ISSN: 2583-9292 (ONLINE)
DOI: 10.21917/ijdsml.2025.0184

A SYSTEMATIC LITERATURE REVIEW OF YOUTUBE COMMENTS SENTIMENT ANALYSIS: CHALLENGES AND EMERGING TRENDS

Raj Kumar Singh[1] and Ani Thomas[2]
[1]*Department of Computer Science and Engineering, Chhattisgarh Swami Vivekanand Technical University, India*
[2]*Department of Information Technology, Bhilai Institute of Technology, India*

Abstract

In the current era of information, many social networking sites have established and evolved into an essential aspect of today's society. As the second most popular social media platform, a huge amount of online content is generated and shared by YouTube and provides novel perspectives for corporations, policymakers and researchers. However, getting valuable information effectively from the massive volume of data has become increasingly difficult. Sentiment analysis offers an automated way to examine opinions, views, and sentiments in written language. Many researchers have sought to improve the effectiveness of several sentiment analyzers or use information from social media to apply them to multiple domains. This research investigates the difficulties experienced by scholars and emerging challenges in examining sentiment analysis in YouTube comments. It provides knowledge regarding the sentiment evaluation task's objectives, the implementation procedure, and how it is applied in various application areas. Additionally, it compares multiple studies and identifies several issues with the analysis techniques, the datasets, the languages of text, and the criteria for evaluation. This study aims to contribute to sentiment analysis research and can assist researchers in adopting an appropriate approach for their potential uses.

Keywords:

YouTube, Comment, Sentiment Analysis

1. INTRODUCTION

1.1 YOUTUBE AND SENTIMENT ANALYSIS

With the increasing popularity of numerous social media sites, YouTube draws much user attention. In recent years, YouTube has rapidly grown and significantly influenced people's everyday lives. Many people use it to post about their lives and leave feedback on various services, goods, events, and organizations. As a result, an unlimited amount of data is generated by viewers [1]. Extracting and using valuable information from user-generated data is crucial for people, organizations, and governments.

The popularity and spread of internet technologies, especially YouTube, have brought forth several challenges for data management [2]. This platform is multi-domain, multi-lingual, and multicultural since users from diverse nations may post videos and comment in various languages on various subjects. According to Aliaksei Severyn's research [3], 60-80% of YouTube comments contain opinions. Therefore, the industry and the research community are very interested in a reliable sentiment analysis technique in such a setting. It is the reason why YouTube is the main subject of our study.

The process of automatically extracting views, opinions, and notions from the written text is called sentiment analysis [4]. The study of sentiment is being used in many different fields, including education [5], [6], movies [7], products [8], [9], politics [10], and more. All of these applications in various sectors demonstrate the value of sentiment analysis in generating insightful data on public sentiment around targeted areas of importance.

In recent literary work, many researchers have attempted to enhance the capabilities of different sentiment classification methods or apply them to other fields using information from social media. This research uses the PRISMA (i.e. Preferred Reporting Items for the Systematic Review and Meta-Analysis) methodology to summarise the outcomes from analyzing the literature on sentiment analysis on YouTube comments. It addresses difficulties and additional probable issues that academics may have encountered and offers answers. This review article can help scholars and professionals solve sentiment analysis-related concerns more successfully and serve as a foundation for future research planning.

1.2 CONTRIBUTIONS

This work contributes to the literature in numerous ways by analyzing and comparing investigations into sentiment analysis approaches and their applications. Firstly, this study provides examples of additional issues that researchers may face, such as data collecting and preparation, models or algorithms, and assessment measures.

Due to the widespread usage of YouTube as a social networking site, the volume of information generated by users continues to rise dramatically, requiring more personnel, effort and time to manage the increasing size and variety of data. On the other hand, corporations and government agencies must monitor public mood quickly. However, most research focuses on model correctness while ignoring practical issues such as execution time. This review article emphasizes, in particular, that it is not enough to assess an algorithm's performance alone, as well as its adaptability and efficiency.

Furthermore, we outline these issues and suggest solutions for future reference. By citing this article, they could focus on resolving these issues and building an improved and successful research strategy.

Subsequently, by summarising the previous works in Table 1, this review study assists researchers in formulating methodological considerations based on earlier research findings. Determining vitally important information about data, methods, assessment measures, and domains is partly made possible by this table. Each factor listed in Table.1 can help researchers more effectively respond to their practical sentiment analysis-related queries and serve as a foundation for future research planning.

1.3 SENTIMENT ANALYSIS PROCESS

The sentiment analysis procedure is outlined in this section. Even though multiple authors adopted different techniques, Fig.1 depicts the basic structure of sentiment analysis. The stages for sentiment analysis in YouTube social media include:

- Selection of a YouTube channel or video uploaded where the user intends to extract sentiments.
- Data extraction process, in which users begin by using certain keywords, depending on their choices, to obtain the information they seek.
- Pre-processing, where the collected data is pre-processed and makes it ready for the subsequent step. This stage involves extracting features, tokenization, and cleaning.
- Analysis of data, in which every pre-processed piece of information is used for what it was created for, such as identification of polarity, opinion analysis and frequency evaluation.

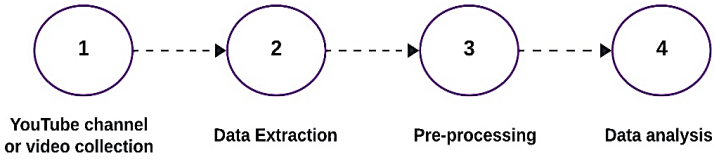

Fig.1. Sentiment analysis steps

The systematic literature review methodology used in this work is outlined in Section 2. Section 3 discusses the various strategies used for the task of sentiment analysis. Section 4 depicts languages used in the reviews, posts or comments assessed by several sentiment classifiers, and Section 5 examines the issues and challenges associated with models, assessment measures and datasets. The paper is concluded in Section 6, which also highlights its contributions and limits.

2. SYSTEMATIC REVIEW PROTOCOL

Scholars may find worldwide evidence, authenticate current practices, settle disagreements, and find new approaches through systematic reviews [11]. Additionally, systematic reviews may be used to explore conflicting findings, define and advise future research areas, and direct decision-making. A systematic literature review was conducted for this study to get a more thorough and better grasp of the context of the investigation on sentiment evaluation in YouTube comments.

There are five steps in the review protocol (Fig.2), including the selection of studies, the search strategy, the extraction of data, and the analysis and discussion phases [12], [13], [14], [15].

2.1 IDENTIFICATION OF RESEARCH QUESTIONS

Developing highly efficient sentiment analyzers and their applications has received much research attention, as it has been proven that sentiment analysis techniques enable people, businesses, and authorities to benefit from the quantity of useful content present in raw social media data [13]. Nevertheless, it might be challenging for those interested in selecting the most significant one for their application or research due to the wide range of data, algorithms, and models employed in these researches and the wide variety of fields covered. To make future sentiment analysis research more useful in resolving practical issues, it is also vital to examine and debate the difficulties experienced by academics and to pinpoint future problems.

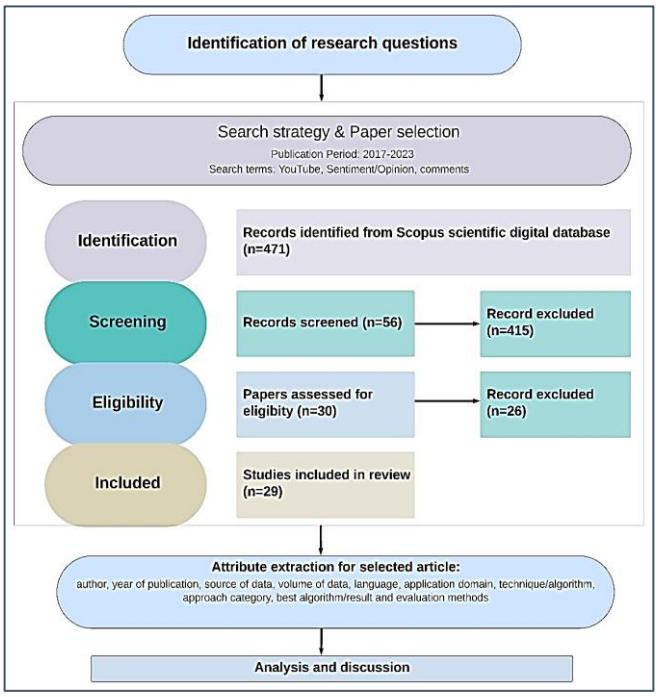

Fig.2. A review protocol

Therefore, we choose the research questions listed below for this literature review to fulfil these objectives:

- What patterns have been discovered in studies on sentiment analysis of YouTube comments?
- Which sentiment classification algorithms, assessment criteria, and datasets are most commonly used while analyzing YouTube comments in diverse disciplines?
- What difficulties do writers encounter when running sentiment analysis on YouTube comments datasets?
- What potential problems may there be with the current research on the sentiment analysis of YouTube comments?

2.2 INFORMATION SOURCE AND SEARCH STRATEGY

The recommended reporting items for systematic review and meta-analysis (PRISMA) statements, presented in Fig.2, served as the foundation for the search approach used in this study [16]. A search phrase has been created to retrieve all publications connected to the study questions. While identifying pertinent papers, we examined Scopus digital databases. The publishing timeframe was limited to the years 2017 through 2023. The search was started in the advanced search fields of the Scopus database. The search terms used include YouTube comments, YouTube comments sentiment, YouTube comments opinion and YouTube comments emotion. Models using conjunctive and disjunctive Boolean operators (i.e., AND, OR) were used to search. The search term, for example, in the Scopus database is: (YouTube comments OR YouTube reviews) AND (Sentiment OR Opinion

OR Emotion). The content was found and filtered based on different publishing articles, conferences, and review articles. This choice effectively covered the most recent and pertinent publications in the review's chosen topic.

