International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Enhanced Classification Model for Detecting Deceptive Content in Social Media using Natural Language Processing Techniques

Mohamed Shenify

Submitted: 28/11/2023 **Revised**: 30/12/2023 **Accepted**: 09/01/2024

Abstract: People may now express their thoughts on products, services, motion pictures and other media thanks to the rise of social networking sites. The emotion of the user is their viewpoint or viewpoint on any issue, event, occasion, or service. People's choices have always been influenced by their mental condition in general. Emotions have been extensively studied in natural language in recent years, but many issues need to be addressed. One of its most serious issues is a lack of exact categorization resources. Researchers discovered an unintentionally bias and unfairness generated by data sets used for training, which resulted in the inaccurate classification of harmful terms in context. Several ways to discover toxicity in text are evaluated and reported in this research, with the goal of improving the general standard of text categorization. Suggested methods included a deep learning model of Long- and Short-Term Memory (LSTM) with Glove word embedding and the LSTM with word embedding created by the representations of Bidirectional Encoder Representation from Transformers (BERT). The results showed that LSTM with BERT, as the word embedding attained a satisfactory precision of 94% and a F1 score of 0.89 in the binary categorization of comments (dangerous and nontoxic). The combined use of LSTM and BERT, as the outperformed both LSTM alone and LSTM with Multimodal word anchoring. This work attempts to overcome the challenge of accurately categorizing comments by relating models to bigger corpus of text (good-quality keyword anchoring) rather than training information alone.

Keywords: *Social Media, Deceptive Content, Long Short-Term Memory (LSTM), Bidirectional Encoder Representations from Transformers (BERT), Natural Language Processing.*

1. Introduction

The scientific technique of teaching a computer to comprehend and interpret human language is known as natural language processing. The advancements in machine learning and research have made Natural Language Processing (NLP) far more significant in recent years [1].

Researchers are working hard to extract fascinating data and statistics from the language of humans and use the findings in various spheres of society, including industry, retail centres, hospitals, and schools. Rule-based systems were once used to handle NLP difficulties. But because text is different all over, machine learning is used to Natural Language Processing (NLP) and has become increasingly popular using Support Vector Machines (SVM) and Naïve Bayes.

Natural language processing (NLP) or text mining is the process of extracting human-generated text from various social media networks using a variety of programs, algorithms, and methods. This is a significant field in AI. Information mining approaches have accomplished the best results across the fields of programmed question answering machines, anaphora goal, programmed deliberation, bioinformatics, and web connection network examination because of progressing research on text mining and regular language handling utilizing information mining calculations, AI, and profound learning. Data mining, text categorization, and Natural Language Processing (NLP) have been shown to be highly beneficial in all aspects of life.

In order to learn order dependency in sequence prediction problems, this research used a number of deep learning techniques, including LSTM networks [2], and redundant neural networks with nodes. Voice recognition and machine translation are two complex problem domains where deep learning technologies are essential. Temporal Convolutional Neural Networks (TCNNs) were first proposed to be segmented visually in the big articles. In these classic methods, time information is captured at high levels using (usually) RNN in the first two stages, where low-level calculating functions are introduced into a classifier and low-level functions are used to encode spatial time-related data using CNN. The main problem is that a comparable procedure requires two different models. A uniform technique for capturing the encoder-decoder (the two information layers) hierarchically is provided by TCN [3].

The rise in social media use and its user base has made it feasible for people to voice their opinions in everyday speech. Research in social media sentiment analysis has been very busy recently. For this natural language system of management to be analysed, a model must be able to

Associate Professor, Department of Computer Science,
Faculty of Computing & Information, Al-Baha University, Al-Baha, Kingdom of Saudi Arabia.

Corresponding Author : maalshenify@bu.edu.sa

distinguish between the many emotional components of social media users.

The thorough examination of sentimental forms of analysis demonstrates that the research may help the user categorize the operator's emotions according to a topic. To determine user sentiments or views, an analysis of their feelings is performed. Each person on social media, like Twitter, has their own area to share thoughts, discuss issues, or leave service-related comments. The user review demonstrates the variety of sentiment evaluation models that have been created for natural language evaluations, including assessments of products, movies, politics, and other media.

Many studies are being done on Twitter to forecast the general sentiment that is used in many domains and applications [4]. In general, feeling is classified into three categories:

(i) Technique for completing information,
(ii) A mixed teaching strategy, and
(iii) Sentiment analysis requires a thorough examination of methods for processing natural language in order to give datasets to train using machine learning and emotion lexical data for statistical or semantically approaches.

The purpose of this research is to create a user-generated sensation framework for tweets in English. Social media platforms like Twitter have had a significant impact on non-native English speakers. When expressing one's an opinion on social media, non-native English speakers face many speech obstacles. The first task is to create grammatical guidelines for categorizing emotions in tweets written in English [5]. The lack of resources, such feeling lexicons and datasets, is the second issue. Enhancing the performance of slang words—words translated in different languages and fields—is the last question.

