific study publications, these techniques are used to find harmful information on social media. The preparation of the material Data and data mining is the area that deals with the analysis and processing of SM

data. The method of collecting knowledge and information from unorganized and cluttered material is called text mining [1]. Since UGC is unorganized, processing SM material might be difficult. The UGC on SM is frequently garbled and published informally [16], with sentences or texts lacking punctuation, other abbreviations, emoticons (like:-), other characters (like "@@Sush", "U9","##happy"), and the use of many repeated characters (like "coool" or "haaa") throughout the text. It is extremely difficult to read material with this unclear form of content. Therefore, pre-processing is an essential step in converting such unstructured input into a form that can be successfully analyzed. Improvement technology an important factor in the success of NLP text classifiers is feature engineering, also known as feature selection and representation [1]. Words, phrases, characters, and unique words [1] that vary based on the type of material to be processed might be among the attributes. The selection of characteristics for SM content is influenced by the lexical, syntactic, and semantic components of the text. At the level of words, the lexical components might be expressed in formal, informal, or subjective ways [25]. The sequence of words and paragraphs that make up a sentence is referred to as its grammatical elements [19]. To determine the meaning of the statement, one of the semantic components is to determine the attributes [7]. Semantic components can be used to analyze the text's feelings. The choice of the extra features is likewise made using the text's associated meta-data. Among these are multimedia information, statistics about consumers and follows, and spatial information that describes the context of the content [26]. The news title and body text of the news piece may be used to extract the lexicon, semantic, and syntactic aspects of false news and rumors. The image/video attribute may be used to retrieve the image characteristics. Negative language is indicative of hate speech. [5]. It is able to derive the vocabulary, grammar, and semantic aspects of an online hate letter from its brief text, usage of distinguishing words that set it apart from other messages, use of special characters, punctuation, information to all stakeholders, and other characteristics [3, 5]. In terms of word choice, typing dependence, and the use of other characters like hashtags (#), (@), grammar, etc. false news and hateful speech material share several lexical, syntactic, and surface features.

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. Figure 5 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.

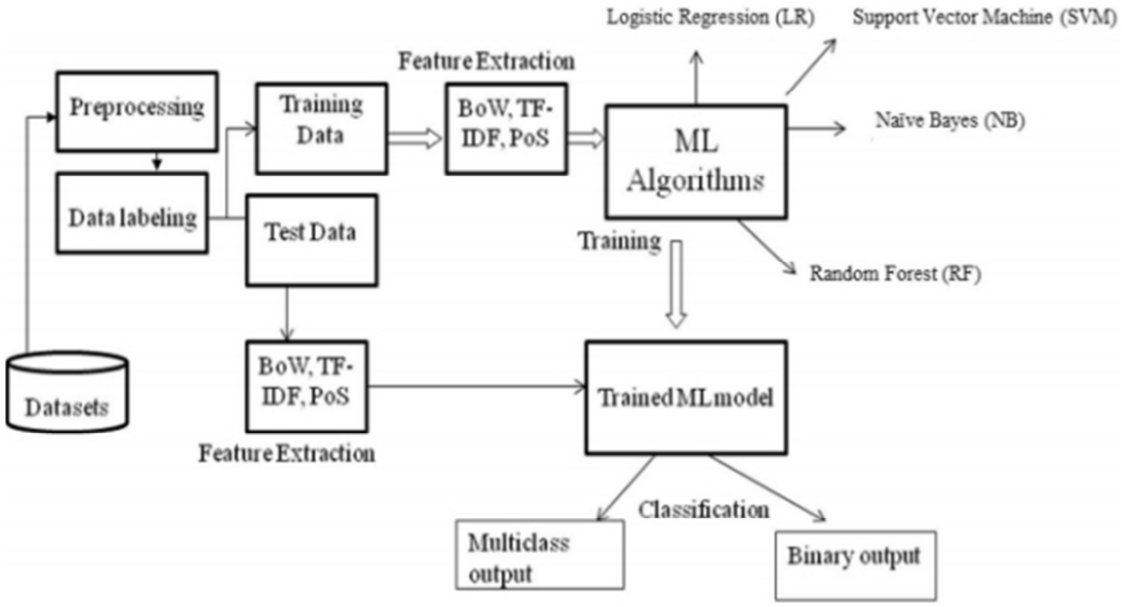

Figure 5. *Detection and classification of a social media data through machine learning.*