2.3 SELECTION CRITERIA

In the identification phase, 471 records were retrieved from Scopus scientific digital database and other sources. To determine the relevance of articles, titles and abstracts were scanned in the next step. All retrieved articles were scrutinized extensively by reading their titles and abstracts to ensure they met the inclusion criteria discussed in section D. In the final round, completely matched articles were included. The full-text reading was completed in the third step. Consequently, 29 publications were ultimately included in the inclusion stage of our assessment for additional and in-depth analysis.

2.4 INCLUSION CRITERIA

Peer-reviewed academic publications from 2017 to 2023 that were gathered from digital sources and have something to do with the tone of comments on YouTube social media are also included. The more particular version was always included if research appeared in multiple publications.

2.5 EXCLUSION CRITERIA

Articles that don't employ machine learning techniques and aren't published in computer science publications are removed, even if they are relevant to the topic. Conference papers that describe ongoing or unfinished research are also not accepted. Also eliminated are journal publications that do not fully or sufficiently describe their methods.

2.6 SUMMARY OF ARTICLE

This step aims to summarise the papers from the inclusion stage. The retrieved data may be used for identifying gaps in the research and, finally, to establish research questions. Each of the chosen studies took information on the author, publication year, data sources, data volume, review language, field of application, algorithm/method utilized, approach type, best algorithm and matrix of evaluation for the proposed approach. The characteristics were then summarised in a summary table with appropriate classes.

Each characteristic was chosen for a specific purpose. Citations were used to identify the author and the publication year for browsing and further referencing by readers who might be interested. To explore the data selection procedures employed by the researcher to analyze comments sentiment, the following four features, namely data source, volume, review language, and application domain, were extracted. Additionally, a deeper knowledge of the applications and approach to sentiment analysis is provided by discussing aspects, such as algorithm/technique used, approach type, best algorithm/result, and assessment metrics.

3. BIBLIOMETRIC ANALYSIS

The statistical data (publication trends, country statistics, keywords, etc.) can be usefully analyzed using bibliometric analysis [17]. This approach is used to identify the best research and scholars in an area of interest. Researchers are helped in recognizing the correlations between explanatory values and other elements of the discovered studies by the approach utilized in the bibliometric analysis and by mapping and graphically displaying the data together with figures. Statistical data analysis is done in this work to help people to understand as many articles as possible.

3.1 PUBLICATION AND CITATION TREND

The yearly trend of articles published and total citations on sentiment analysis of YouTube comments is shown in Fig.3. The graph also shows that interest in sentiment analysis of YouTube comments has increased. There has been a significant rise in research publications between 2017 and 2022. In 2023, the articles were collected only till July. The percentage of publications in 2023 might exceed that in 2021 and 2022.

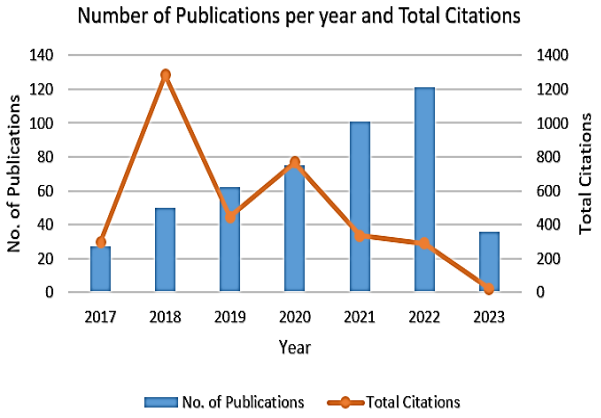

Fig.3. Publication per year and total citations

3.2 MOST RELEVANT WORDS AND TREND TOPICS

In general, systematic reviews examine the conceptual framework of the articles under consideration.

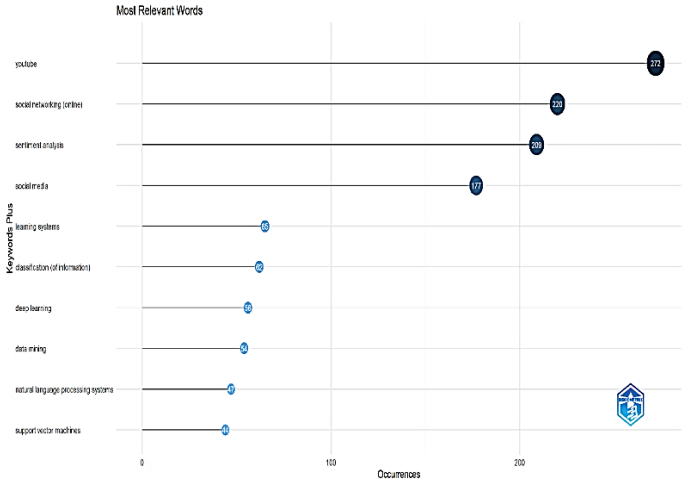

Fig.4. Most relevant words

Bibliometrix includes several more features, such as creating matrices for co-citation, coupling, cooperation, and co-word analysis [18]. The analysis was carried out in R using RStudio

[19]. We have utilized the Biblioshiny app that enables users to execute relevant bibliometric and visual studies via an interactive web interface, significantly decreasing user information input intensity and use threshold [20]. It allows the determination of the most frequently used keywords by the researcher in the study. Fig.4 shows the top ten most popular terms in studying YouTube comment sentiment. According to the chart, the authors' keywords mostly focus on youtube, social media, sentiment analysis and machine learning. The commonly used terms were utilized to create the word cloud in Fig.5.

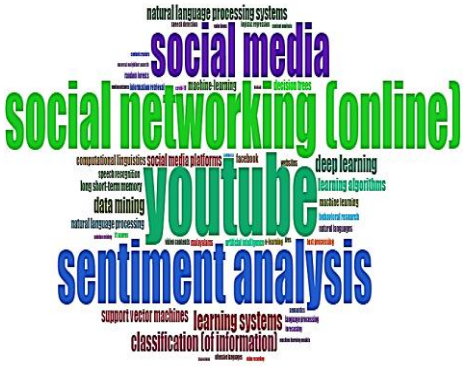

Fig.5. Words cloud

A trending topic analysis was also carried out based on the keywords from the articles extracted from the Scopus database. This research provides insight into the year-to-year trends in term appearances in YouTube comment sentiment analysis literature. The article's keywords are shown in Fig.6 in a topical hierarchy. Trend topics can help researchers to identify new and emerging research areas that may be worth exploring. For instance, in 2021, "youtube" and "sentiment analysis" were the most discussed topics. However, after one year, in 2022, "youtube comment" was the most discussed topic.

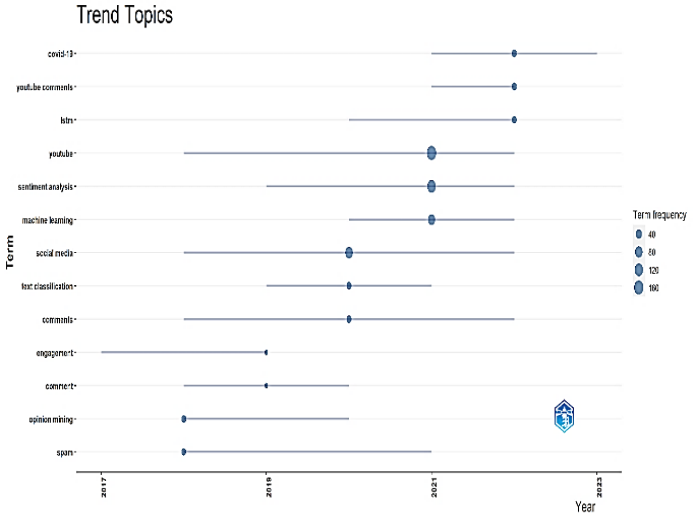

Fig.6. Trend topics

3.3 ANALYSIS OF CO-OCCURRENCE (CO-WORD ANALYSIS)

The development of network theory, an advancement of network analysis, illustrates the links among authors, keywords, and publications to reveal their relationships. Additionally, network analysis aids in determining the degree of linkage between domains and potential consequences [21]. The CSV file was retrieved from the Scopus repository and utilized as input data for network analysis in the VOSviewer program [22].

To understand the fundamental intellectual issue addressed by the existing study, a co-occurrence study was performed using 548 authors and 3434 keywords, with the minimum number of occurrences of a keyword being 10. The keywords shown in Fig.7, youtube and social networking, appeared important in the most integrated network, which can be understood by examining the interlinking lines among the keywords. These lines show the thickness and significance of the linkages between each node (keywords) [23].

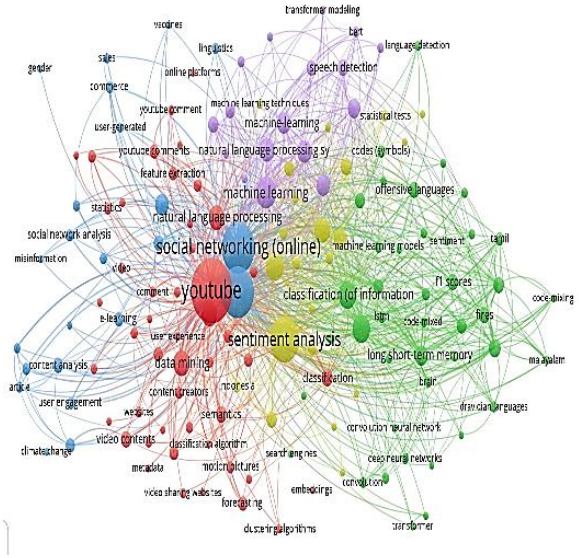

Fig.7. Network diagram for all keywords. Note: Threshold criteria of a minimum of 10 keywords

3.4 COUNTRY CO-AUTHORSHIP ANALYSIS

The co-authorship analysis of papers by country reveals that 37 countries can be grouped into 9 main clusters. According to the node size shown in Fig.8, the United Kingdom (cluster 5), Indonesia (cluster 6), the United States (cluster 3), and India (cluster 7) are the most influential countries in this study. The fact that these clusters are more firmly connected in the network also suggests that the articles from these countries are cited more frequently.