Researchers have identified the presence of unfairness in Artificial Intelligence (AI) models as one of the most significant obstacles dealing with users of ML technology, given the increased reliance on these models for various purposes and tasks. This is because the majority of these models are developed using human-generated data that implies that human bias will manifest itself clearly in these models. Put differently, machine learning algorithms exhibit bias much like the people who created the training data [6].

The creators of machine learning models need to be proactive in identifying and eliminating these biases, or else the models might reinforce unjust categorization practices. The age distribution of the internet users, the implicit or overt prejudices of those labeling, or the choice and sampling processes might all contribute to an unintentional bias in the models. The aim of this endeavor is to provide a novel approach to sentiment analysis on Twitter, organized into two phases. The first and most common kind of jargon on Twitter is emoji's and other symbols. The techniques that are utilized to transform emoji's to plain text are not reliant on language being used. On the other hand, it can be easily translated into other languages. Secondly, the topic matter of the produced tweets is used to classify them [7]. One benefit of using BERT as a language model is that it was retrained using plain text instead of tweets. The models need fewer resources and hours to construct since they are easily available in several languages and depend on plain text. The ensuing benefits are available:

(1) Models could have been trained directly on tweeting from the beginning.
(2) Plain text is available. Corpora are larger than just tweets corpora, enabling for better performance.

Social media's versatility has led to its enormous rise in relevance in recent decades. It goes without saying that those who use social media platforms often will produce enormous amounts of data [8]. Expressions of hate are also on the rise as a result of the massive amount of data that social media users collect. When a motion picture is released, for instance, reviews and opinions from the public will range from positive to negative or indifferent. Scholars have also conducted a great deal of research on hate speech, and this body of information is growing every day [9].

The use of Natural Language Processing (NLP) in hate speech activities involves automating the process of identifying and capturing hateful social media information. Since these studies use naturally occurring human-generated information, NLP is involved. Social media data produced by users of social media platforms is a valuable resource for a variety of industries; include healthcare, science, and policy-making. Diverse information may be found in UGC (User Generated Contents) on various evaluation platforms and websites. Opinion extract algorithms and sentiments analysis methods are used to extract text from this material.

Additionally, these algorithms perform better when it comes to text classification's feature extraction stage [10]. The process known as "transfer learning" refers to the ability to apply knowledge from unlabeled data to related tasks with a little amount of labeled material. And with the aid of earlier data, that little labeled dataset attains high accuracy. Comparing NLP transformers to ML and DL approaches, they have shown promising accuracy gains in all practices. According to what they have said in their study, the main concept behind TL is to use data from related domains to support machine learning-based systems to achieve greater accuracy in the target domain. Therefore, in comparison to active learning as well as supervised learning, we can also state that transfer learning may be utilized to obtain outstanding outcomes with less human supervision.

A. Objectives

- Develop a strong neural networks or deep learning technique for classifying social media material into misleading and non-deceptive subcategories.
- Use methods like oversampling, under-sampling, or synthetic data production to lessen the effect of abnormalities.
- Choose appropriate assessment metrics to assess the model's performance, taking into account aspects such as accuracy, recalled, F1 score, and the area beneath the ROC curve.

2. Literature Review

False information [11] released under the pretence of being real news, usually with the intention of influencing political opinions, is referred to as fake news or created news. False news reports are a danger to public confidence in the government and, therefore, represent one of the main challenges confronting modern democratic nations This essay examines the developments made so far to address the problem. The article also offers a number of group methods for classifying news stories binary. Furthermore, one of the trendiest topics in the world of communication and language technology is the processing of natural languages, or NLP, and the use of Machine Learning (ML) is capable of understanding how to carry out crucial NLP jobs. In situations when hand programming is not feasible, this is often feasible and economical.

Fake news identification [12] in social networking with deep learning (DL) and Machine Learning (ML) models has received a lot of interest in recent studies. This Bio-inspired Artificial Intelligence (AI) with Natural Language Processing for Misleading Content Detection (BAINLP-DCD) method for social networking is presented in the current study paper. The suggested BAINLP-DCD approach aims to identify false or misleading information on social media. The BAINLP-DCD method uses data pre-processing to change the input dataset into a format that makes sense in order to achieve this. The BAINLP-DCD method makes use of a Multi-Head Self-awareness Simultaneous Short-Term and Long-Term Memory (MHS-BiLSTM) model for misleading content identification. Lastly, the African Vulture Optimizing Algorithm (AVOA) is used to choose the MHS-BiLSTM model's ideal hyper parameters.

Spam identification [13] has become essential due to the massive increase in the prevalence of spam material on social media. As more individuals utilize social media sites like Facebook, Twitter, YouTube, & email, the amount of spam material increases. People's use of social media is excessively increasing, particularly during this epidemic. Through social media, users get a large volume of text messages, and they are unable to identify the spam contents in these communications. Malicious links, applications, fake accounts, bogus reviews, rumors, and false news are all included in spam communications. The identification and management of spam material are critical to enhancing social media security. This study provides an in-depth analysis of recent advancements in social media spam detection and categorization.