As illustrated in Figure 5, the trained machine learning system is applied for the testing information for classification. The categorization might be manual, such as labeling content as offensive, hateful, or non-hateful [21], or it can be binary. Consider the categorization of true and false news [21]. The performance of the ML algorithm is assessed using datasets including massive amounts of data gathered from prominent social media networks such as Facebook, Twitter, and Instagram. Traditional approaches for the identification of social media information are machine learning algorithms. The constructed features employed by Algorithms are time consuming, insufficient,

and labor costly, and the effectiveness of an ML algorithm is reliant on the characteristics used for categorization. Deep Learning is a subset of machine learning, has piqued the interest of business for a variety of reasons. The fundamental structure of deep learning is a neural network comprising an input layer, one or more hidden layers, and an output layer [4].

4.2.3. Multimodal Approach of Detecting Detrimental Content on SM

Multimedia is a crucial modality and characteristic that can help with the regulation of social media material. The media files consist of pictures, videos, text and other form of format. The growth of multimedia technology has transformed the paradigm of text-only news stories into news articles that also contain photos and videos that draw in more readers [27]. For instance, a tweet with a picture receives 89% more likes and receives 11 times as many retweets as a post without the need for an image [27]. Fake photos that are linked to news stories have become more in recent years. According to [27] incorrect visual material can take the shape of altered pictures, deceptive pictures, and visual effects with false claims, as seen in Figure 6.

Figure 6. *Some images of fake news.*

Investigating the many features of the false picture [20], as these features differ from those of a real picture, is necessary for the identification of fake news from visual content. These traits make up the characteristics, which are retrieved to assess the accuracy of the picture and comprise statistical features [20]. DCT was used to research with forensics features, converting the picture from the video and image to the frequency response. CNN was used to collect the image's many semantic characteristics, and a bidirectional GRU network was used to simulate the sequential relationships among such features. With the use of those two qualities together, false news could be identified with an accuracy of 84.6%. To identify bogus news, [14] experimented with forensic characteristics and gathered descriptive and inferential statistics. The forensic features were merged with user-based features and content-based features, and the results indicated recall, accuracy, and F1-score of 0.749, 0.994, and 0.854 respectively. On a media platforms, the user posts information in a variety of formats, including text, pictures etc. This modality is also seen in the spreading of hate speech and bogus news on SM platforms. A combination of text and visual graphics offers more information and aids in the detecting procedure.

5. Moderation of Detrimental Content on SM Platforms

Every year, there is a significant increase in the use of social media for illegitimate purposes, which poses problems for several industries as well as the public sector and civil society [14]. The spread of this content has continued despite legal restrictions the government has imposed to curb the devastatingly harmful information on SM. On SM platforms, content identification and moderation are therefore crucial. The academic focus has been attracted to content moderation on online platforms because of the publication of several study publications in scholarly journals. Publishing industry platforms use text filtration to prevent the posting of certain words or types of content, as well as other explicit moderation techniques, to detect modest content. They do this by checking the content against known facts [14], reducing the existence of harmful content [14], removing offensive or disrespectful material, deleting or removing posts, banning users by their username and Internet protocols address, and other clear and specific moderation actions. Government- and civil society-established government agencies participate in content moderation [14].

Content moderation is implemented by SM companies in three discrete phases namely [14].