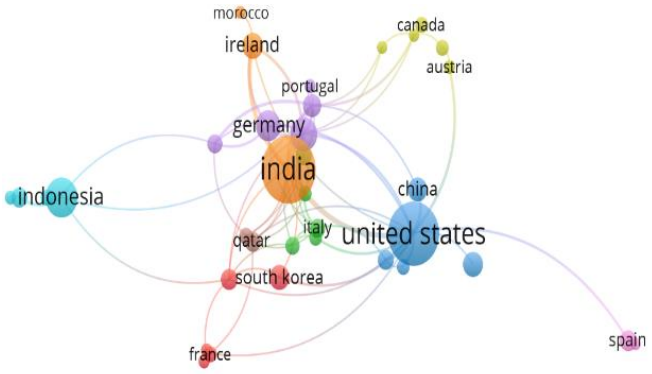

Fig.8. Country contribution analysis from the Scopus database

4. TAXONOMY

This section gives our taxonomy, which summarises the findings of our search. The articles were first searched, examined, and filtered. Next, reading the entire text of all chosen articles n = 29 was performed. All papers were divided into three main groups: Machine Learning (ML), lexicon-based, and hybrid approaches combining the first two categories.

4.1 MACHINE LEARNING-BASED APPROACHES

This category included the first collection of research articles in the taxonomy that covered the machine learning-based method and its use in sentiment analysis of YouTube comments. For the most part, ML-based classification was used for research, and according to this study's taxonomy, 20 out of 29 studies (69% of all studies) were of this category.

In the first investigation, Nguyen et al. [24] suggested the CoNBiLSTM word embedding (i.e., convolutional N-gram BiLSTM) approach, which expresses a term in long and short-distance intervals with contextual and semantic information. Authors have also applied CoNBiLSTM word embedding to identify comment type and its negative, neutral or positive polarity. This study compared the model's effectiveness using the SenTube dataset, including comments on automobile and tablet domains in English and Italian. The research's goal was to extract opinions from multi-domain and multi-lingual platforms. The experimental result indicates that CoNBiLSTM often performs better than the most recent cutting-edge sentiment analysis on the SenTube dataset.

For automatically gathering, filtering, and analyzing YouTube comments for a certain product, Mai et al. [25] presented a unique system that can help to perform comment-type categorization, create a combined sentiment analysis model to achieve both sentence and aspect levels, and finally build Vietnamese datasets. The authors made the YCSA dataset, i.e., YouTube Comments for Sentiment Analysis, which contains 2153 Vietnamese comments about smartphone products. Several experiments were performed on various architectures, viz. ELMO, BERT and BGRU to choose the best technique for JSA (Joint Sentiment Analysis). These experiments showed that the BERT model delivered the best results that achieved an accuracy of 91.45%.

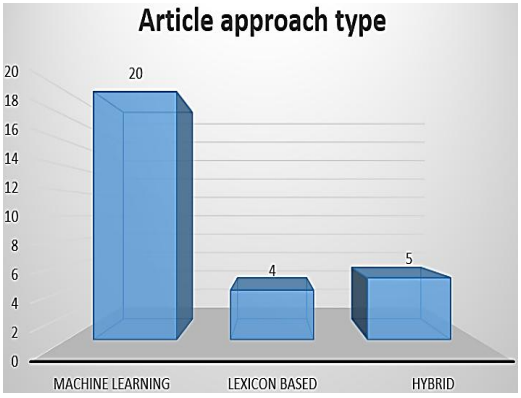

Fig.9. Number of articles for each type of strategy

Ahmad et al. [7] proposed a model that can be applied to predict movie revenue. This research indicates a film's box-office revenue by analyzing people's intent to buy a movie ticket based on trailer reviews. According to the authors, this study has practical significance because it may be utilized for decision-making and business intelligence. Furthermore, the strategy takes advantage of trailer opinions before the premiere of a film. As a result, it allows moviemakers to modify their film or distribution plans. Reviews for 29 movies released in 2016 and 2017 were collected and uploaded on the official YouTube channel of the film production firm. The comments on movies released between 2016 and 2017 and uploaded by official YouTube channels were extracted with the online YouTube Comment Scrapper tool. This experiment assesses the effectiveness of SVM, MLR, multilayer perceptron NN, and RF in predicting box office revenue relative to the suggested strategy. Based on the correlation coefficient with box office revenue, the experiment has shown that MLR has the strongest correlation.

In social media analysis, spam is defined as information or contact that creates an unpleasant experience and makes it difficult to locate more relevant and significant stuff. On YouTube, it can occasionally be used to deliver unwanted messages randomly. Oh [26] suggested an approach to identify YouTube spam comments. YouTube has a mechanism to filter spam, but it is frequently ineffective. The author used an openly available dataset that contains reviews from five well-known music videos. The experiment uses six machine learning algorithms and ensemble models with soft and hard voting, as the author tests several strategies to determine the best classification algorithm. The experimental findings demonstrated that the ensemble with soft voting (ESM-S) model suggested in this study performed optimally across most assessment metrics.

In another study, Röchert et al. [27] examined the informational uniformity of misleading data on YouTube concerning the ongoing COVID-19 outbreak. This study examines a dataset of 10,724 videos and 2,585,367 comments about COVID-19 obtained on YouTube between January and March 2020 (corresponding to the outbreak's start). Random YouTube comments and videos that matched the search term "coronavirus" were annotated by the author. The author used a mix of NLP and network analysis to calculate the homogeneity of information. The research found that despite minor variations in the amount of disinformation on YouTube during the three-month study, around one-third of the videos contained misleading data of some kind. The findings suggest that the users or channels spreading misinformation were generally part of diverse conversation networks, where most of the content discussed was not misinformation.

In non-English speaking regions, individuals on social networks usually write and interact using a combination of languages known as code-mixing [28]. In another research, Subramanian et al. [29] conducted various examinations to identify possibly abusive comments on YouTube, provided by the HASOC-Offensive Language Identification track in Dravidian Code-Mix FIRE 2021. Different machine learning techniques and neural network models were examined, including BNB, SVM, LR, and KNN. KNN demonstrated the highest accuracy at 81.651%. Feature extraction was performed using TF-IDF.

Additionally, transformer-based and BERT-based models that don't require explicit feature extraction were created. These models were fine-tuned or equipped with adapters to reduce the

number of parameters to be trained. XLM-RoBERTa (Large) had the highest accuracy at 88.53%. The proposed models reveal that transformer models, with a special emphasis on pre-trained adapter models, exhibit better performance when compared to machine learning models.

Chakravarthi [30] discusses the importance of promoting positivity in online forums, specifically through hope speech, to foster compassion and acceptable social behaviour. As a result, the author developed the HopeEDI dataset, which comprises user-generated comments from YouTube. These comments, totalling 28,451 in English, 20,198 in Tamil, and 10,705 in Malayalam, were manually categorized as either containing hopeful speech or not. They seek to further study hope speech and positive material in online social media to benefit equality, diversity, and inclusion. They feel that this dataset will help future research promote positivity. The experiment showed that the decision tree produced higher macro-F-Scores for English and Malayalam, while the logistic regression model demonstrated good performance, specifically with Tamil.

A technique that can extract and classify opinions from real-time YouTube comments on recipes by Benkhelifa Randa and Laallam [31] has been proposed in another study. The system collects comments on culinary recipe videos and then uses a Support Vector Machine (SVM) classifier to differentiate them from objective texts, reaching a precision rate of 93.4%. Subjective texts or opinions can be effectively classified as positive or negative using a separate SVM classifier with an impressive 95.3% accuracy. According to the author, the system's functionality includes organizing cooking recipes into two categories: "recommended" and "not recommended," and the ability to compare different recipes.

Akhter et al. [32] have created a dataset of offensive comments in Urdu from social media. They used different feature extraction techniques at the word and character levels. They tested seventeen classifiers from seven machine-learning approaches to detect abusive language in Roman Urdu and Urdu text comments. The results indicate that using character-level n-grams and certain models like LogitBoost and SimpleLogistic achieve high accuracies in detecting offensive language. The dataset is publicly available on GitHub for further research.

In another research, Chakravarthi et al. [33] suggest the utilization of MPNet and CNN in a multi-lingual model to identify offensive language targeted at individuals or groups in low-resource Dravidian languages. This research aims to categorize social media code-mixed postings and comments in Kannada, Tamil, and Malayalam, evaluating their neutral or offensive content across various levels of severity. The model can handle information with mixed coding, including situations where Tamil and Latin scripts are used together. The author's study found that by validating their model on multiple datasets, they were able to beat the other conventional models in detecting objectionable language, with weighted average F1-scores of 0.85, 0.98, and 0.76 for Tamil, Malayalam, and Kannada, respectively. The proposed model outperformed the standard models of Ensemble With Decision Tree (EWDT) and Ensemble Without Decision Tree (EWODT) in terms of accuracy by 0.02, 0.02, and 0.04 for Tamil, Malayalam, and Kannada, respectively.

The abundance of offensive content on social media platforms is a concern that adversely affects internet users. Homophobia/transphobia is the term used to describe feelings of fear, aversion, unease, or prejudice towards individuals who identify as lesbian, gay, transgender, or bisexual. The spread of hate speech online against LGBT+ individuals has raised significant alarm, underscoring the presence of homophobia and transphobia in the virtual realm. To address such issues, Chakravarthi et al. [34] have proposed a new hierarchical taxonomy and created expert-labelled datasets for automatically identifying homophobic/transphobic content. The dataset consists of 15,141 annotated multi-lingual comments, made using comprehensive annotation rules and educated annotators due to the topic's sensitivity. This study describes the process of building the dataset, qualitative analysis of the data, inter-annotator agreement, and baseline models are described in this paper. According to the author, this dataset is the first of its kind. The study finds that Random Forest with BERT embedding performs the best for English and Tamil-English code mixed settings. In contrast, for the Tamil language, Random Forest with fastText embedding performs the best with an accuracy rate of 0.994.

