International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Social Media Sentiment Analysis with Multi-Token Concatenated Embedding and Semantixpert Probabilistic Classifier for Fake News Detection

Mrs. B. Hemalatha*[1], Dr. M. Soranamageswari[2]

Submitted:15/03/2024 **Revised**: 25/04/2024 **Accepted**: 02/05/2024

Abstract: In the field of false news identification, sentiment analysis is an effective strategy for enhancing detection accuracy. Sentiment analysis is critical for understanding user emotions and intents, but traditional methods frequently neglect the complex interplay between text and emojis, particularly in recognizing sarcasm. To address this limitation, a novel approach, the "Multi-Token Concatenated Embedding and SemantiXpert Probabilistic Classifier" is proposed which is aimed at improving sarcasm detection, polarity prediction accuracy, and overall sentiment prediction accuracy. The existing sentiment analysis approaches fail to capture the indirect interaction between text and emojis, resulting in the loss of crucial sentiment cues and misunderstanding of sarcasm in news content. Hence, Multi-Token Concatenated Embedding approach is introduced, which includes Multi-Run Byte Pair Encoding (MR-BPE) to capture sub-word patterns and the Concat-ViLT model to encapsulate the textual representation of emojis, thus enhancing the model's ability to understand the multimodal context of text-emoji pairs. Moreover, existing semantic classification methods are based on lexical patterns and mean-zero feature assumptions, which leads to misclassifications and lower accuracy in recognizing the genuine emotional tone of mixed sentiment comments. For semantic classification, a SemantiXpert Probabilistic Classifier is proposed which includes a Contrastive Semantic Clause Filter Network to understand user intentions and domain context and a Polarized Probabilistic Classifier is utilized to enhance the polarity prediction thereby reducing misclassifications and improving accuracy in sentiment prediction based fake news detection. As a result, the proposed model outperforms existing methods, achieving higher accuracy, sensitivity, and AUC values.

Keywords: *Byte Pair Encoding, Sentiment analysis, polarity prediction, fake news detection, social media.*

1. Introduction

Fake news detection in the realm of social media necessitates advanced tools capable of deciphering the intricate web of emotions and opinions expressed through textual content. Sentiment analysis has developed as an essential technique for assessing the wide variety of emotions represented in textual material in the dynamic world of social media, where millions of people communicate in real time. As the digital world evolves, emojis and words become increasingly important in understanding the complex emotions that individuals seek to communicate. Sentiment analysis on social media, which incorporates both text and emojis, allows for a more in-depth evaluation of user opinions, attitudes, and emotions. This multidisciplinary strategy includes the emotional and visual features of emojis while also capturing the intricacies of English, providing a complete understanding of the continuously altering emotions that saturate online interactions. Researchers, advertisers, and social media platforms now have additional ways to learn about the complex range of human expression in the digital age due to the convergence of text and emojis in sentiment analysis.

[1]*Government Arts and Science College, Avinashi, Tamilnadu – 641 654.*
[2]*Associate Professor, Computer science,*
Government Arts College (Autonomous), Coimbatore, Tamil Nadu 641018.
** Corresponding Author Email: sabahema99@gmail.com*

Emojis are an essential component of sentiment analysis because they help to accurately represent various aspects of human emotion and expression in digital communication. Emojis are pictorial symbols that indicate emotions. Emojis add a nonverbal element to written material by giving sentiments expressed online more context [1-4].

Emojis communicate certain emotions, sarcasm, and tone in a more obvious and visually understandable way than standard text analysis provides. Furthermore, emojis' enormous variety makes it possible to express a wide range of emotions more accurately than words alone, which can be difficult at times. Because of this diversity, sentiment analysis is more accurate and sensitive to the subtle differences and cultural quirks of various user groups. The methods now in use for handling emojis in sentiment analysis are often categorized into two categories: embedding-based methods and lexical methods. Lexical techniques frequently use sentiment lexicons, in which the perceived emotional tone of each emoji is used to assign a predetermined sentiment score. Although these techniques provide an inexpensive method to add emojis to sentiment analysis, they frequently fail to convey how context-dependent and complex emoji usage is. Conversely, embedding-based techniques use pre-trained word embeddings or emoji-specific models to try and understand the semantic connections between emojis and words in context. Nevertheless, these methods have difficulties due to language's constant change and the quick creation of new emojis, which aren't well-represented in the embeddings that are now in use [5-8].

Sentiment analysis pre-processing techniques that combine text and emoji are essential and ask for the careful blending of linguistic and visual components. To ensure consistency in the textual data, textual pre-processing first entails common techniques like tokenization, stemming, and lowercasing. But when emojis are introduced, extra consideration is required due to their unique function in communicating emotion. Emojis hold important sentiment information, thus it's important to identify and preserve these during tokenization. In this sense, using a specific emoji tokenizer is possibly useful since it ensures that every emoji is handled as a distinct token, maintaining its semantic meaning. Therefore, it's common practice to remove stop words from text and emoji components to reduce noise. Although this is essential for textual material, if it is applied carelessly to emojis, important sentiment-bearing emoticons can be unintentionally lost. Consequently, a targeted strategy for stop word removal that takes into consideration the special qualities of emojis needs to be used. Using a variety of methods to interpret the meaning and emotion conveyed in both text and emoticons is known as semantic categorization for combined text and emoji [9–12].

One prevalent way for semantic classification is to use deep learning models, such as recurrent neural networks (RNNs) or long short-term memory networks (LSTMs), to capture the sequential connections in text and emoji sequences. More precise semantic categorization is made feasible by these algorithms' ability to recognize intricate patterns and connections between words and emojis. Using attention mechanisms is another method that enables the model to highlight words and emojis that are most pertinent to sentiment analysis by focusing on particular portions of the input. When handling lengthy sequences where certain aspects are more important than others in defining sentiment, attention methods come in handy. In addition, combining text and emoji semantic classification can aid in transfer learning, which is the process of fine-tuning pre-trained models on huge data sets for specific tasks. Models including BERT (Bidirectional Encoder Representations from Transformers) have successfully extracted contextual information from both textual and visual components. Although these methods have advanced, there are still several shortcomings in the semantic categorization of combined text and emoji [13–15]. Therefore, to increase the robustness of the semantic analysis, a unique technique for pre-processing and semantic categorization is required.