1) Creation: Creation refers to the process of creating the terms of service (or regulations) that platform employments to regulate how the user interacts.
2) Enforcement: Enforcement involves flagging problematic content that is a problem, determining if the content breaches the guidelines established during the development stage, and then deciding what action should be taken for the tough situation.
3) Reaction: Reaction details the inner appeals procedure employed by platforms as well as the strategies of action activists may employ to alter the platform from afar. For instance, social media corporations took note of the controversy surrounding the live streaming of

murder and announced the employment of extra moderators to better monitor such occurrences. This phase explains the manual, somewhat automated, or fully automated ways of moderating.

5.1. Manual Approach of Moderating Detrimental Content on Social Media

According to [29], considered alongside is the process of using administrators or moderators who have the power to delete material, block users, and decide how a society's members interact with each other. For Social media, material control is regarded as essential [20]. The safety of SM platforms is ensured in large part by content moderators [20]. The content moderators make the decisions on what sort of material belongs on SM and what stuff has to be taken off.

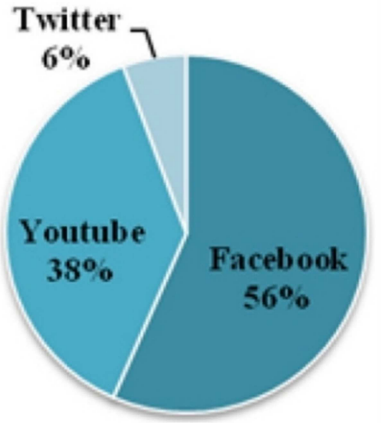

Figure 7. *Content moderation on social media.*

The statistics of moderator employed by well-known SM platforms are displayed in Figure 7 [2]. According to Figure 7, Facebook has the most moderators with over 15,000, then YouTube with 10,000, and Twitter with about 1500 [14]. The numbers reflect both the volume of content posted on these sites and the number of moderators assigned to content vetting. Social media companies have marginalized people to scale up with the growing volume of content and have outsourced the task of moderate amounts to third-party vendors who operate in different countries, including the United States, the Philippines, India, Ireland, Portugal, Spain, Germany, Latvia, and Kenya [1]. Additionally, online services like Amazon Mechanical Turk are used for moderation [29].

The user base uses flagging, a detection mechanism, to alert SM platforms to objectionable or graphically violent content [30]. Artificial intelligence based techniques are used to detect harmful material to scale with the content uploaded on SM [21]. The SM platforms frequently employ the fagging function, which enables users to voice their concerns about the information uploaded on these platforms [3]. The platform's community guidelines and regulations are then examined through the content moderators, who determine if a flagged content violates violation of them [3]. Many SM platforms value user-generated content since it supports the maintenance of their brand [3]. Because content administrators only have to assess flagged content rather than all comments, the flagging method lessens their workload.

5.2. Semi-Automated Technique of Moderating Detrimental Content on Social Media

Amount, veracity, and pace of the tough situation that must be assessed, as well as cultural, historic and geographical context among the content, provide several difficulties for the manual method of content moderation. Automation techniques are being proposed by several businesses and governments to aid in the detection and analysis of problematic content, such as misinformation, hate speech, and terrorist propaganda [30]. People moderators then examine the flagged content. The effort of human reviewers will reduced by the automatic flagging system. The review procedure for human moderation is made much easier by AI-based technologies like hashing match, in which an image's fingerprint is checked with a database that contains known hazardous pictures, and "keyword filtering," in which terms that signal potentially dangerous material are used to flag content [31]. Microsoft's Cloud content moderator is an AI-based content moderation solution that automatically applies content filters to text, images, and videos. The content is saved and shown via the web-based Effective online so that human moderators may evaluate it [12]. A tool contain moderating Programming Interface that scans for unacceptable material such as foul language, sexually explicit or suggestive material, and language. It also scans for adult or racy photos and videos in photographs and movies. According to up opportunities or experience level, the review tool allocates or elevates material evaluations to several review teams [31].