The spread of toxic information on social media channels outweighs medical experts' important advice regarding the rapidly spreading pandemic like COVID-19. In a study, Obadimu et al. [35] discuss the issue of toxic content on social media platforms, particularly regarding the COVID-19 pandemic. They describe strategies for analyzing harmful information and identifying those who spread it on YouTube. The authors employed topic modelling and social network analysis to identify dominant subjects, emerging patterns, and important YouTube commentators. They also performed toxicity studies to assess the network's overall health. The report emphasizes the importance of social media firms and policymakers taking action against toxic users. It presents experimental proof that eliminating these users can lessen the network's toxicity. Overall, the article shows how social media platforms may address the issue of toxic COVID-19 content and promote community well-being. The dataset used in the study includes 544 channels, 3,488 videos, 453,111 commenters, and 849,689 comments. To identify the key topics and evolving trends present in the comments of these videos, the researchers employed the topic modelling technique using Latent Dirichlet Allocation (LDA).

Social media platforms like YouTube now provide more ways to make money and, as a result, have attracted a larger user base, including malicious actors who leverage automated bots to distribute spam messages across multiple channels. In another investigation, Aiyar et al. [36] describe a new technique for recognizing comments on YouTube that either contain promotional material or are irrelevant to the video. These spam comments damage a channel's reputation and disrupt the experience of normal users. YouTube's current blocking of comments with links has proven ineffective as spammers find ways to bypass them. Although standard machine learning classification algorithms have some effectiveness, there is still room for improvement. The authors applied machine learning algorithms like Random Forest, Support Vector Machine, and Naive Bayes, along with custom heuristics like N-Grams, which have proven successful in detecting and combating spam comments. It is evident that Support Vector Machines and Random Forests outperform other traditional Machine Learning algorithms in classification performance and are particularly effective in handling datasets with many dimensions. The

experimental result shows that SVM has the highest accuracy of 0.9841.

Research on political deliberation in the digital sphere has become a priority in various disciplines. The study proposed by Luengo et al. [37] examines how social media has influenced political discussions on COVID-19 in Spain, Italy, and the United Kingdom. The researchers collected and analyzed 111,808 YouTube comments using automated analysis and machine-learning techniques. They hypothesized that polarization would be higher in Southern European countries like Spain and Italy compared to the United Kingdom, which follows a more liberal model. The results confirmed this hypothesis, with Spain and Italy showing higher levels of polarization. Additionally, the study found that other users in Mediterranean countries often support and reward strongly polarised attitudes. In this study, the authors utilized an unsupervised technique called the latent Dirichlet allocation (LDA) algorithm.

In another research, Gao et al. [38] examined the difficulties of using natural language processing (NLP) to predict suicide risk in Cantonese social media posts. Cantonese, a combination of Traditional Chinese, borrowed characters, and English, is a language that is undeveloped in NLP. The study investigates various text mining approaches, including SVM, AdaBoost, Random Forest, and LSTM, to classify Cantonese comments on YouTube for suicide risk. Re-sampling and focus loss approaches are applied to overcome the data imbalance. The LSTM algorithm produces the most promising results, with a testing classification accuracy of 84.3% and g-mean accuracy of 84.5%. This study illustrates the feasibility of recognizing suicide risk in Cantonese social media messages automatically.

Several researchers are using the attention mechanism to collect exceptional feature vectors and improve the effectiveness of deep learning algorithms. A study by Uddin et al. [39] focuses on sentiment analysis, an essential aspect of evaluating online content, mainly on YouTube. The researchers propose a new approach combining deep learning techniques, specifically using an Encoder-decoder-based Attention model with a squeeze-and-excitation attention layer. Totalling 700,000 user comments from 8,000 YouTube channels were collected to train and test the algorithm. Positive, negative, or neutral sentiments are assigned to each statement in the dataset. Results of the experiment suggest that the model performs better than other machine learning-based approaches, with accuracy and F1-scores of 92.8% and 91.9%, respectively. Further, the study shows how the attention mechanism affects sentiment analysis. The results of this study provide a broad paradigm for future sentiment analysis across languages.

In another study, Saifullah et al. [40] have prepared a dataset that aims to detect public anxiety about government programs related to the COVID-19 pandemic by analyzing their social media comments. The authors have applied Machine learning techniques such as K-NN, Bernoulli, Decision Tree Classifier, Support Vector Classifier, Random Forest, and XG-Boost to analyze sentiments in the comments. The data sample consists of 4862 YouTube comments, with 3211 negative (indicating anxiety) and 1651 positive (indicating hope) comments. Count-vectorization and TF-IDF techniques are used in the feature extraction stage of the training of machine learning models. The experimental findings suggest that Random Forest outperforms competing models, with an accuracy rate of 84.99% using count-vectorization and 82.63% using TF-IDF. XG-Boost achieves the best recall, whereas K-NN achieves the best precision. Therefore, Random Forest is the most effective model for detecting anxiety in social media data.

The use of online social networks (OSNs) to find news and information has grown. They have negative aspects in addition to benefits like simple communication and rapid update. For OSNs, recognizing fake news is a serious difficulty yet unsolved [41]. Choi et al. In another research, Choi and Ko [40] describe a unique technique for recognizing fake YouTube content that conveys incorrect information. The system estimates the subject distribution of a video based on its title/description and comments using adversarial learning and topic modelling. The model uses the BTM (Biterm Topic Model), which is more capable because conventional topic models cannot perform well with brief texts. Additionally, it efficiently uses an adversarial neural network to extract topic-neutral characteristics. The suggested model performs better than earlier models for identifying fake news videos, scoring 3.41% higher on the F1 scale.

In another research, Savigny et al. [42] conducted experiments on emotion classification on Indonesian Youtube comments. The authors collected 8,115 YouTube comments and manually labelled them with six fundamental emotions (happy, sad, angry, surprised disgusted, and fearful) and one neutral label. This study compared the average word vector, average word vector with TF-IDF, paragraph vector, and the convolutional neural network (CNN) algorithm as ways to use word embedding in a classification job. The tests revealed that word embedding using the CNN technique, which had an accuracy of 76.2%, gave the best results. The study's findings suggest that word embedding using the CNN method is a potential way for categorizing emotions in Indonesian YouTube comments.

From the text, determining emotion or mood has advanced from a straightforward frequency distribution analysis to more intricate learning techniques [43]. Another research by Ezpeleta et al. [38] shows that analyzing a text's mood makes it possible to identify spam comments more accurately. The authors' experimentation using a social spam dataset improved the best accuracy attained with the original dataset from 82.50% to 82.58% using the YouTube Comments Dataset and from 93.97% to 94.38% using the validation dataset. Additionally, an average of 13.76% and 11.47% fewer false positives were detected. The experimental outcome indicates that mood analysis can distinguish between spam and real social media comments. The authors suggested including the mood feature in all video categories, whether a positive or negative sentiment, enables classifiers to identify spam comments and enhance the overall outcome effectively.

Another research on Indian cooking channels was carried out by Kaur et al. [44] to assist channel creators in growing their channels by including commonly asked questions in the videos. The study was conducted on two YouTube channels, one of which featured just vegetarian Indian food and the other of which included both vegetarian and non-vegetarian Indian food. The major goal was to identify a suitable classifier that can aid in determining the emotions of YouTube comments.

Table 1. YouTube comments sentiment analysis Taxonomy

References	Language of Study	Source / Name of Dataset	Dataset Volume	Domain	Technique Applied	Algorithm	Result
Oh [26]	English	YouTube Spam Collection	1983	Music	Machine Learning	BOW/TF-IDF/CART/ LR/NB-B/RF/SVM-L/SVM-R/ ANN/ESM-H/ ESM-S	ESM-S (Accuracy: 0.864)
Akhter et al. [32]	Urdu	Roman Urdu Dataset	147000	Multiple YouTube Videos	Machine Learning	NB/BayesNet/k-NN/Hoeffding Tree/J48/REPTree /SVM/ RF/RT/ Logistic/SimpleLogistic/ LogitBoost/ OneR/JRip	LogitBoost (F-Measure: 0.992)
		Urdu Abusive Dataset	2171				SimpleLogistic (F-measure: 0.958)
Choi and Ko [45]	English	Fake Video Corpus	2505	Multiple YouTube Videos	Machine Learning	BTM/LDA	BTM (F-1 Score: 0.8812)
		Volunteer Annotated Video Dataset	448				
		Misleading YouTube Video Cor-pus	1805				
Livas et al [46]	English	NA	663	Invisalign patient testimonials	Lexicon	SentiStrength	Mean ICS of 3.78 (SD: 0.97)
Osztian et al. [47]	English	NA	1922	Algorithm Visualisation Videos	Lexicon	Mozdeh	-
Schneebeli [48]	English	-	20287	Humorous Videos	Lexicon	-	-
Saifullah et al [40]	Indonesian	Government Free Electricity Program during Covid-19	4862	Multiple YouTube Videos	Machine Learning	CV/TF-IDF/k-NN/Bernoulli/Decision Tree/ Support Vector Classifier/Random Forest/XG-Boost	Random Forest (Accuracy: 0.85)
Luengo et al. [37]	Spanish, Italian and English	Comments on COVID or Coronavirus	111808	Multiple YouTube Videos	Machine Learning	LDA	
Gao et al. [38]	Cantonese	Suicide-Related Comments	5051	Multiple YouTube Videos	Machine Learning	Jieba/word2vec/SVM/ AdaBoost/RF/LSTM/	LSTM (g-mean: 0.843)
Alakrot et al. [49]	Arabic	Offensive Language in Arabic	167549	Multiple YouTube Videos	Lexicon	Corpus-Based Approach	-
Uddin et al. [39]	English	-	700000	Multiple YouTube Videos	Machine Learning	LSTM	LSTM (Accuracy: 0.928)
Chaithra [9]	English	-	6248	Mobile	Hybrid	Naïve Bayes	Naïve Bayes (Accuracy: 0.797)
Chauhan et al.[50]	English	-	3263	Multiple YouTube Videos	Hybrid	SentiWordNet/Naïve Bayes	Naïve Bayes (F-measure: 0.74)
Aiyar et al [36]	English	-	13000	Music Videos	Machine Learning	BOW/TF-IDF/WORD2VEC/ M-NB/RF/SVM	SVM (F-1 Score: 0.9774)
Mai et al. [25]	English	YCSA Corpus	2153	Smartphone Product	Machine Learning	BERT/ELMO/WORD2VEC	BERT (Accuracy: 0.9145)
Savigny et al [42]	English	-	8115	Multiple YouTube Videos	Machine Learning	TF-IDF/Word2Vec/ Doc2Vec/GloVe/CNN/SVM	CNN (Accuracy: 0.762)
Ahmad et al. [7]	English	Movie Trailer Reviews/Comments	318783	Movies	Hybrid	SVM/MLR/MLP NN/RF	MLR (RAE: 0.2965)
Nguyen et al. [24]	English	SenTube	38000	Automobiles & Tablets	Machine Learning	STRUCT/BiLSTM/CoNBiLSTM	CoNBiLSTM (Accuracy: 0.5381(AUTO), 0.6128(TABLET))
	Italian		10000				