For the processing of Natural Languages (NLP) applications [14], Deeper Learning (DL) models often handle sensitive data, necessitating security against intrusions and disclosures. Privacy is thus enforced by data protection legislation, such as the General Data Protection Regulation (GDPR) of the European Union. It is challenging to track the development of the literature despite the large number of privacy-preserving Natural Language Processing (NLP) approaches that have been published recently and the lack of established categories to group them into. This paper addresses the gap by providing a systematic assessment of over sixty DL approaches for secure privacy NLP published from 2016 and 2020. It covers the theoretical underpinnings of the methods, a study of privacy-enhancing technology, and a suitability analysis for real-world situations. First, we provide a new taxonomy that divides the current approaches into three groups: verification techniques, trusted methods, and data protecting methods. Secondly.

Utilizing online entertainment [15] to consume news has its downsides. Individuals search out and ingest media from the web due to its minimal expense, usability, and fast transmission of data. Notwithstanding, it additionally makes it more straightforward for "counterfeit news," or bad quality news that contains deliberately erroneous material, to broadly spread. The discovery of phony news via online entertainment presents particular elements and troubles, delivering recognition calculations got from customary news sources lacking or non-arousing. To begin with, it is testing and nontrivial to distinguish counterfeit news dependent exclusively upon news content; all things considered, we want to consolidate helper information, like client collaborations via online entertainment stages, to support the appraisal. Counterfeit news is deliberately written to hoodwink perusers into trusting mistaken data.

Since social media [16] is essential to human existence and contains news from all around the world, it might be referred to as the news powerhouse. Social media has made it simple to share information, which has improved our quality of life. However, it also diverts awareness from the speedier dissemination of false information. The objective of fake broadcast writing is to deliberately mislead readers into believing untrue facts. It is necessary to have an application that can distinguish among fake and true news in order to reduce the overuse of news in society. Many researchers have created applications over the years utilizing various tools, strategies, and algorithms to achieve the highest accuracy. As a result, in order to improve the accuracy of identifying mendacity broadcast beyond what the current

system offers, we are creating the Track Mendacity Broadcasting survey report.

The main platform [17] for rumour's to propagate these days is social media, and in this setting, the accuracy of the information is becoming more and more crucial. A number of investigators have been developing techniques to enhance the classification of rumour's in recent years, with good results, particularly in the identification of bogus news in social media. However, this activity gives many research opportunities as well as challenging obstacles because to the intricacy of spoken language. Out of 1333 contenders, 87 unique papers were carefully chosen for analysis in this survey. This study includes the primary techniques, text and user attributes, and datasets utilized in literature, covering a period of eight years of studies on false news utilized on social media.

Our social [18] lives are being profoundly impacted by the phenomena of fake news, especially in the political sphere. The detection of fake news is a rapidly developing field of study that presents some difficulties because there aren't many available resources (such as published literature and datasets). In this research, we offer a machine learning and n-gram analysis based fake news detection methodology. We examine and contrast six distinct machine classification methods as well as two distinct feature extraction methods. With an accuracy of 92%, the most successful experimental assessment makes use of Term Frequency-Inverted Documents Frequency (TF-IDF) as a feature extraction strategy and Linear Support Vector Machines (LSVM) as a classifier.

Social media networks [19] are a major source of fake or false material, which can purposefully mislead consumers by including propaganda, remorse, or incorrect data about a person, business, or service. Twitter is a highly popular social media tool, particularly in the Arab world, where a growing number of users are also experiencing a rise in the spread of false information. Researchers became aware of the need to create a secure internet space devoid of false information as a result. The purpose of this research is to offer a smart classification model that uses Machine Learning (ML) models, Natural Language Processing (NLP) approaches, and the Harris Hawks Optimizer (HHO) as a wrapper-based selection of features methodology for the early identification of fake news in Arabic tweets. This study used an Arabic Twitter corpus with 1862 already annotated tweets to evaluate the effectiveness of the suggested model.

For many years [20], identifying internet spamming has been a difficult task. Numerous methods have been successful in identifying bogus emails or spam comments on social media. However, because these messages mimic actual communications, it is challenging to develop a suitable technique for message filtering. After text has been pre-processed, Deep Learning models offer a viable alternative for text classification from the standpoint of Natural Language Processing (NLP). Specifically, among the models that function well for single and multi-label text classification issues are Long Short-Term Memory (LSTM) networks. This research presents a method for combining two distinct data sources: one for classifying fraud in emails and another for identifying spam in social network posts.

3. Methodology

An explanation is provided of the overall framework for the sentiment assessments of English tweets using several algorithms Fig.1 [21]. After describing the architecture of the systems supporting the task, each stage of the plan associated with this work is provided.