1.1. Main contribution of this study

The following methodological and experimental contributions have been achieved by this paper:

- The challenge of capturing complex sentiment cues from text-emoji interactions is addressed by the proposed pre-processing technique, named Multi-Token Concatenated Embedding, which integrates Multi-Run Byte Pair Encoding (MR-BPE) and Concat-ViLT models to effectively capture both textual and visual representations.

- To accurately capture semantic relationships and user intentions, a Contrastive Semantic Clause Filter Network is introduced, the CLIP model is used to understand semantic relationships in user comments and the Mixed-Clause Semantic Separator Algorithm to identify domain-specific contexts and user intentions. This contributes to a more precise semantic analysis.

- Improved Polarity Prediction the Polarized Probabilistic classifier is used, in which ELECTRA distinguishes important semantic representations to determine polarity accurately and the Naive Bayes classifier, which combines semantic connections and polarization scores, decreases misclassifications and improves overall sentiment prediction accuracy.

The paper is arranged as follows: Section 2 discusses the existing literature on sentiment analysis on social media, including the role of emotions and language in sentiment analysis, and summarizes the available approaches. Section 3 explains the suggested approach and workflow for this work. Section 4 explains the datasets, as well as the experimental methodology, the analytical results, and the comparision to previous investigations. Finally, Section 5 sums up the study.

2. Literature Survey

Due to the rich diversity of emotions expressed on digital communication platforms and their exponential expansion, there has been a rise in interest in the field of sentiment analysis in social media. This literature review investigates new developments and creative approaches in social media sentiment analysis for fake news detection put forth by various researchers for analyzing the intricate relationship between words, images, emojis, and emotions on social media platforms.

Amira Samy Talaat [16] presented four deep learning models that combine the bidirectional Long Short-Term Memory (BiLSTM) and Bidial Gated Recurrent Unit (BiGRU) algorithms with the BERT algorithm. Pre-trained word embedding vectors were employed in the suggested methods to improve accuracy and analyze the influence of BIGRU in text sentiment classification without and with emojis. This made model fine-tuning much easier. The proposed methods were compared to two pre-trained BERT models and seven additional models developed for the same purpose using classical machine learning. To improve overall performance, however, effective techniques must be applied throughout the feature extraction and feature selection stages.

Lydia Bryan-Smith et al [17] developed a unique flood forecasting and monitoring model that employed a sentiment analysis transformer network based on text and images to estimate flooding levels. Sentiment research was conducted, comparing VADER, RoBERTa, a Transformer encoder, and CLIP utilizing text and photos from Twitter. By integrating visual input, the CLIP-based model considered both text and images, possibly reducing the gap between AI and human raters. Two different methodologies were used to explain the overlap of dots indicating tweets from the same Place area: first, the size of the points was enlarged by the number of tweets included inside; second, the points overlaying on a single coordinate were augmented with Gaussian noise. However specific communities or events didn't receive much activity on social media, which could result in an inaccurate or biased portrayal of the real situation.

Chuchu Liu et al [18] evaluated the impact of emoji introduction and the ambiguity of emoji tags using well-known supervised and unsupervised learning approaches, including neural network algorithms (LSTM) for supervised learning, rule-based algorithms for unsupervised learning, and classification algorithms (SVM). It was shown that compared to unsupervised learning algorithms, supervised learning algorithms frequently produce higher accuracy. Furthermore, deep learning algorithms consistently produced the greatest results. Moreover, it was shown that emoji usage improves the performance of SA algorithms and that emoji tag words were directly used in feature creation for classifier training. But to strengthen the strategy's resilience, further study needs to be done in terms of platform and scenario comparisons as well as usage patterns of more comprehensive emojis.

Murtuza Shahzad et al [19] predicted the emotive resonance of Facebook posts about study articles. "Like" replies were found to be the most common when Facebook users' reactions to research articles and postings were examined. Machine learning techniques were used to predict the tone of Facebook posts related to scientific papers. The attitude expressed in the research article's title, the abstract's sentiment, its length, the number of authors, and the study domain were all taken into account while building the models. The five classifiers used were Random Forest, Decision Tree, K-Nearest Neighbours, Logistic Regression, and Naïve Bayes. The models were evaluated using metrics such as F-1 score, recall, accuracy, and precision. In terms of accuracy, the Random Forest classifier was the best model, with values of 86% and 66% for two- and three-class labels, respectively. Nonetheless, disparities in sentiment interpretation affect how reliable the model's predictions are, even while the results exactly match users' actual emotional reactions.

Shelley Gupta et al [20] presented an emoji-based paradigm for the cognitive-conceptual-affective processing of emotion polarity that was based on text linguistic patterns and emojis. A text-based parser that combined a variety of emojis with suggested linguistic elements to depict emotions converted voice segments into n-gram patterns. A total of 1,68,548 tweets from 650 well-known people were downloaded worldwide. The findings demonstrated how the suggested framework for natural language processing suggested that adding emojis often appears to shift the overall polarity of the view. By extension, a vocabulary of emotions was used to assess the polarity of text, and the CLDR name of each emoji was used to assess the appropriate polarity of emoji patterns. However, additional domain knowledge has to be provided for this model to recognize textual patterns more reliably.

Alsayat [21] provided a language model for group deep learning that utilised a complex word embedding approach and an LSTM network for sentiment analysis. The word embedding layer transformed the Twitter data into a feature set, which was then fed into a proprietary deep-learning language model. To detect predicates and suffixes in training data, the word embedding layer employs an LSTM network and FastText word embedding. Create a hybrid ensemble model for sentiment analysis that makes use of the considerable disparities between existing methods.