Subramanian et al [29]	Tamil	-	6534	Movies	Machine Learning	TF-IDF/BNB/LR/SVM/KNN/M-BERT/ MuRIL(Base/Large)/XLM-RoBERTa(Base/Large)	XLM-RoBERTa (Large) (Accuracy: 0.8853)
	Tamil and Malayalam	HASOC-DravidianCodeMix	11892				
Chakravarthi [30]	English	HopeEDI	28451	Multiple YouTube Videos	Machine Learning	TF-IDF/SVM/MNB/KNN/DT/LR	DT (F-1 Score: 0.9 for English)
	Tamil		20198				
	Malayalam		10705				
Chakravarthi et al [34]	English	YouTube Multilingual Comments	4946	Homophobic and Transphobic	Machine Learning	TF-IDF/countvec/fastText/LR/NB/RF/ SVM/DT/BiLSTM/BERT/MBERT	RF (Accuracy: 0.944)
	Tamil		4161				
	Tamil-English		6034				
Chakravarthi et al. [51]	Tamil	DravidianCodeMix	43919	Movie	Machine Learning	SVC/MNB/DT/RF/LGBM/ EWDT/EWODT/CNN/MBERT	CNN (F-1 Score: 0.85)
	Malayalam		20010				
	Kannada		7771				
Naz et al. [52]	English	-	2000	Educational	Hybrid	SentiStreangth/TextBlob/NB	
Obadimu et al [35]	English	-	849689	Covid-19	Machine Learning	LDA	-
Benkhelifa Randa and Laallam [31]	English	-	20000	Cooking	Machine Learning	TF-IDF/SVM	SVM (Accuracy: 95.3)
Röchert et al. [27]	English	-	2585367	Covid-19	Machine Learning	BERT/LSTM/CNN/LR/SVM	BERT (F-1 Score: 0.81)
Swain et al. [53]	English	-	17000	Multiple YouTube Videos	Hybrid	TextBlob/LR/RF/SVM/NB	LR (Accuracy: 0.8915)
Ezpeleta et al. [54]	English	YouTube Spam Campaigns	6431471	Multiple YouTube Videos	Machine Learning	CNB/NBM/NBMU	CNB (Accuracy: 0.9438)
Kaur et al. [44]	English	NishaMadulika's andKabita's Kitchen dataset	9800	Cooking	Machine Learning	TF-IDF/CV/TFV/CART/LR/NB-B/NB-G/NB-M/RF/SVM-L/SVM-P/SVM-R	LR (Accuracy: 0.7537)

The experimental outcome applied to Nisha Madhulika and Kabita datasets, respectively, LR with term frequency vectorizer yielded 74.01% and 75.37% of the results.

4.2 LEXICON BASED MODELS

This method is based on the text's polarity score based on its positive and negative aspects, which were brought up by dictionaries of words. Fig.9 shows that 14% of papers used lexicon-based strategies. Lexicon-based sentiment analysis method includes both dictionary and corpus-based approach. WordNet and SentiWordNet are preset dictionaries used in dictionary-based sentiment categorization [55], [56], [57]. However, sentiment analysis based on corpus relies on a statistical study of the texts instead of predetermined dictionaries to do the research [58].

The study of anti-social online communication practices, such as abusive language and cyberbullying, has gained attention recently. However, most of these studies gather data from English sources. Alakrot et al. [59] raise concerns about a lack of datasets targeting Arabic text. The authors filled this gap in the literature by providing a dataset of Arabic-language YouTube comments created expressly for machine learning-based abusive language identification. The dataset contains a wide range of offensive language and flaming in the form of YouTube comments. The paper also discusses the methodology used to label the dataset, considering regional variations in Arab dialects and attitudes towards offensive language. Statistical analysis of the dataset is also offered to make the dataset suitable for use as a training dataset for predictive modelling.

An investigation by Osztián et al. [47] focuses on assessing comments on the AlgoRythmics channel using the Comment Term Frequency Comparison (CTFC) social media analytics approach to find and explore potential future research directions. Using the CTFC social media analytics tool, the authors examined the comments from the AlgoRythmics YouTube channel in this study. Understanding how the videos on the AlgoRythmics channel are being perceived by the audience and identifying areas for development may be done with the help of YouTube comment analytics. Based on 10 videos of algorithm visualization from the AlgoRythmics channel, the authors' experiment emphasizes the value of user input. They looked at the key elements of user feedback using the CTFC social media analytics approach to discover proof of how the channel operates and fresh suggestions for improvement. YouTube Studio Analytics and the Mozdeh Big Data Analysis tool analyzed the comments.

Since information technology has allowed students to obtain the best educational content anywhere globally, the emergence of YouTube videos has coincided with a substantial revolution in

education. In another research, Naz et al. [52] used sentiment analysis to investigate the opinions expressed in YouTube comments for educational videos. The sentiments were analyzed using SentiStrength, TextBlob, and Naive Bayes Classifier. The results showed that the majority of the comments had positive emotions. The experiment result suggests that students are satisfied with YouTube educational videos and consider YouTube to be a beneficial learning tool.

On YouTube, there is a lot of activity related to learning more about treating numerous diseases. Another study by Livas et al. [46] suggests that Invisalign patient testimonials on YouTube can be valuable information for potential patients. The authors collected data from YouTube videos about Invisalign experiences and then analyzed the data to understand the different perspectives on Invisalign treatment. Videos were given an information completeness score (ICS), and comments were categorized according to their source and topic. The average ICS for the 40 evaluated testimonies was 3.78 (SD 0.97). They discovered that a video's length, subject matter, and age did not impact its audience or interaction. However, comments on sponsored videos or those from well-known YouTube creators were more favourable, showing viewers believe in and are affected by sponsorships or videos from reliable sources.

Although "lol" is frequently abbreviated for "laughing out loud" in social media, it can also mean other things. In another research, Schneebeli [48] investigates how the position of "lol" in YouTube comments affects its function. The study's data was gathered from the discussion threads of three well-known Miranda Sings-posted YouTube videos. A total of 20,287 comment texts and 886 unique non-lexicalized "lol" occurrences were extracted and stored. According to the study, "lol" is more frequently employed in clause-final contexts for social interaction and clause-initial contexts for discourse organization.

4.3 HYBRID APPROACHES

Among the 29 papers evaluated, around 17% employed hybrid strategies that combined lexicon- and machine-learning-based approaches.

Unboxing videos, especially ones featuring mobile phones, are currently the most popular trend on video platforms. Viewers' perceptions of unboxed phones may be obtained from analyzing the comments on these videos. It is possible to assess if the feedback is good or negative by looking at the view conveyed in these comments. Chaithra [9] utilizes a hybrid technique that combines the lexical approach Sentiment VADER with the machine learning algorithm Naive Bayes (NB) to determine the sentiment of these comments. The findings reveal that VADER has a good influence on the NB classifier, obtaining an accuracy of 79.78% and an F1 score of 83.72%.

Since abusive video risks public safety, effective detection algorithms are urgently needed. Here, sentiment analysis-based video categorization is suggested by Swain et al. [53] to increase detection accuracy. Contents are classified into two categories, abusive and nonabusive, using a sentiment analysis-based video classification system. The proposed sentiment analysis algorithm analyses YouTube comments for a given video as a source of input, and the model then decides the category to which that particular video belongs. Numerous methods are employed, including lemmatization, logistic regression, bag of words, and NLP. The suggested approach yields competitive outcomes for the identification of offensive content. The experimental results demonstrate the simplicity and effectiveness of the methodology applied.

In another research, the influence of several factors (extracted from user comments) on the retrieval of YouTube videos is demonstrated by Chauhan et al. [50]. The authors have noticed that while the popularity of the video is a key factor, it's also crucial to retrieve videos based on the users' interests, skill levels, and understanding. Collect comments on YouTube videos from five different search categories. They extracted YouTube comments and pre-process to remove spam and other irrelevant comments, enriched each topic using SentiWordNet., extracted aspect terms from the comments using part-of-speech tagging and Stanford NLP Parser, determined the aspect categories, classified the sentiment of each aspect term and category using supervised machine learning algorithms and finally rank videos based on the polarity of different aspects. The experimental results suggest that aspect-based sentiment analysis can be used to improve the accuracy of video retrieval.