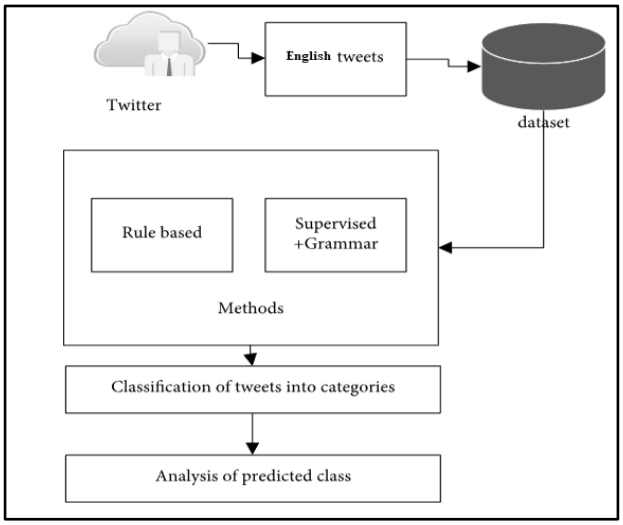

Fig. 1 Framework for sentiment analysis of English movie tweets [21].

A. The Proposed Model's Architecture

Four stages are used to identify user sentiment from tweets about English-language films. Tokenizers for each user tweet must be gathered and ready for usage as the first step. The second step is using tokenized keywords to identify segments of speech tags Fig. 2. Lastly, the genre category will be determined by applying a natural language toolkit to the tokenized content methods.

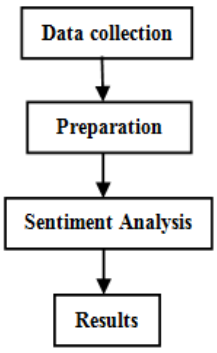

Fig. 2 Flow diagram for the suggested project [22].

B. Input Data Gathering

Gathering the data expected for opinion investigation — a classification technique — is the initial step. A few deep rooted datasets in English and related spaces are accessible for opinion examination. Just couple of information bases for feeling investigation can be found for genuine dialects other than English [22]. All datasets utilized in this study are taken from Twitter by using the hashtag (#), trailed by the name of the film or item utilizing Twitter's Programming interface. For trial investigation, it is an impressive issue to get an unlabelled dataset for English movies, since there is certainly not a present dataset accessible.

C. Task for Pre-processing

The subsequent stage is the pre-medicines of tweets. To eliminate clashing, blemished, and glowing data, the pre-handling of information is finished. To play out all information mining usefulness, information should be pre-handled. The primary occupation is to erase URLs. Typically, the Uniform Asset Finder doesn't assist in that frame of mind with evaluating the inclination.

D. Models of Sentiment Analysis

Solo strategies like as Credulous Bayes are much of the time utilized related to supporting vector machines to distinguish feeling subjects at the record level Fig. 3. In any case, more refined models — like etymological standards — are expected to order the (extremity) sentiments and thoughts communicated in casual messages (orientation) [23]. There are three usefulness based models in the proposed opinion system. Figure 3 represents two unique ways to deal with feelings research.

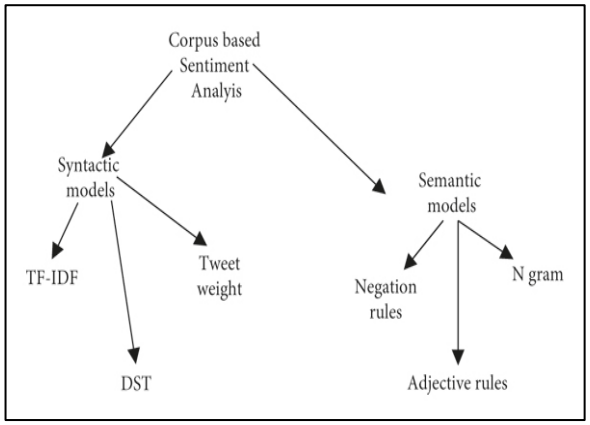

Fig. 3 Three different kinds of sentiment models [23].

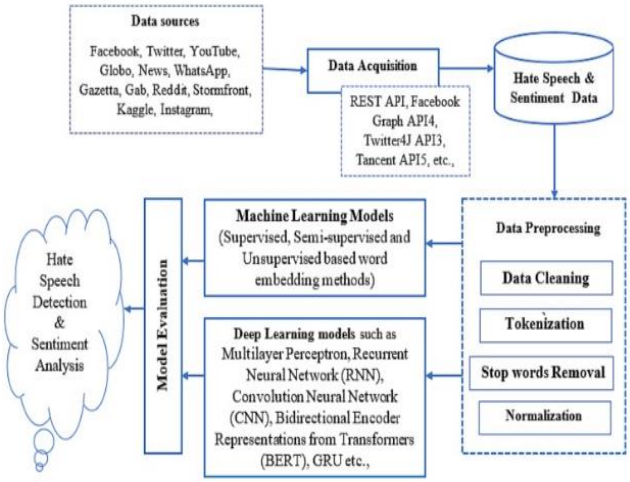

Fig. 4 The processes of Hate Speech Detection and Sentiment Analysis [23].

In order to precisely and reliably anticipate false and genuine news as well as equally hateful or not so hate material, the suggested model has to contain diverse elements because to the intricate nature of social networking site data on COVID-19 hate speech and fake news. The processes of Hate Speech Detection and Sentiment Analysis are shown in detail in Figure 4.