Pathak et al [22] provided a topic-level attention method for an LSTM network that used online latent semantic indexing with regularization restrictions to retrieve a topic. For it for the suggested model to function online, small content extracted from social media sites was serially processed. The online latent semantic indexing model, which was based on deep learning, continuously constructed the topic model based on streaming input. The previously learned term-topic matrix was used to predict the topic vector of a newly provided phrase. Subsequently, the anticipated topic vector and any additional sentences were used to update the term-topic matrix. Nonetheless, this model oversimplifies complex user-generated content where many topics and emotions coexist by assuming that each streaming syllable only has a single topic.

Bagus et al [23] developed a supervised machine-learning technique to identify sarcasm in text and emojis Emoji polarity scores were determined by using the CNN to compute sentiment polarity values from text and emoji weighting. Emojis and linguistic characteristics were highlighted in this strategy as important clues for identifying sarcastic tweets. The achieved accuracy rate of 87.5% for the CNN model and an f1-score of 87.59% for the sarcasm detection engine signify a high level of effectiveness in identifying sarcastic expressions. The incorporation of sarcasm detection based entirely on textual data, investigating subtleties like the polarity distance between words in a tweet and their nearest neighbors, represents an area that was found for development.

Xu et al [24] provided the latest multi-view deep learning technique for sentiment analysis that took into account non-textual components like emojis. The results acknowledged textual and emoji perspectives' individual and combined contributions to sentiment analysis. The proposed technique regarded textual and emoji perspectives as distinct views of emotional information for the sentiment classification model. three emoji handling techniques such as Emoji Replacement, Emoji Scoring, and Emoji embedding were suggested. Additionally, either separately or in combination, the sentiment classifiers enhanced the performance of the classifiers when handling emoji features processed by the three approaches. However, additional domain knowledge has to be provided for this model to recognize textual patterns more reliably.

Naresh Kumar et al [25] proposed an Intelligent Senti-net-based Lexicon Generation Metho to accurately categorizes feelings on social media. The first phase was pre-processing, which required eliminating stop words and extracting the keywords. Following keyword extraction, a holoentropy-based lexicon was constructed and utilized to investigate the link between characteristics and cluster structure. The extracted features were then used in the categorization process. This work used optimized NN to classify phrases, with training carried out by a new SI-SLnO algorithm that selected the appropriate weight. Finally, during Int SentiNet sentiment classification, the lexicon and NN outputs were combined to produce the final sentiment. Furthermore, static lexicons created with holoentropy offer flaws in sentiment classification for new language patterns, slang, and cultural shifts, limiting their adaptability and usefulness.

The literature reviews existing various methodologies and approaches for sentiment analysis, each with its drawbacks and limitations. In [16] reliance on pre-trained word embeddings and static lexicons may limit adaptability to evolving language patterns and [17] lead to biased or incomplete representations of real situations. [18] investigates emoji usage in sentiment analysis, yet the effectiveness of supervised learning algorithms and deep learning models vary across different platforms and scenarios. In [19] disparities in sentiment interpretation affect the reliability of predictions, [20] the model's reliance on language features and emojis require additional domain knowledge for improved reliability, [21] the effectiveness of ensemble deep learning methods varies based on dataset characteristics, and in [22] oversimplification of user-generated content may lead to inaccurate topic extraction. [23] the model's effectiveness depends on the dataset's linguistic characteristics and sarcasm shades. In [24] additional domain knowledge is required for reliable recognition of textual patterns and in [25] static lexicons limit adaptability to new language patterns and cultural shifts.

3. Social Media Sentiment Analysis with Multi-Token Concatenated Embedding and Semantixpert Probabilistic Classifier for Fake News Detection

In the field of false news authentication, using emojis in sentiment analysis significantly enhances the efficacy of spotting misinformation. The integration of emojis in sentiment analysis assists in identifying fake news more effectively than models that rely solely on text. Text by itself does not always convey the whole context and emotional tone that emojis offer. The enhanced sentiment analysis model accurately detects deceptive or misleading content by capturing sentiment cues from text and emojis, enhancing understanding of user intent and emotional undertones in social media posts, leading to more accurate fake news detection. Emojis transmit complex and deep feelings and sentiments, hence their use in natural language processing tasks, such as sentiment analysis, is an attractive path for investigation. The intricate link between text and emojis has made integrating emoticons into text extremely difficult for sentiment analysis, frequently resulting in misinterpretations of sentiment. Hence a **"Multi-Token Concatenated Embedding and SemantiXpert Probabilistic Classifier"** for fake news detection is proposed to improve sarcasm detection, polarity prediction accuracy, and overall sentiment prediction. Existing pre-processing methods, such as Emoji translation or removal, fail to capture the intricate interplay between text and emojis, resulting in the loss of valuable sentiment cues. Particularly, instances where text and emojis express contradictory sentiments or are used sarcastically present contextual dissonance, confounding traditional pre-processing techniques reliant on lexical and syntactic patterns. To address this, the proposed methodology introduces Multi-Token Concatenated Embedding, which utilizes Multi-Run Byte Pair Encoding (MR-BPE) to capture both common sub-words and specialized patterns in text data iteratively. Additionally, a Concat-ViLT model is employed to merge textual and visual representations of emojis, enhancing the understanding of the multimodal context of text-emoji pairs and improving accuracy in recognizing sarcasm or deception in news content.

Additionally, semantic classification in sentiment analysis is crucial for accurately interpreting and categorizing the emotional tone within textual data. However, existing models often overlook the significance of less influential features, assuming their negligible contribution to overall variation. This oversight can lead to reduced content awareness and accuracy in tone and polarity prediction, particularly in cases where comments contain a mix of positive and negative words. Hence, for the semantic classification, a novel "SemantiXpert Probabilistic Classifier" is proposed which reduces misclassifications and improves the overall accuracy in sentiment prediction. In this classification model, a Contrastive Semantic Clause Filter Network is introduced, which includes CLIP to capture semantic connections, and a Clause Separator Algorithm to break user comments into clauses, allowing for more exact semantic analysis. This network enhances polarity prediction accuracy by comprehending the specifics of domain and user intention. Furthermore, to improve the polarity prediction a Polarized Probabilistic classifier, utilizing ELECTRA, distinguishes important semantic representations to determine polarity accurately. The integration of the Naive Bayes classifier further enhances sentiment classification by considering both semantic relationships and polarization scores, ultimately reducing misclassifications and improving overall accuracy in sentiment prediction.