5. DATASETS LANGUAGE AND EXTENT

This section covers the languages used in the datasets of reviews, comments, and postings related to the publications under consideration in this study. The datasets include a variety of languages. Sentiment analysis has been used in English, Urdu, Indonesian, Mixed, Cantonese, and Arabic, as illustrated in Fig.10. This review study examined 29 studies, 19 of which used datasets for sentiment analysis in English, followed by mixed with 6 articles. The remaining languages, Arabic, Cantonese, Indonesian and Urdu, have one paper each.

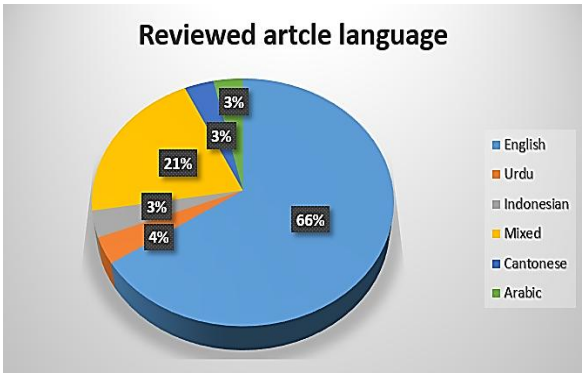

Fig.10. Reviewed research publication language

Significant progress has been made in the field of sentiment analysis in English, not just adapting cutting-edge theories in the domains of lexicon approaches [26], [46], [47] and machine learning strategies [25], [26], [27], [44] but, at the application level, taking into account several domains: education, health, tourism and product. This systematic review also aims to cover the challenges related to the language used. Most authors mentioned that they preferred English due to the availability of datasets in English. Several authors have tried to prepare a corpus on other languages and found great accuracy in their experiments. It indicates that sentiment analysis should not depend on a specific language. There is still a need for more investigation into creating non-English lexicons and using efficient ML classifiers.

The utilized datasets were divided into two parts: primary and secondary datasets. Out of the 29 articles, 16 used the primary datasets manually gathered from various YouTube channels. On the other hand, 14 out of 29 papers also examined the secondary datasets utilized by other authors in different studies. An investigation by Ezpeleta et al. [54] gathered comments from the YouTube spam campaign as a primary dataset, and Oh et al. [26] used YouTube spam collection as a secondary dataset.

Another investigation by Ahmad et al. [60] gathered movie trailer comments/reviews as the primary dataset to predict movie revenue. On the other hand, Subramanian et al. [29] used the secondary dataset HASOC-DravidianCodeMix collected from various YouTube movie channels. The databases' domains include tourism, health, films, products, businesses, education, and psychology.

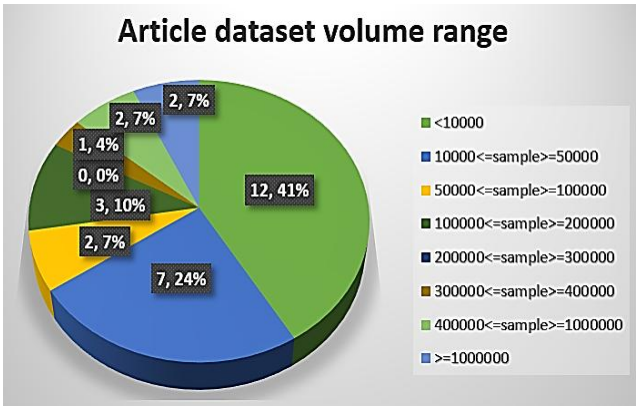

Fig.11. Reviewed article dataset volume range

Additionally, as shown in Fig.11, only 7% of articles utilized datasets with more than 1 million samples, while 41% used datasets with less than 1000 samples. There is still a need for more investigation into creating non-English lexicons and using efficient ML classifiers. More research is still required on extraction techniques and the usefulness of different YouTube comments datasets in English and non-English.

6. DISCUSSION

This section attempts to emphasize and go through three important aspects of this research. They can be divided into two categories: (1) challenges dealing with the datasets, (2) challenges associated with the techniques and (3) challenges regarding the assessment metrics. The data's overall quality is crucial in conducting an efficient analysis.

6.1 CHALLENGES DEALING WITH THE DATASETS

A huge volume of data from numerous YouTube channels offers fresh perspectives for people, companies, and governments. The issues associated with the use of data samples may be further subdivided into two classes based on data gathering and data preparation or pre-processing stages. One of the initial difficulties in the data-collecting process is the data size. It is essential to provide a sufficient training data set for sentiment analysis to increase the classifier accuracy [7]. A significant dataset can help prevent overfitting by providing the model with more information to learn from [29]. Due to the limited availability of relevant datasets and the need for enough data samples, researchers in certain works on sentiment analysis of languages other than English texts frequently choose to develop their training datasets employing supervised learning and human annotation [38], [59]. Despite huge data benefits, academics cannot keep up with the amount and speed of social networking data [50]. Finding and utilizing a suitable technique for collecting data is essential.

However, for researchers utilizing primary data, the variation in data [34] and varying lengths of comment [59] are some of the difficulties frequently faced beyond the problems caused by large-scale data. Additionally, the author could create bias while manual comment labelling is done after data collection to apply supervised approaches. It is especially difficult to annotate comments that contain figurative language. Figurative language relies highly on setting, topic and context, making it challenging for annotation experts to identify the true opinion expressed in text. As a result, various annotators may choose to annotate the same statement differently [25], [30].

Data diversity can make pre-processing a challenging task after data collection. Social media data is typically unstructured, noisy, and irrelevant. It cannot be easy to extract meaningful insights from this data. Social networking platform data is unstructured [29], noisy [60], ineffective [32], unnecessary [52], and typically lacks textual emotional substance. Swain et al. [53] presented that pre-processing is crucial in detecting sentiments when dealing with a variable stream of textual information. As a result, to complete the necessary tasks, various techniques for pre-processing the data are required to collect relevant contents from the enormous volume of data and transform the text to a consistent and identifiable form [38], [42].

Certain studies additionally omit nonalphabetic keywords during data pre-processing. Numerous YouTube comments also contain emoticons as nonalphabetic phrases and are closely linked to the users' emotions. The accuracy of the classification techniques may be impacted if they are eliminated from the dataset. Benkhelifa Randa and Laallam [31] used an emoticons corpus to address this problem, and Saifullah et al. [40] replaced emoticons and smilies with suitable phrases. Future research can focus on creating a complete emoticon vocabulary that integrates the most popular emoticons from social networking sites and powerful ML classifiers, considering that emoticons also convey people's views and thoughts about certain subjects.

6.2 CHALLENGES RELATED TO TECHNIQUES APPLIED

Two methods are typically used to analyze sentiment: lexicon-based and machine learning. One significant problem is the method's adaptability. The existence of previously specified terms in dictionaries is a prerequisite for the lexicon-based approach. Since rules, usage patterns, and other linguistic constructions are included in dictionaries, this impacts their correctness [7], [52]. Additionally, a word's emotional direction in one dictionary could be completely different from that in another. Many researchers use supervised learning techniques in their machine learning-based approaches to train the sentiment classifications with labelled data samples from a particular field. Fig.12 depicts the percentage of articles utilizing various feature

extraction methods. The subject of the data samples substantially affects the practical area of sentiment predictors.

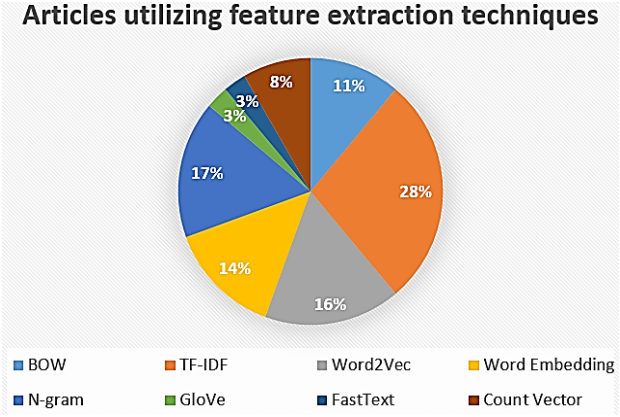

Fig.12. Article utilizing the techniques for feature extraction

Despite having excellent applications in the research sector, many methodologies fall short in other fields. The SenTube lexicon used by Nguyen et al. [24] is more focused on products, and the lexicon-based strategy can outperform the CNN and LSTM classification methods for a dataset for product review, less on the dataset for hotel reviews. In addition, while the hybrid CNN algorithm may score higher on the dataset for product review, the hybrid models beat other classification algorithms on both datasets. However, the hybrid LSTM can outperform the other models on the dataset for hotel review. Subramanian et al. [22] built a Dravidian Code-Mix to address the issue with domain-specific sentiment analysis. They evaluated it against Machine Learning methods on datasets from various disciplines, and the greatest accuracy, 88.5%, was discovered for XLM-RoBERTa (Large) against other ML models.