The findings of nine models that use transfer learning are shown in Tables 1 and are verified on the COVID-19 English tweet the data set COVID-19 false news the data set and extremist-non-extremist information set, respectively, using the performance metrics specified above: recall, precision, F1-score, and accuracy Fig. [24].

Table 1 Results of transfer learning-based methods for COVID-19 false news [25].

Model	Accuracy	Precision	Recall	F1 score
BERT-base	98.63	98.63	97.61	97.65
BERT-large	89.64	93.16	93.89	95.69
RoBERT-base	97.69	86.48	97.64	86.49
RoBERT-Large	92.46	76.66	94.65	85.64
Distil BERT	97.64	72.69	97.64	87.69
ALBERT-base-v2	93.46	98.46	97.68	84.69
XLM-RoBERTa-base	94.67	96.48	91.64	85.64
BERT-large	93.69	94.69	96.88	96.78

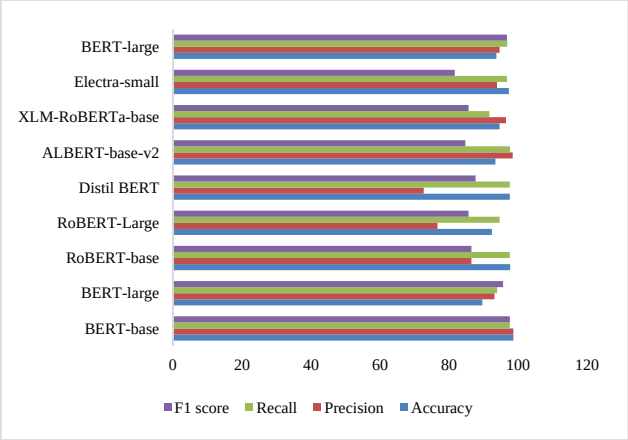

Fig. 5 Results of transfer learning-based methods for COVID-19 false news.

The findings unmistakably demonstrate that learned transfer classification models perform better when tested with test datasets from reputable sources Table 2 [25].

Table 2 Comparing suggested methods with cutting-edge methods.

Model	Accuracy	Precision	Recall	F1 score
BERT-base	98.87	89.64	98.68	96.48
BERT-large	96.79	87.69	87.95	86.49
RoBERT-base	87.69	86.79	78.69	85.69
RoBERT-Large	75.69	87.96	85.69	78.96
Distil BERT	98.63	74.98	87.96	75.69
ALBERT-base-v2	97.58	78.96	98.96	71.96
XLM-RoBERTa-base	94.68	86.97	48.97	75.89
BERT-large	96.98	87.56	96.49	89.64

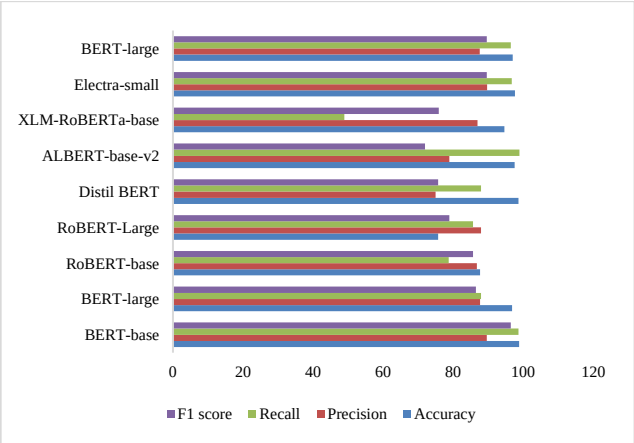

Fig. 6 Comparing suggested methods with cutting-edge methods. [25].

4. Results

A tool for analysis was created that integrates all algorithms based on Python and NLTK. The application automatically displays the sentiment values of tweets about English-language movies at the genre and polarity levels [26]. Figure 6 depicts the emotions around the English movie. The grammatical performance is shown in Table 3.

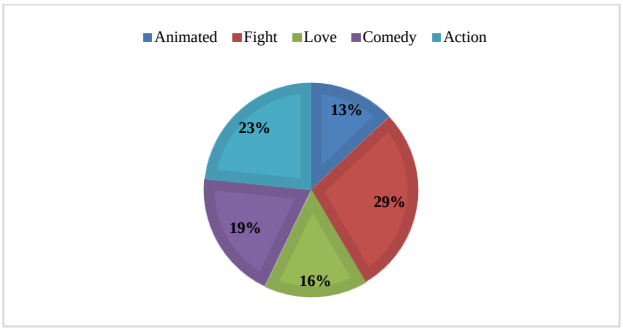

Fig. 7 Outcomes of analysing feelings using grammatical rules.

Table 3 Grammar performance in English SA [26].