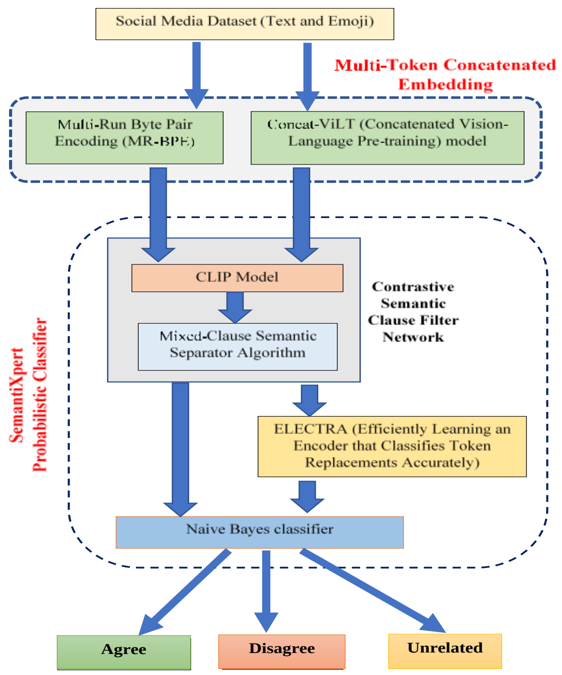

Fig. 1. Architecture of the proposed framework

Fig. 1 depicts the construction of the proposed framework. MR-BPE is first used in this method to tokenize text input iteratively, catching both specialized patterns and common sub-words. The comprehension of multimodal text-emoji pairs is then improved by combining textual and visual representations of emojis using a Concat-ViLT model. A Contrastive Semantic Clause Filter Network is used by the SemantiXpert Probabilistic Classifier, which combines a Clause Separator Algorithm for domain knowledge with a CLIP for semantic connection modeling. This network allows for exact semantic analysis by dividing user comments into clauses. After that, the Polarized Probabilistic classifier uses ELECTRA to identify significant semantic representations to determine polarity. Lastly, to reduce misclassifications and increase overall accuracy in sentiment prediction, a Naive Bayes classifier integrates both semantic links and polarization scores for sentiment classification. The details of every part of the framework are discussed in detail in the remaining parts of this section

3.1. pre-processing

To enhance the accuracy of polarity prediction in sentiment analysis tasks, mainly in the context of social media datasets containing both text and emojis, a novel preprocessing technique known as Multi-Token Concatenated Embedding is developed.

3.1.1. Multi-Run Byte Pair Encoding (MR-BPE)

A Multi-Run Byte Pair Encoding (MR-BPE) technique is developed to manage the complexities of text data, especially with uncommon words and specialized patterns. BPE is a data compression technique used for text data, particularly in linguistic processing tasks such as tokenization. The input data consists of text from social media posts along with emojis. Each post is represented as a sequence of characters, including both textual content and emojis. Text and emoji are considered as characters or bytes, and the collection of data is initially divided into smallest unit tokens denoted as L. Create a priority queue P containing all token pairs in L, ordered by descending frequency and the initial vocabulary is V_o. In the first run, MR-BPE performs the standard BPE merging process. This involves identifying the most frequent pair of adjacent characters or tokens and merging them into a new token t, and include t in new vocabulary (V). In each iteration, MR-BPE identifies the token pair (a, b) with the highest frequency, which is represented in equation (1)

$$(a, b) = arg \max_{(x,y)\in M} f(x, y) \tag{1}$$

Where $f(\cdot)$ as the frequency of a token or token pair within L. The pair with the highest frequency $f(x, y))$ is selected and merged, this process is repetitive for a set number of steps (M). After the first run, the resulting vocabulary (V) is analyzed to identify any remaining specialized tokens that are not adequately captured. All significant subword patterns cannot be captured by a single merging sequence, particularly in noisy and mixed datasets including social media content. Hence, the BPE extends this concept by performing multiple runs of the BPE merging process. Run multiple BPE runs, concentrating on various textual parts or patterns. The algorithm gives merging pairs that are less common or ignored in earlier rounds priority in each run. Iteratively carry out this practice, improving the vocabulary with each attempt. The algorithm for Multi-Run Byte Pair Encoding is given in algorithm 1.

Algorithm 1: Multi-Run Byte Pair Encoding

Input: Text from social media posts along with emojis.

1. Split corpus C into a sequence of smallest unit tokens L
2. Initialize an empty priority queue Q, Text corpus C containing a mixture of text and emoji data. Predefined vocabulary size N, Number of runs M.
3. For each run r from 1 to M
   a. Perform BPE merging:
      i. Identify the token pair (a, b) with the highest frequency
      ii. Merge the token pair (a, b) into a new token t
      iii. Update the vocabulary V by adding the new token t
      iv. Update the priority queue P with the updated frequencies of token pairs in V
4. Analyze resulting vocabulary (V):
   - Identify any remaining specialized tokens that are not adequately captured
   - Consider extending the vocabulary further if necessary
5. Repeat steps 3 and 4 for the desired number of runs (M)

Output: dynamic vocabulary that captures both common sub word patterns and specialized patterns in the text data

Instead of a single pass of merging, MR-BPE iteratively merges pairs of characters in the vocabulary over multiple runs (M). This iterative merging process allows MR-BPE to efficiently create a dynamic vocabulary that captures both common subword patterns and specialized patterns in the text data, thereby enhancing the model's ability to understand and represent the multimodal context of text-emoji pairs. This results in improved sentiment analysis performance, efficient tokenization, and better handling of uncommon words particularly in the setting of social media, where language and sentiment expression are extremely varied and dynamic.