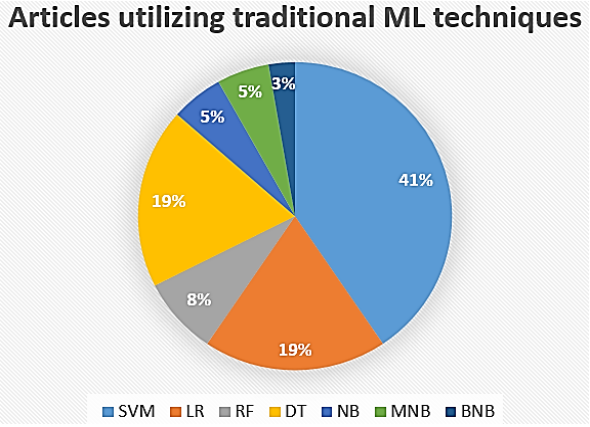

Fig.13. Article utilizing the traditional machine learning techniques

Therefore, developing a generic sentiment analysis approach is essential to make sentiment analysis more useful. Additionally, manually annotating a large dataset or lexicon with various subjects sometimes takes a lot of resources and time [59]. Fig.13 depicts the percentage of articles utilizing machine-learning techniques. The focus of future research needs to shift from single-domain applications to multi-domain applications progressively. The Table.2 shows that only 12 researchers have been used deep learning techniques. Various other deep learning methods can also be considered as future work.

Aside from the diversity of approaches, another issue for using sentiment analyzers in practical applications is the effectiveness of the methods. A possible reason is that user data has grown dramatically because of the YouTube channel's increasing attractiveness and complexity, necessitating more labour, time, and effort to handle the rising amount and diversity of data kinds [9]. Another factor is that corporations and governments rely considerably on timely public opinion monitoring. As a result, the time required for the execution of sentiment evaluation is a critical aspect of their practical application in reality. Although many studies have focused on the accuracy of their methods, few have considered the trade-off between accuracy and running time.

Table.2. Number of articles utilizing deep learning technique

DL technique	Total number of articles
CNN	4
LSTM	6
MLP	2

6.1 CHALLENGES RELATED TO THE EVALUATION MATRIX

After examination of the papers that are a part of this review study, it can be seen that the sentiment classifiers have been evaluated using a range of metrics. Some examples of evaluation matrices are accuracy, precision, recall, F1-score, RMSE, LRAP score, MAE, etc. Accuracy, precision, recall and f-measure are the most often used assessment metrics for these publications and reports. These metrics may be calculated using the confusion matrix, which displays and summarises a classification algorithm's performance. However, the two measures that draw the greatest attention are accuracy and F-measure. To address the problem of the trade-off between accuracy and recall, the term "F-measure" is used. In addition to precision and F-measure, Oh [26] employed the Receiver Operating Characteristic (ROC) curve to assess their suggested model's performance. Ahmad et al. [7] utilized the root mean square error (RMSE) and Mean Absolute Error (MAE) to compare model outputs with true data. Furthermore, a review of model performance assessment metrics from reviewed articles shows that accuracy and F-score are frequently provided concurrently.

It is crucial to analyze their practical utility and assess the efficacy of sentiment classifiers using the various statistical metrics mentioned above to understand the application of such classifiers in real situations. Sentiment analysis is part of machine learning, a tool for resolving business, government, and private issues. Processing vast volumes of data is challenging in the information age, and new ways have been developed to help. As a result, these approaches must be effective rather than making the situation worse [61].

Although the different sentiment classifiers used in research exhibit excellent F1-score, accuracy, or similar statistical assessment metrics, few were assessed from a practical standpoint, such as execution time. An experiment by Akhter et al. [25] showed that the regression-based models perform better than others but need more time to create the model.

As mentioned in Section B, the effectiveness of a sentiment analysis model is essential for solving real-world issues, especially in situations when there is a lot of data flow and decisions must be made immediately. Therefore, future studies should assess sentiment classifiers more thoroughly, considering their usefulness in the real world, including their efficiency, extensibility, and scalability.

7. CONCLUSION

This research aims to systematically review and analyze the existing research on analyzing sentiment in YouTube user comments. The paper describes the difficulties in current sentiment analysis research in addition to having a complete understanding of how sentiment analysis is used on user-generated data. The study discusses the goals of the sentiment evaluation process, the overall procedure to implement, the algorithms used, and how they have been employed in various domains using the PRISMA framework. The research then highlights several obstacles and issues relating to datasets, reviews text languages, evaluation methodologies, and assessment metrics in current literature by comparing the characteristics of various studies.

During the investigation and evaluation process, it was discovered that there were certain study gaps in the previous work:

- Utilizing a single lexicon or dataset with a range of topic phrases and emoticons for sentiment analysis tasks on both English and non-English texts is advised.
- Future studies should progressively shift from applications in one domain to multi-domain implementations by creating generic sentiment classifiers to make sentiment analysis more useful.
- When assessing the effectiveness of their models, researchers should pay close attention to how each assessment statistic is used. There are many distinct assessment metrics, but it is not required to report all of them at once because they could have various use cases.

Additionally, this study has some limitations. Our objective is to review the most recent sentiment analysis studies on YouTube comments, and the intended publishing window was from 2017 to 2023. Literature does not often examine sentiment analysis methods for mixed languages, such as (Hindi+English) or Hinglish. In the future, the timeline can be extended to catch additional research variety and comprehend more sentiment analysis methodologies. Non-English-speaking researchers and multi-lingual academics will be welcomed to collaborate better to grasp the present status of worldwide sentiment analysis research.

REFERENCES

[1] A.F. Ibrahim, M. Hassaballah, A.A. Ali, Y. Nam and I.A. Ibrahim, "COVID19 Outbreak: A Hierarchical Framework for User Sentiment Analysis", *Computers, Materials and Continua*, Vol. 70, No. 2, pp. 2507-2524, 2022.

[2] Y.L. Chen, C.L. Chang and C.S. Yeh, "Emotion Classification of YouTube Videos", *Decision Support System*, Vol. 101, pp. 40-50, 2017.

[3] A. Severyn, A. Moschitti, O. Uryupina, B. Plank and K. Filippova, "Multi-Lingual Opinion Mining on YouTube", *Information Processing Management*, Vol. 52, No. 1, pp. 46-60, 2016.

[4] H.T. Nguyen and M. Nguyen, "Multilingual Opinion Mining on YouTube - A Convolutional N-Gram BiLSTM Word Embedding", *Information Processing Management*, Vol. 54, pp. 451-462, 2018.

[5] A. Shoufan and F. Mohamed, "YouTube and Education: A Scoping Review", *IEEE Access*, Vol. 10, pp. 12557-125599, 2022.

[6] S. Naz, S. Hina, U. Fatima and H. Tabassum, "A Hybrid Approach to Measure Students' Satisfaction on YouTube Educational Videos", *International Journal of Emerging Technologies in Learning*, Vol. 18, No. 9, pp. 131-147, 2023.

[7] I.S. Ahmad, A.A. Bakar and M.R. Yaakub, "Movie Revenue Prediction Based on Purchase Intention Mining using YouTube Trailer Reviews", *Information Processing Management*, Vol. 57, No. 5, pp. 1-15, 2020.

[8] Y. Rashid and M. Zeeshan, "Customer Attitude towards Online Ads of Smartphone Brands: A Netnographic Analysis of User Generated Comments on YouTube", *Journal of Management Sciences*, Vol. 5, pp. 40-64, 2018.

[9] V.D. Chaithra, "Hybrid Approach: Naive Bayes and Sentiment VADER for Analyzing Sentiment of Mobile Unboxing Video Comments", *International Journal of Electrical and Computer Engineering*, Vol. 9, No. 5, pp. 4452-4459, 2019.

[10] Z. Acharoui, A. Alaoui, B. Ettaki, J. Zerouaoui and M. Dakkon, "Identifying Political Influencers on YouTube during the 2016 Moroccan General Election", *Procedia Computer Science*, Vol. 170, pp. 1102-1109, 2020.

[11] Z. Munn, M.D.J. Peters, C. Stern, C. Tufanaru, A. McArthur and E. Aromataris, "Systematic Review or Scoping Review? Guidance for Authors when Choosing between A Systematic or Scoping Review Approach", *BMC Medical Research Methodology*, Vol. 18, No. 1, pp. 1-17, 2018.

[12] A. Kumar and G. Garg, "Systematic Literature Review on Context-Based Sentiment Analysis in Social Multimedia", *Multimed Tools and Applications*, Vol. 79, No. 21-22, pp. 15349-15380, 2020.

[13] A.H. Alamoodi, "Sentiment Analysis and Its Applications in Fighting COVID-19 and Infectious Diseases: A Systematic Review", *Expert System with Applications*, Vol. 167, pp. 1-27, 2021.

[14] S. Windisch, S. Wiedlitzka and A. Olaghere, "PROTOCOL: Online Interventions for Reducing Hate Speech and Cyberhate: A Systematic Review", *Campbell Systematic Reviews*, Vol. 17, pp. 1-17, 2020.

[15] Q.A. Xu, V. Chang and C. Jayne, "A Systematic Review of Social Media-Based Sentiment Analysis: Emerging Trends and Challenges", *Decision Analytics Journal*, Vol. 3, pp. 100073-100098, 2022.

[16] I.S. Ahmad, A.A. Bakar, M.R. Yaakub and S.H. Muhammad, "A Survey on Machine Learning Techniques in Movie Revenue Prediction", *SN Computer Science*, Vol. 1, pp. 1-37, 2020.

[17] E. Caglayan Akay, N.T. Yılmaz Soydan and B. Kocarik Gacar, "Bibliometric Analysis of the Published Literature on

Machine Learning in Economics and Econometrics", *Social Network Analysis and Mining*, Vol. 12, No. 1, pp. 1-18, 2022.

[18] M. Aria and C. Cuccurullo, "Bibliometrix: An R-Tool for Comprehensive Science Mapping Analysis", *Journal of Informetrics*, Vol. 11, No. 4, pp. 959-975, 2017.

[19] RStudio Team, "RStudio: Integrated Development Environment for R", Available at: http://www.rstudio.com/, Accessed in 2020.

[20] A. Massimo and C. Corrado, "Biblioshiny Bibliometrix for No Coders", Available at: https://www.bibliometrix.org/biblioshiny/, Accessed in 2025.

[21] E. B. Kemp, "Mapping Systematic Reviews of Breast Cancer Survivorship Interventions: A Network Analysis", *Journal of Clinical Oncology*, Vol. 40, No. 19, pp. 2083-2093, 2022.