Movie	Method	Accuracy
English	TF-IDF ranking	28.96
	TF-IDF+DST	24.98
	Tweet Weight	24.96
	Negation rule	28.98
	Adjustment rule	89.69

The average accuracy of the metaphorical grammatical model is 64.72 percent higher than that of the TF-IDF and other morphological models, as shown in Figure 7 (A. Onan, 2019) [27]. Using suggested grammatical rules, the sentiment model has examined tweets to determine polarity, genre classifications, and other algorithms.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}. \quad \ldots 1$$

$$Precision = \frac{TP}{TP+FP}. \quad \ldots 2$$

$$Recall = \frac{TP}{TP+FN}. \quad \ldots 3$$

$$F1 = \frac{2}{precision^{-2}+recall^{-1}}. \quad \ldots 4$$

$$= 2.\frac{Precesion.recall}{Precesion+recall} \quad \ldots 5$$

$$= \frac{TP}{TP+\left(\frac{1}{2}\right)(FP+FN)}. \quad \ldots 6$$

The findings show that when complicated sentences are taken into account and semantic elements are better integrated, general grammar for unfavorable rules and adjectives rules is improved. The suggested grammatical guidelines analyze all types of tweets to ascertain feelings (simple, compound, and complicated). The best emotion

model is the grammar rule-based model, which has an accuracy of 64.72 percent. In comparison to previous sentiment models, the grammar-based algorithm may perform 20% better when TF-IDF, tweet thereby affecting their and regulatory modeling results are evaluated. Findings indicate that machine learning techniques by themselves are not beneficial to emotions.

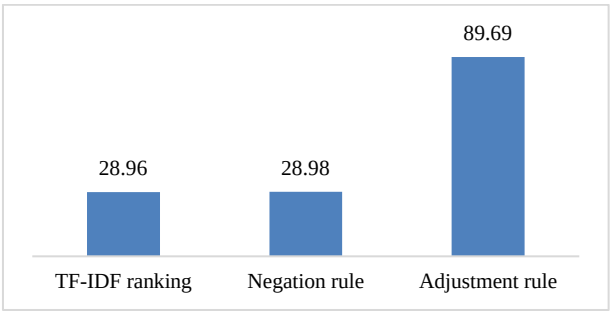

Fig. 8 SVM classifier performance using grammatical rules [27].

Adjectives and negative-based rule of grammar are thus employed as a feature for classifiers for evaluating device study techniques with grammar rules, hence improving machine study classifier performance. Table 4 displays the SVM classification results when combined with grammatical rules.

Table 4 SVM classifier performance using grammatical rules.

Movie	Method	Accuracy
English	TF-IDF ranking	26.69
	Domain-specific Tags	48.69
	Tweet Weight	89.64
	Negation rule	87.69
	Grammar Rule	89.65
	SVM + Grammar rules	69.69

It is shown that the grammar rules technique performs better when paired with the SVM method of classification than it does when combined with any other emotion models. The accuracy varies by 7% between the learning frameworks and the grammar rule Fig. 9. The outcome demonstrated once again how the vocabulary-based machine learning model handled both the clarity and ambiguity of English grammar [28]. Grammar rules are used to support the excellent promise given for future progress.

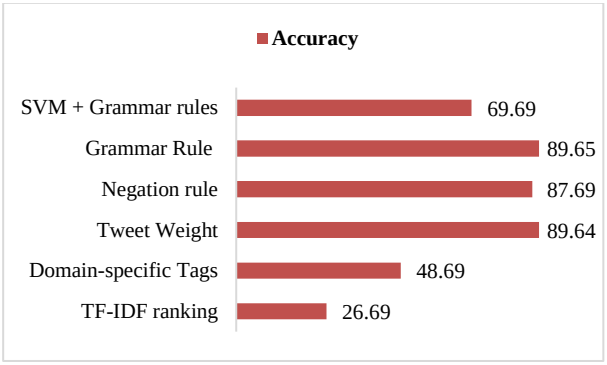

Fig. 9 SVM classifier performance using grammatical rules [28].

This is due to the fact that there are as many movie domain options as there are tweets accessible on this domain. The objective is to validate the efficacy of machine learning techniques and grammatical rule algorithms across domains, especially in cases when the volume of the product field tweets is limited Table 5 [29]. For master classification, the training's domain-independent characteristics are extracted. The terms that are common to all domains are traits that are independent of domains.

Table 5 Product domain performance analysis [29].

Corpus	Method	Accuracy
Mobile Phone	TF-IDF ranking	28.96
	Domain-specific Tags	27.89
	Tweet Weight	28.79
	Negation rule	24.69
	Grammar Rule	89.67
	N-Gram Model	29.78
	Naïve Bayes	58.96
	SVM	47.89
	SVM + Grammar Rule	58.98

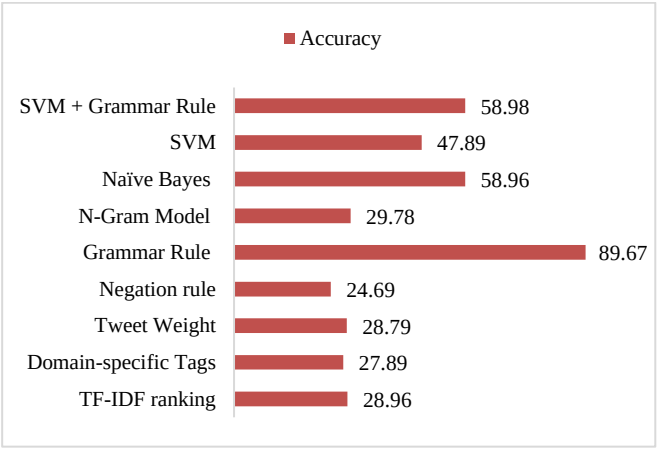

Fig. 10 Product domain performance analysis [29].