3.1.2. Concat-ViLT

To capture the interdependencies and relationships between text and emojis and enhance the model's ability to understand the multimodal context of text-emoji pairs, a novel Concat-ViLT is used. The flow diagram of the Concat-ViLT is illustrated in the following fig. 2.

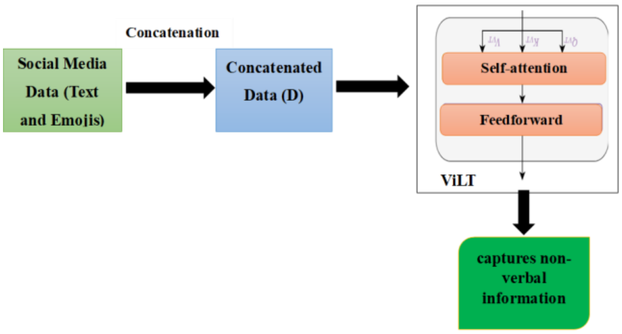

Fig. 2. Flow diagram of Concat-ViLT

The input data contains a sequence of both emoji and text tokens, initially concatenating the text and emoji sequences together to form a single input sequence. Let's denote the input text as $T = \{t_1, t_2, \dots, t_n\}$ and the corresponding emoji as $E = \{e_1, e_2, \dots, e_m\}$. The concatenated input sequence is denoted as D, which is expressed in the following equation (2)

$$D = T \oplus E \qquad (2)$$

The concatenated input sequence D captures the interdependencies and relationships between the text and emojis. This concatenated representation allows the model to consider both (emoji and textual data) modalities simultaneously during pre-training. An embedding layer is used to embed each token in a high-dimensional vector space.

The concatenated embeddings are then passed through the ViLT architecture. ViLT is a specialized pre-training framework capable of jointly processing and understanding emoji and textual information within the same embedding space. ViLT integrates text and visual data (emojis) into a shared embedding space, allowing it to capture non-verbal cues and their interaction with the text. ViLT is composed of feedforward neural networks and self-attention processes. Self-attention techniques in the transformer layer enable the model to assess the importance of various tokens in the sequence, including text-emoji interactions. The self-attention mechanism is defined by the following equation (3)

$$A(Q,K,V) = softmax(\frac{QK^T}{\sqrt{dk}})V \quad (3)$$

After the attention mechanism, each token's representation proceeds through a feed-forward network. This network is made up of one or more fully linked layers with activation functions, which enable the model to represent complicated interactions between tokens. By this process, it captures non-verbal cues conveyed by emojis and their interaction with the text. Concat-ViLT learns to represent both textual and visual components (emojis) cohesively through pre-training on concatenated text-emoji pairs, which improves the model's capacity to comprehend the subtle meaning communicated by text-emoji combinations.

Multi-Token Concatenated Embedding helps the model understand the multimodal context of text-emoji pairs, resulting in more accurate sentiment analysis. By resolving contextual mismatch generated by opposing text and emoji pairs, the technique efficiently recognizes and handles instances of sarcasm or fraud.

3.2. Semantic classification and prediction

The SemantiXpert Probabilistic Classifier is an innovative technique that makes use of powerful semantic analysis approaches to minimize misclassification rates and increase sentiment prediction based fake news detection accuracy. To process the complex interaction between text and emojis, comprehend user intentions, and precisely forecast sentiment polarity, it combines a Polarized Probabilistic classifier with a Contrastive Semantic Clause Filter Network.

3.2.1. Contrastive Semantic Clause Filter Network

A new Contrastive Semantic Clause Filter Network is introduced to understand the domain and user intention by capturing the semantic relationships within user comments and separating them into meaningful clauses. The architectural diagram of Contrastive Semantic Clause Filter Network is shown in the following fig. 3.

Initially, CLIP uses a contrastive learning objective to align emoji and text representations in the same embedding space. CLIP helps the SemantiXpert Probabilistic Classifier's ability to capture and represent the combined semantics of sentences and emojis seen in user comments. This is critical for comprehending the relationships between distinct entities, actions, and modifiers, and is especially useful for sentiment analysis, such as identifying sarcasm and complex sentiments in the news content.

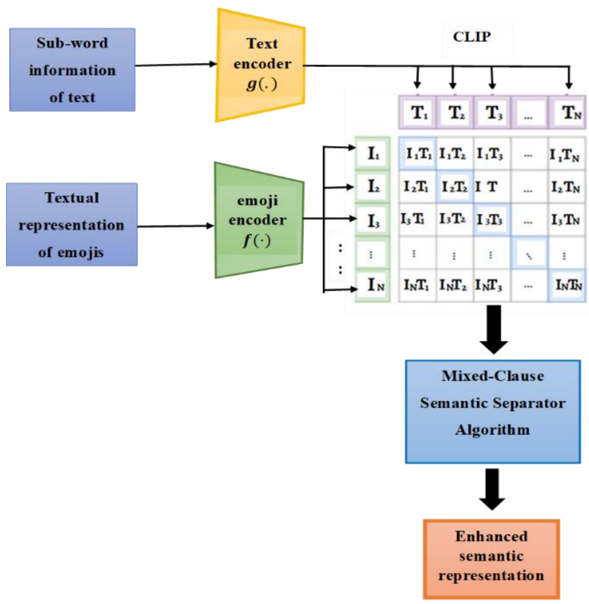

Fig. 3. Architectural diagram of Contrastive Semantic Clause Filter Network

The CLIP model accepts as input both textual representations of comments (after tokenization and processing) and textual representations of emojis collected during the pre-processing phases. CLIP uses a dual-stream approach with emoji encoders $f(\cdot)$ and text $g(\cdot)$. The shared encoder $h(.)$ processes both emoji and text inputs through the same transformer layers. CLIP model maps each textual input x_i^T and emoji input $x_i^I$to a shared embedding space, enabling it to understand the semantic relationship between the two modalities (where similar semantic impressions are represented by nearby vectors and unrelated impressions are farther apart). Once the inputs are embedded into the shared space, the network calculates the similarity between text-emoji pairs. The cosine similarity between a text embedding $g(x_j^T)$ and an emoji embedding $f(x_i^I)$ is defined as in the following equation (4)