The effectiveness of semi-automated material filtering strategies depends more on the precision of the AI algorithms used to manipulate material and images. The level of variety employed in social networking UGC should also be detected by AI algorithms, as this is difficult to do and need additional study. It is necessary to put the automatic fagging mechanism and then in actual evaluate what those technologies help human being moderating method. More AI-based fagging methods should be used to identify damaging words or images and provide a flag that indicates horrifying or obscene material that should be reviewed by a person moderator.

6. Conclusion

However, automated social media content analysis is where artificial neural network-based NLP models fall short, despite their impressive results in object identification, sentimental analysis, and translation software at the moment. It is crucial to create various types of concepts can convey all subtleties of speech in the unique contexts that the research needs to look at. Harmful social media posts have already damaged the business. The present approaches focus on reducing or eliminating it after the damage has already

occurred. However, scientists have to check outside content moderation and then go further to avoid this when a user gets some fagging after choosing a threshold for the number of improper postings. A platform that tracks a user's record of sharing potentially dangerous material; sets a limit on the number of inaccurate postings, and sends alerts when the limit is surpassed may be developed to build a secure social media environment. The exponential growth of hazardous social media sites is characterized in large part by the ability to precisely detect such information. The volume of harmful information is increasing, and technological detection mechanisms cannot keep up. Recent advances in AI have paved the way for the automation of online media identification thanks to modern algorithms, computing power, and the ability to manage large volumes of data. NLP techniques have done a good job of parsing the particular format of the social media material. To obtain character and word-level properties from the source material and convert them into input vectors, NLP mainly depends on feature engineering techniques and tagging.

References

[1] V. U. Gongane, M. V. Munot, and A. D. Anuse, Detection and moderation of detrimental content on social media platforms: current status and future directions, vol. 12, no. 1. Springer Vienna, 2022. doi: 10.1007/s13278-022-00951-3.

[2] M. S. Ahmed, M. H. Sharif, N. Ihaddadene, and C. Djeraba, "Detection of Abnormal Motions in Multimedia," Chania ICMI-MIAUCE, vol. 8, no. May 2014, 2008.

[3] J. Ma, W. Gao, Z. Wei, Y. Lu, and K. F. Wong, "Detect rumors using time series of social context information on microblogging websites," Int. Conf. Inf. Knowl. Manag. Proc., vol. 19-23-Oct-, no. October, pp. 1751–1754, 2015, doi: 10.1145/2806416.2806607.

[4] A. Al-Hassan and H. Al-Dossari, "Detection of Hate Speech in Social Networks: a Survey on Multilingual Corpus," pp. 83–100, 2019, doi: 10.5121/csit.2019.90208.

[5] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus," ACM Int. Conf. Proceeding Ser., pp. 1980–1984, 2012, doi: 10.1145/2396761.2398556.

[6] F. Alkomah and X. Ma, "A Literature Review of Textual Hate Speech Detection Methods and Datasets," Inf., vol. 13, no. 6, pp. 1–22, 2022, doi: 10.3390/info13060273.

[7] J. Robinson et al., "Social media and suicide prevention: A systematic review," Early Interv. Psychiatry, vol. 10, no. 2, pp. 103–121, 2016, doi: 10.1111/eip.12229.

[8] C. Emma Hilton, "Unveiling self-harm behaviour: what can social media site Twitter tell us about self-harm? A qualitative exploration," J. Clin. Nurs., vol. 26, no. 11–12, pp. 1690–1704, 2017, doi: 10.1111/jocn.13575.

[9] C. Laorden, B. Sanz, G. Alvarez, and P. G. Bringas, "A threat model approach to threats and vulnerabilities in on-line social networks," Adv. Intell. Soft Comput., vol. 85, no. October, pp. 135–142, 2010, doi: 10.1007/978-3-642-16626-6_15.

[10] G. E. Hine et al., "Kek, cucks, and god emperor Trump: A measurement study of 4chan's politically incorrect forum and its effects on the web," Proc. 11th Int. Conf. Web Soc. Media, ICWSM 2017, pp. 92–101, 2017.