[22] N.J. Van Eck and L. Waltman, "Software Survey: VOSviewer, A Computer Program for Bibliometric Mapping", *Scientometrics*, Vol. 84, No. 2, pp. 523-538, 2010.

[23] N. Donthu, S. Kumar and D. Pattnaik, "Forty-Five Years of Journal of Business Research: A Bibliometric Analysis", *Journal of Business Research*, Vol. 109, pp. 1-14, 2020.

[24] H.T. Nguyen and M. Le Nguyen, "Multilingual Opinion Mining on YouTube - A Convolutional N-Gram BiLSTM Word Embedding", *Information Processing Management*, Vol. 54, No. 3, pp. 451-462, 2018.

[25] L. Mai and B. Le, "Joint Sentence and Aspect-Level Sentiment Analysis of Product Comments", *Annals of Operations Research*, Vol. 300, No. 2, pp. 493-513, 2021.

[26] H. Oh, "A YouTube Spam Comments Detection Scheme using Cascaded Ensemble Machine Learning Model", *IEEE Access*, Vol. 9, pp. 144121-144128, 2021.

[27] D. Rochert, G.K. Shahi, G. Neubaum, B. Ross and S. Stieglitz, "The Networked Context of COVID-19 Misinformation: Informational Homogeneity on YouTube at the Beginning of the Pandemic", *Online Social Networks and Media*, Vol. 26, pp. 1-23, 2021.

[28] C. Vasantharajan and U. Thayasivam, "Towards Offensive Language Identification for Tamil Code-Mixed YouTube Comments and Posts", *SN Computer Science*, Vol. 3, No. 1, pp. 94-108, 2021.

[29] M. Subramanian, "Offensive Language Detection in Tamil YouTube Comments by Adapters and Cross-Domain Knowledge Transfer", *Computer Speech and Language*, Vol. 76, No. 3, pp. 1-13, 2022

[30] B.R. Chakravarthi, "HopeEDI: A Multilingual Hope Speech Detection Dataset for Equality, Diversity, and Inclusion", *Proceedings of 3rd Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pp. 41-53, 2020.

[31] F.Z. Benkhelifa Randa and Laallam, "Opinion Extraction and Classification of Real-Time YouTube Cooking Recipes Comments", *Proceedings of International Conference on Advanced Machine Learning Technologies and Applications*, pp. 395-404, 2018.

[32] M.P. Akhter, Z. Jiangbin, I.R. Naqvi, M. Abdelmajeed and M.T. Sadiq, "Automatic Detection of Offensive Language for Urdu and Roman Urdu", *IEEE Access*, Vol. 8, pp. 91213-91226, 2020.

[33] B.R. Chakravarthi, M.B. Jagadeeshan, V. Palanikumar and R. Priyadharshini, "Offensive Language Identification in Dravidian Languages using MPNet and CNN", *International Journal of Information Management Data Insights*, Vol. 3, No. 1, pp. 1-26, 2023.

[34] B.R. Chakravarthi, "Dataset for Identification of Homophobia and Transophobia in Multilingual YouTube Comments", *Proceedings of International Conference on Computation and Language*, pp. 1-44, 2021.

[35] A. Obadimu, T. Khaund, E. Mead, T. Marcoux and N. Agarwal, "Developing A Socio-Computational Approach to Examine Toxicity Propagation and Regulation in COVID-19 Discourse on YouTube", *Information Processing Management*, Vol. 58, No. 5, pp. 102660-102678, 2021.

[36] S. Aiyar and N. Shetty, "N-Gram Assisted Youtube Spam Comment Detection", *Procedia Computer Science*, Vol. 132, pp. 174-182, 2018.

[37] O. Luengo and E. De-Blasio, "COVID-19 on YouTube: Debates and Polarisation in the Digital Sphere", *Comunicar*, Vol. 29, No. 69, pp. 9-19, 2021.

[38] J. Gao, Q. Cheng and P.L.H. Yu, "Detecting Comments Showing Risk for Suicide in YouTube", *Proceedings of International Conference on Advances in Intelligent Systems and Computing*, pp. 385-400, 2024.

[39] M.R. Uddin, N.S. Prova and A. Jawad, "Qualitative Sentiment Analysis of YouTube Contents based on User Reviews", *Proceedings of International Conference on Applied Artificial Intelligence and Computing*, pp. 1099-1102, 2023.

[40] S. Saifullah, Y. Fauziyah and A.S. Aribowo, "Comparison of Machine Learning for Sentiment Analysis in Detecting Anxiety based on Social Media Data", *Jurnal Informatika*, Vol. 15, No. 1, pp. 1-45, 2021.

[41] E. Aimeur, S. Amri and G. Brassard, "Fake News, Disinformation and Misinformation in Social Media: A Review", *Social Network Analysis and Mining*, Vol. 13, No. 1, pp. 1-30, 2023.

[42] J. Savigny and A. Purwarianti, "Emotion Classification on Youtube Comments using Word Embedding", *Proceedings of International Conference on Applied Artificial Intelligence*, pp. 1-5, 2017.

[43] S. Bhattacharya, A. Agarwala and S. Roy, "Mood Detection and Prediction using Conventional Machine Learning Techniques on COVID19 Data", *Social Network Analysis and Mining*, Vol. 12, No. 1, pp. 139-154, 2022.

[44] G. Kaur, A. Kaushik and S. Sharma, "Cooking is Creating Emotion: A Study on Hinglish Sentiments of Youtube Cookery Channels using Semi-Supervised Approach", *Big Data and Cognitive Computing*, Vol. 3, No. 3, pp. 1-19, 2019.

[45] H. Choi and Y. Ko, "Using Adversarial Learning and Biterm Topic Model for an Effective Fake News Video Detection System on Heterogeneous Topics and Short Texts", *IEEE Access*, Vol. 9, pp. 164846-164853, 2021.

[46] C. Livas, K. Delli and N. Pandis, "'My Invisalign Experience': Content, Metrics and Comment Sentiment Analysis of the most Popular Patient Testimonials on

YouTube", *Progress in Orthodontics*, Vol. 19, No. 1, pp. 1-26, 2018.

[47] P.R. Osztian, Z. Katai and E. Osztian, "Investigating the AlgoRythmics YouTube Channel: the Comment Term Frequency Comparison Social Media Analytics Method", *Acta Universitatis Sapientiae, Informatica*, Vol. 14, No. 2, pp. 273-301, 2022.

[48] C. Schneebeli, "Where lol Is: Function and Position of lol Used as a Discourse Marker in YouTube Comments", *Discours*, Vol. 27, pp. 1-29, 2020.

[49] A. Alakrot, L. Murray and N. Nikolov, "Dataset Construction for the Detection of Anti-Social Behaviour in Online Communication in Arabic", *Procedia Computer Science*, Vol. 142, pp. 174-181, 2018.

[50] G.S. Chauhan and Y.K. Meena, "YouTube Video Ranking by Aspect-Based Sentiment Analysis on User Feedback", *Proceedings of International Conference on Advances in Intelligent Systems and Computing*, pp. 1-6, 2019.

[51] B.R. Chakravarthi, M.B. Jagadeeshan, V. Palanikumar and R. Priyadharshini, "Offensive Language Identification in Dravidian Languages using MPNet and CNN", *International Journal of Information Management Data Insights*, Vol. 3, No. 1, pp. 100151-100167, 2023.

[52] S. Naz, S. Hina, U. Fatima and H. Tabassum, "A Hybrid Approach to Measure Students' Satisfaction on YouTube Educational Videos", *International Journal of Emerging Technologies in Learning*, Vol. 18, No. 9, pp. 131-147, 2023.

[53] D. Swain, M. Verma, S. Phadke, S. Mantri and A. Kulkarni, "Video Categorization Based on Sentiment Analysis of YouTube Comments", *Proceedings of International Conference on Advances in Intelligent Systems and Computing, Springer Science and Business Media*, pp. 59-67m 2021.

[54] E. Ezpeleta, M. Iturbe, I. Garitano, I. V. De Mendizabal and U. Zurutuza, "A Mood Analysis on Youtube Comments and A Method for Improved Social Spam Detection", *Lecture Notes in Computer Science*, pp. 514-525, 2018.

[55] G. Jain, M. Sharma and B. Agarwal, "Optimizing Semantic LSTM for Spam Detection", *International Journal of Information Technology*, Vol. 11, No. 2, pp. 239-250, 2019.

[56] E.R. Mahalleh and F.S. Gharehchopogh, "An Automatic Text Summarization based on Valuable Sentences Selection", *International Journal of Information Technology*, Vol. 14, No. 6, pp. 2963-2969, 2022.

[57] N.V. Babu and E.G.M. Kanaga, "Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review", *SN Computer Science*, Vol. 3, No. 1, pp. 1-14, 2022.

[58] N.C. Dang, M.N. Moreno Garcia and F. De La Prieta, "Sentiment Analysis based on Deep Learning: A Comparative Study", *Electronics*, Vol. 9, No. 3, pp. 1-24, 2020.

[59] A. Alakrot, L. Murray and N.S. Nikolov, "Dataset Construction for the Detection of Anti-Social Behaviour in Online Communication in Arabic", *Procedia Computer Science*, Vol. 142, pp. 174-181, 2018.

[60] A. Bakar, I. Said Ahmad and M.R. Yaakub, "Movie Revenue Prediction Based on Purchase Intention Mining Using YouTube Trailer Reviews", *Information Processing Management*, Vol. 57, pp. 1-20, 2020.

[61] S. Saifullah, Y. Fauziyah and A.S. Aribowo, "Comparison of Machine Learning for Sentiment Analysis in Detecting Anxiety based on Social Media Data", *Informatika*, Vol. 15, No. 1, pp. 1-45, 2021.