$$s_{i,j} = h\left(f(x_i^I)\right)^T h\left(g(x_j^T)\right) \quad (4)$$

CLIP's training objective involves maximizing the similarity between corresponding pairs of text and image embeddings while lessening the similarity between mismatched pairs. Given a batch of N text-emoji pairs $\{x_i^I, x_j^T\}_{i=1}^N$, the goal is to maximize the similarity for matching pairs and minimize it for non-matching pairs. The total contrastive loss $\mathcal{L}$ of the mini-batch is represented as the sum of emoji-to-text and text-to-emoji losses, which is expressed as follows in equation (5):

$$\mathcal{L} = -\frac{1}{2}\sum_{1=1}^{N}(log\frac{\exp(s_{i,j})/\tau}{\sum_j \exp(s_{i,j})/\tau} + log\frac{\exp(s_{i,j})/\tau}{\sum_j \exp(s_{i,j})/\tau}) \quad (5)$$

Where N is the number of training samples, $s_{i,j}$ denotes the cosine similarity between two embeddings, and τ is a temperature parameter that controls the scale of the similarity scores. It learns to maximize the similarity between positive pairs (e.g., text and corresponding emoji) and minimize the similarity between negative pairs (e.g., text and unrelated emoji). The CLIP model captures the combined semantics of the sentences in user comments. This representation includes the meanings and relationships between different parts of the text. By understanding the semantic content of user comments, CLIP contributes to the accurate identification of sentiment and domain-specific intentions. Additionally, CLIP's contrastive learning framework aids in fine-tuning semantic representations, ensuring that the model effectively distinguishes between different aspects of user comments.

After obtaining the semantic representations from CLIP, a Mixed-Clause Semantic Separator Algorithm is employed in the final layer of the CLIP. This algorithm is designed to improve semantic analysis by accurately splitting user comments into clauses based on semantic relationships captured by the CLIP model. This distinction is critical for figuring out the relationships between different portions of the text and boosting the precision of semantic analyses.

Common-sense reasoning: The algorithm starts by utilizing external information bases and common-sense reasoning to improve comprehension of the relationships and context inside the comment. This stage improves the text's semantic understanding.

Clause separation: Then this algorithm examines the user comment's content for linguistic clues such as punctuation marks (e.g., periods, commas), conjunctions (e.g., "and," "but"), and sentence structure. These cues are used to determine possible boundaries between clauses in a user comment. Based on the identified potential boundaries, the algorithm segments the user comment into preliminary clauses. Semantic consistency is maintained within each segment to ensure that each clause represents a coherent unit of meaning. The extraction of clauses is important for understanding the relationships between entities, actions, and modifiers within the user comments. The algorithm outputs a set of clauses, each containing a coherent semantic unit extracted from the user comments. These clauses are then used for further analysis, such as polarity prediction.

By splitting user comments into clauses, the algorithm enables a more granular analysis of the semantic content. Each clause represents a distinct unit of meaning within the text, facilitating a deeper understanding of the relationships between different components of the input. By effectively identifying and delineating clauses within user comments, the algorithm enables a deeper understanding of the underlying semantics, contributing to improved polarity prediction and overall model performance in identifying fake news.

3.2.2. Polarized Probabilistic classifier

To increase the accuracy of semantic categorization and polarity prediction, a Polarized Probabilistic classifier is presented. In this classifier, initially, an ELECTRA model is used to determine polarity in comments by discriminating between important and less important semantic representations in the relevant semantic sentences.

These networks are designed to accurately classify token replacements based on the semantic output of the Contrastive Semantic Clause Filter Network. The ELECTRA model consists of two neural networks, a generator and a discriminator. The generator acts similarly to a masked language model (MLM). Both parts use transformer-based encoding networks to generate vector representations of input word sequences. However, in this adaptation, the discriminator takes on a larger role in finding relevant semantic representations for sentiment analysis. The architectural diagram of the ELECTRA is shown in the following fig. 4.

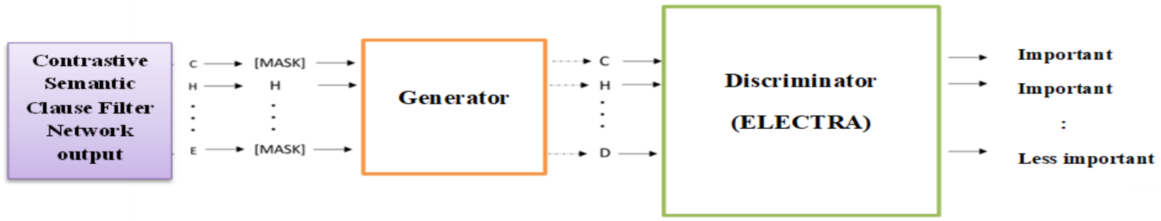

Fig. 4. Architectural diagram of ELECTRA

In its operation, the generator is tasked with predicting the original tokens from sequences where certain tokens have been masked out. The generator uses a SoftMax layer to determine the probability of creating a specific token, $x_t = [MASK]$, for every given location t, where x_t is a masked token. This is presented in the following equation (6):

$$P_G(x_t|x) = \exp\left(e(x_t)^T h_G(x)_t / \sum_{x'} \exp(e(x')^T h_G(x)_t\right) \quad (6)$$