Transferring the semantic context from a particular field to another is made possible by this function. In this study, separate domain adjectives are extracted using grammatical rules in order to analyse cross-domain sentiments. The outcomes of the three theories of emotion are presented in Fig 10.

5. Discussions

The text categorization problem is contrasted between the suggested work and the most modern techniques available. Modern methods make use of deep learning and machine learning-based methods like CNN, HDLTex, and Naïve Bayes. The goal of the project is to classify tweets using the most advanced methods and suggested, refined ways based on learning transference.

Using the extremist-non-extremist dataset, the accuracy performance of the tested methods is compared. With an accuracy of 76.5% of on the extremist-non-extremist dataset, HDLTex had the lowest efficiency. When BERT-large and BERT-base were used, the suggested method had the maximum accuracy of 99.71%.

In order to do this, spatial input, such as images and videos, is predicted using the LSTM + CNN model. In this model, CNN is used for feature extraction, while LSTM aids with prediction. The model that is currently being used is just one-dimensional (1D) due to the way it has been fitted to the content of the dataset. The LSTM + CNN 1D architectural model overview is shown in Figure 10 [30].

Based on the findings, we may conclude that word embedding's could increase classification efficiency. Given that BERT analyzes every phrase without any particular direction in mind, it performs a better job figuring out the implications of homophones than before NLP techniques, such as Glove integrating methods. Word embedding created using the BERT model has been proven to be far more effective than unmoved Glove word integration when used with LSTM.

Although they were found to be less successful than BERT (in identifying toxicity in text documents), word embedding's trained on big corpuses, such as Glove based on Wikipedia, for example, Gig word, and Twitter, were also proven to be beneficial in improving classification accuracy.

6. Conclusion

The current dataset was used to test many algorithms, including SVM and Nave Bayes, and the outcomes were monitored. The findings indicate that the SVM model outperform syntactic approaches in the classification of film genres. Thus, the paper proposed integrating both models and tracing the outcomes. Even if the suggested techniques for setting up a feeling framework operate well, it is important to evaluate how well they function in real time given the makeup of the system. As part of further development, the suggested model could be subsequently tweeted in actual time.

The COVID-19 fake news dataset, COVID-19 English tweet dataset, extremist-non-extremist dataset, and BERT-base, BERT-large, RoBERTa-base, RoBERTa-large, DistilBERT, ALBERT-base-v2, Electra-small, and BART-large are the nine transfer learning models that are used in this study to classify binary text. These datasets, which are extracted from trustworthy sources, are used for the experiments. Accuracy, recall, precision, and the F1 score was are the assessment measures used to assess all transfer learning models.

Discrimination in machine learning algorithms for toxicity categorization that results in erroneous classification has been identified as an issue to alleviate in several previous research studies. This is clearly seen in the way that public chat pages as well as internet message boards classify toxicity.

Future work

The objective of our research is to conduct experiments on larger and more multiclass classification data in the future. For categorizing texts, we may also utilize several language datasets. Since emoticons are often used on social media to symbolize expressions, it would be beneficial to add them.

References

[1] H. Aldabbas, A. Bajahzar, M. Alruily, A. A. Qureshi, R. M. Amir Latif, and M. G. Farhan, "Google play content scraping and knowledge engineering using natural language processing techniques with the analysis of user reviews," Journal of Intelligent Systems, vol. 30, no. 1, pp. 192–208, 2020.

[2] S. Soni and K. Roberts, "An evaluation of two commercial deep learning-based information retrieval systems for COVID-19 literature," Journal of the American Medical Informatics Association, vol. 28, no. 1, pp. 132–137, 2021.

[3] A. B. Tufail, I. Ullah, R. Khan et al., "Recognition of ziziphus lotus through aerial imaging and deep transfer learning approach," Mobile Information Systems, vol. 2021, Article ID 4310321, 10 pages, 2021.

[4] Y. Dai, J. Wu, Y. Fan et al., "MSEva: a musculoskeletal rehabilitation evaluation system based on EMG signals," ACM Transactions on Sensor Networks, 2022.

[5] R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, and M. Esposito, "A novel covid-19 data set and an effective deep learning approach for the de-identification of Italian medical records," IEEE Access, vol. 9, Article ID 19097, 2021.

[6] J. L. Izquierdo, J. Ancochea, J. B. Soriano, S. C. R. Group, and others, "Clinical characteristics and

prognostic factors for intensive care unit admission of patients with COVID-19: retrospective study using machine learning and natural language processing," Journal of Medical Internet Research, vol. 22, no. 10, Article ID e21801, 2020

[7] R. U. Mustafa, P. M. Saqib Nawaz, and F. viger, "Early detection of controversial Urdu speeches from social media," Data Sci. Pattern Recognit, vol. 1, no. 2, pp. 26–42, 2017.