[11] A. T.. Shahjahan and K. U. Chisty, "Social Media Research and its Effect on Our Society," Soc. Media Res. Its Eff. Our Soc., vol. 8, no. 6, pp. 1–5, 2014.

[12] A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin, "Identification and characterization of cyberbullying dynamics in an online social network," Proc. 2015 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2015, pp. 280–285, 2015, doi: 10.1145/2808797.2809398.

[13] R. Produced and O. N. Behalf, "Use of ai in online content moderation 2019," 2019.

[14] Y. Wang et al., "Eann," pp. 849–857, 2018, doi: 10.1145/3219819.3219903.

[15] S. Shama*, S. W. Akram, K. S. Nandini, P. B. Anjali, and K. D. Manaswi, "Fake Profile Identification in Online Social Networks," Int. J. Recent Technol. Eng., vol. 8, no. 4, pp. 11190–11194, 2019, doi: 10.35940/ijrte.d9933.118419.

[16] Y. Chen, S. Zhu, Y. Zhou, and H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescents," Proc. Int. Conf. Privacy, Secur. Risk Trust, p. 71_80., 2012, [Online]. Available: http://www.cse.psu.edu/~sxz16/papers/SocialCom2012.pdf

[17] P. Burnap and M. Williams, "Hate Speech, Machine Classification and Statistical Modelling of Information Flows on Twitter: Interpretation and Communication for Policy Decision Making," Internet, Policy Polit., pp. 1–18, 2014, [Online]. Available: http://orca.cf.ac.uk/id/eprint/65227%0A

[18] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook," CEUR Workshop Proc., vol. 1816, pp. 86–95, 2017.

[19] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on Twitter," WebSci 2017 - Proc. 2017 ACM Web Sci. Conf., pp. 13–22, 2017, doi: 10.1145/3091478.3091487.

[20] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial behavior in online discussion communities," Proc. 9th Int. Conf. Web Soc. Media, ICWSM 2015, pp. 61–70, 2015.

[21] H. Ahmed, I. Traore, and S. Saad, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10618 LNCS, no. December, pp. 127–138, 2017, doi: 10.1007/978-3-319-69155-8_9.

[22] B. Srinandhini and J. I. Sheeba, "Online social network bullying detection using intelligence techniques," Procedia Comput. Sci., vol. 45, no. C, pp. 485–492, 2015, doi: 10.1016/j.procs.2015.03.085.

[23] D. Ramalingam and V. Chinnaiah, "Fake profile detection techniques in large-scale online social networks: A comprehensive review," Comput. Electr. Eng., vol. 65, pp. 165–177, 2018, doi: 10.1016/j.compeleceng.2017.05.020.

[24] S. Singhania, N. Fernandez, and S. Rao, "3HAN: A Deep Neural Network for Fake News Detection," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10635 LNCS, pp. 572–581, 2017, doi: 10.1007/978-3-319-70096-0_59.

[25] O. De Clercq, S. Schulz, B. Desmet, E. Lefever, and V. Hoste, "Normalization of Dutch user-generated content," Int. Conf. Recent Adv. Nat. Lang. Process. RANLP, pp. 179–188, 2013.

[26] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? The challenging case of long tail on Twitter," Semant. Web, vol. 10, no. 5, pp. 925–945, 2019, doi: 10.3233/SW-180338.

[27] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," 25th Int. World Wide Web Conf. WWW 2016, pp. 145–153, 2016, doi: 10.1145/2872427.2883062.

[28] C. Van Hee et al., "Automatic detection and prevention of cyberbullying," Int. Conf. Hum. Soc. Anal. (HUSO 2015), no. c, pp. 13–18, 2015, [Online]. Available: https://biblio.ugent.be/publication/7010768/file/7010781.pdf

[29] J. M. Xu, K. S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," NAACL HLT 2012 - 2012 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf., pp. 656–666, 2012.

[30] P. Gal, I. Santos, and P. Garc, "Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network : Application to a Real Case of Cyberbullying".

[31] W. Akram and R. Kumar, "A Study on Positive and Negative Effects.