Where e donates token embeddings. The discriminator uses a sigmoid output layer to determine if a token is from the input data (original) or a replacement from the generator distribution (less important semantic representation), which is expressed in equation (7)

$$D(x,t) = sigmoid(w^T h_D(x)_t \quad (7)$$

Here, $h_D(x)_t$ represents the contextualized vector representation of the token x_t produced by the discriminator, and w is the weight vector. In the discriminator, a new pre-training task known as the token replacement mechanism is employed, as a novelty. For a given original input, $x = [x_1, \ldots, x_n]$, MLM-selected random tokens in the input sequence are masked, and the generator predicts the original tokens in their place. The predicted tokens are used to replace certain tokens in the input sequence. This produced the vector $m = [m_1, \ldots, m_k]$. The residues at these locations are substituted with a [MASK] token, which is represented as follows in equation (8)

$$x^{masked} = Replace(x, m, [MASK]) \quad (8)$$

The discriminator predicts whether a word at each position of the input sequence has been replaced by the generator, classifying it as original (important) or replaced (less important semantic representations). The ELECTRA model uses a token replacement mechanism to differentiate between important and less important semantic representations in semantic clauses. The ELECTRA model uses the output of the Discriminator Network to determine the polarity of the text-emoji pairs. The polarization score P_i for the i^{th} text-emoji pair is calculated as the probability of the token being the original token in equation (9)

$$P_i = D(x_i, t) \quad (9)$$

It assigns higher polarity scores to important tokens and lower polarity scores to less important tokens. ELECTRA employs its learned representations to assign polarity scores to tokens and phrases, identifying agree, disagree, or unrelated sentiments.

Finally, the Naive Bayes classifier is utilized, which uses semantic relationships from the Contrastive Semantic Clause Filter Network and polarization scores from the ELECTRA model as input features. It is assumed that the presence of any feature (semantic linkages and polarisation scores) is unrelated to the presence of any other characteristic within the class (sentiment). Next, given the input features the classifier applies the Bayes theorem to calculate the posterior probability of each sentiment class (A). The steps involved in the Naive Bayes classification are as follows:

1. Compute the prior probabilities $P(A)$ for each sentiment class (e.g., agree, disagree, or unrelated) based on the training data.
2. Calculate the likelihood $P(S, P_i|A)$ by assuming independence between the semantic features S and polarization scores P_i given the sentiment class A: in equation (10)

$$P(S, P_i|A) = P(S|A)(P_i|A) \quad (10)$$

3. Compute the marginal probability $P(S, P_i)$ by summing the joint probabilities over all possible sentiment classes: in equation (11)

$$P(S, P_i) = \sum_C P(S, P_i|A)P(A) \quad (11)$$

4. Apply the Naive Bayes formula to compute the posterior probabilities $P(A|S, P_i)$ for each sentiment class: in equation (12)

$$P(A|S, P_i) = \frac{P(S,P_i|A)P(A)}{P(S,P_i)} \quad (12)$$

5. The Naive Bayes classifier then selects the sentiment class A that has the highest posterior probability $P(A|S, P_i)$ as the final sentiment prediction.

The predicted emotion is then determined by choosing the class with the highest posterior probability. The combination of the semantic relationship and the polarization scores allows the Naive Bayes classifier to capture both the contextual meaning and the sentiment-specific information, which improves the overall accuracy of the sentiment classification task for fake news detection.

By tackling the intricacies of text-emoji data and enhancing semantic and polarity understanding, the SemantiXpert Probabilistic Classifier decreases misclassifications and improves overall sentiment prediction accuracy in social media fake news detection scenarios. This holistic strategy ensures that the model manages complexities of human communication, such as sarcasm and mixed sentiments, resulting in more accurate and informative sentiment analysis results for fake news detection.

4. Result and Discussion

This section contains a full description of the implementation findings and the proposed system's performance, as well as a comparison section to ensure that the proposed approach is appropriate for social media sentimental analysis for false news detection.

4.1. Dataset description

In this proposed model Fake News Detection (XLMRoberta) dataset is used. A collection of labeled samples is used to train and evaluate algorithms for false news detection tasks. It includes a wide range of textual data, such as news stories, social media postings, and internet material, gathered from a variety of sources and domains. Each sample in the collection is labeled with a binary classification that indicates whether the content is real or fraudulent. In this research, 1000 samples are taken. The dataset is divided into training and testing sets, with 90% of the samples used for training and 10% for testing. The data set link is https://huggingface.co/Sajib-006/fake_news_detection_xlmRoberta.

4.2. Experimental setup

The proposed system is simulated in Python, and this section includes an extensive discussion of the implementation results and performance, as well as a comparison section to confirm that the proposed setup functions properly.

Software :Python

OS :Windows 10 (64-bit)

Processor:Intel i5

RAM :8GB RAM

4.3. Performance evaluation of the proposed model

This section provides a comprehensive assessment of the experimental findings. The proposed social media sentiment analysis for the fake news detection technique's performance was evaluated by calculating total recall, precision, f-score, sensitivity, specificity, and accuracy.

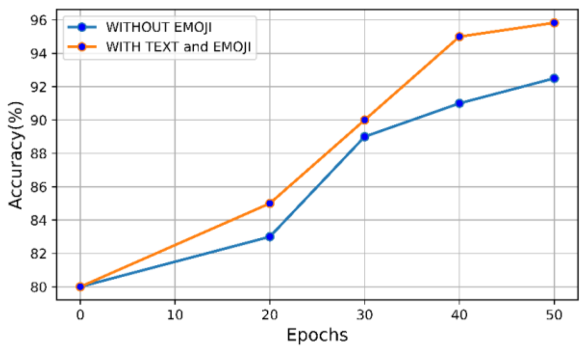

Fig. 5. Accuracy of the proposed model

Fig. 5. shows the proposed model's accuracy without emoji as well as with text and emoji. The proposed method obtains the greatest accuracy of 95.83% with text and emoji, as well as 92.5% without emoji, at epoch 50. When the epoch is 1, the proposed approach achieves a low accuracy of 80% for both text and emoji. MR-BPE improves accuracy by better-catching text subtleties, It helps the model to better represent words and phrases, resulting in more accurate sentiment analysis.