[8] K. Kumar, B. S. Harish, and H. K. Darshan, "Sentiment analysis on IMDb movie reviews using hybrid feature extraction method," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 5, no. 5, p. 109, 2019.

[9] Wang, "Stock market forecasting with financial micro-blog based on sentiment and time series analysis," Journal of Shanghai Jiaotong University, vol. 22, no. 2, pp. 173–179, 2017.

[10] H. Xia, Y. Yang, X. Pan, Z. Zhang, and W. An, "Sentiment analysis for online reviews using conditional random fields and support vector machines," Electronic Commerce Research, vol. 20, no. 2, pp. 343–360, 2020.

[11] Sharifani, K., Amini, M., Akbari, Y., &AghajanzadehGodarzi, J. (2022). Operating Machine Learning across Natural Language Processing Techniques for Improvement of Fabricated News Model. International Journal of Science and Information System Research, 12(9), 20-44.

[12] Albraikan, A. A., Maray, M., Alotaibi, F. A., Alnfiai, M. M., Kumar, A., & Sayed, A. (2023). Bio-Inspired Artificial Intelligence with Natural Language Processing Based on Deceptive Content Detection in Social Networking. Biomimetics, 8(6), 449.

[13] Kaddoura, S., Chandrasekaran, G., Popescu, D. E., &Duraisamy, J. H. (2022). A systematic literature review on spam content detection and classification. PeerJ Computer Science, 8, e830.

[14] Sousa, S., & Kern, R. (2023). How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing. Artificial Intelligence Review, 56(2), 1427-1492.

[15] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter, 19(1), 22-36.

[16] Prabhu, S. N., Singh, A., Pasanna, A., & Ray, A. (2021, December). Track Mendacity Broadcast using Natural Language Processing. In 2021 International Conference on Forensics, Analytics, Big Data, Security (FABS) (Vol. 1, pp. 1-6). IEEE.

[17] De Souza, J. V., Gomes Jr, J., Souza Filho, F. M. D., Oliveira Julio, A. M. D., & de Souza, J. F. (2020). A systematic mapping on automatic classification of fake news in social media. Social Network Analysis and Mining, 10, 1-21.

[18] Ahmed, H., Traore, I., &Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. In Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1 (pp. 127-138). Springer International Publishing.

[19] Thaher, T., Saheb, M., Turabieh, H., &Chantar, H. (2021). Intelligent detection of false information in Arabic tweets utilizing hybrid Harris hawks based feature selection and machine learning models. Symmetry, 13(4), 556.

[20] Baccouche, A., Ahmed, S., Sierra-Sosa, D., &Elmaghraby, A. (2020). Malicious text identification: deep learning from public comments and emails. Information, 11(6), 312.

[21] A. Krishna, V. Akhilesh, A. Aich, and C. Hegde, Sentiment Analysis of Restaurant Reviews Using Machine Learning Techniques, vol. 545, Springer, Singapore, 2019.

[22] Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, vol. 2012, 10 pages, German, 2017.

[23] P. Kirilenko, S. O. Stepchenkova, H. Kim, and X. Li, "Automated sentiment analysis in tourism: comparison of approaches," Journal of Travel Research, vol. 57, no. 8, pp. 1012–1025, 2018.

[24] W. H. Bangyal, J. Ahmad, H. Tayyab, and S. Pervaiz, "An improved bat algorithm based on novel initialization technique for global optimization problem," International Journal of Advanced Computer Science and Applications, vol. 9, no. 7, pp. 158–166, 2018.

[25] W. H. Bangyal, J. Ahmad, and H. T. Rauf, "Optimization of neural network using improved bat algorithm for data classification," Journal of Medical Imaging and Health Informatics, vol. 9, no. 4, pp. 670–681, 2019.

[26] F. Ahmadi, Sonia, G. Gupta, S. R. Zahra, P. Baglat, and P. Thakur, "Multi-factor biometric authentication approach for fog computing to ensure security perspective," in Proceedings of the 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), pp. 172–176, IEEE, New Delhi, India, March 2021.

[27] A. Onan, "Topic-enriched word embeddings for sarcasm identification," Advances in Intelligent Systems and Computing, Springer, New York, NY, USA, 2019.

[28] H. Bulut, S. Korukoğlu, and A. Onan, “Ensemble of keyword extraction methods and classifiers in text classification,” Expert Systems with Applications, vol. 57, pp. 232–247, 2016.
[29] T. Bolukbasi, K. W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, “Man is to computer programmer as woman is to homemaker? debiasing word embeddings,” Advances in Neural Information Processing Systems, vol. 29, pp. 4349–4357, 2016.
[30] S. Korukoğlu, “A feature selection model based on genetic rank aggregation for text sentiment classification,” Journal of Information Science, vol. 43, no. 1, pp. 25–38, 2016.