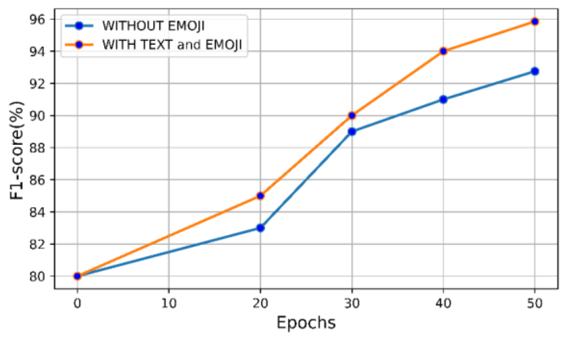

Fig. 6. F1-score of the proposed model

Fig. 6. demonstrates the F1-score of the proposed model with text and emoji and without emoji. The proposed model achieves an extreme F1-score value of 95.85% with text and emoji and 92.75% without emoji when the epoch is 50. By lowering misclassifications, the classifier increases its F1 score by generating more accurate predictions about the sentiment of social media posts.

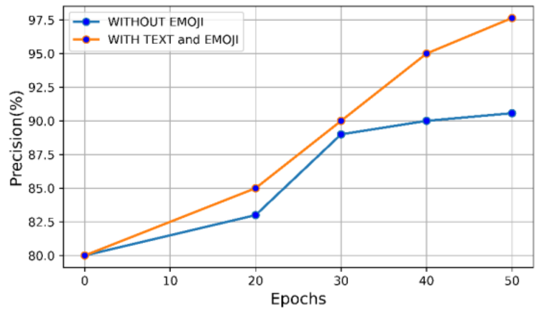

Fig. 7. Precision of the proposed model

The above fig. 7. depicts the proposed model's precision with text and emoji and without emoji. From the figure, it is illustrated that when the epoch is 50, the proposed model reaches a maximum precision of 97.64% without emoji, and also achieves a precision of 90.57% with text and emoji. When the epoch is 1, the proposed model achieves a minimum precision of 80% in both types. By accurately capturing the semantic relationships and domain intentions, the clause separator algorithm model makes fewer false positive predictions, thus improving precision.

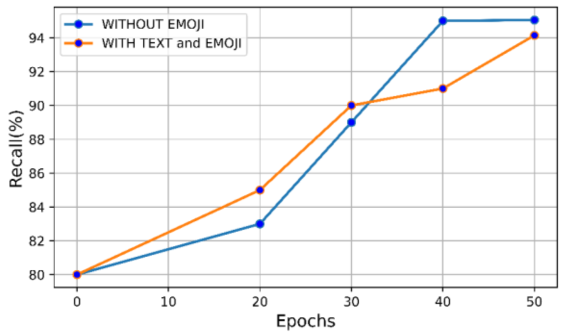

Fig. 8. Recall of the proposed model

Fig. 8. shows the recall of the suggested model. The proposed model achieves a low recall value of 80% when the epoch value is low. Also achieves a maximum recall value of 94.14% with text and emoji, and 95.05% without emoji, when the epoch is 50. The Contrastive Semantic Clause Filter Network improves in extracting more relevant information from user comments, resulting in increased recall by eliminating false negatives.

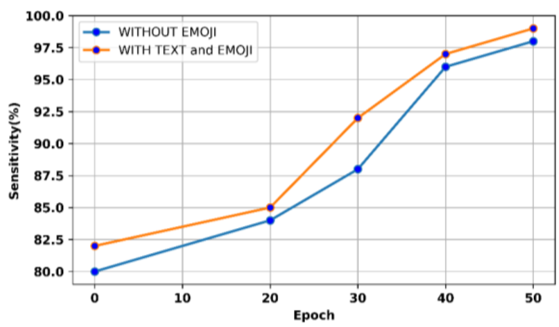

Fig. 9. Sensitivity of the proposed model

The fig. 9. above shows the proposed model's sensitivity performance with and without text and emoji. When the epoch is 50, the proposed approach has a maximum specificity value of 97% without emoji and 98% with text and emoji. When the epoch is low, the proposed model achieves low sensitivity. The Concat-ViLT Model captures the interdependencies between text and emojis, which improves the model's sensitivity to recognize subtle details found in social media data.

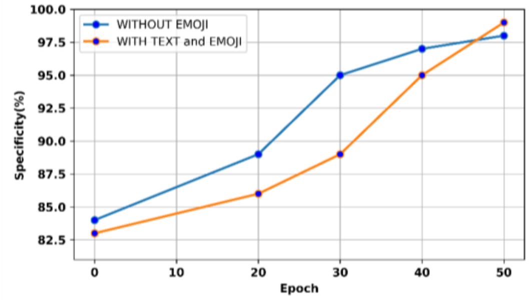

Fig. 10. Specificity of the proposed model

The proposed model's specificity is demonstrated in the fig. 10. The proposed approach obtains low specificity values of 83% and 82.6% with text and emoji and without emoji, respectively, when the epoch value is low. At 50 epochs, it also reaches a maximum specificity value of 97% without emoji, and 98% with text and emoji. Higher specificity is a result of improved comprehension of semantic relationships and the polarization of significant semantic representations.

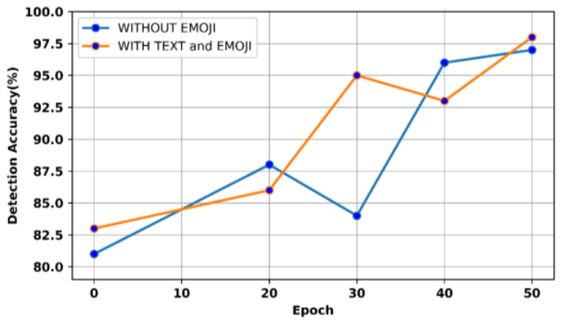

Fig. 11. Detection accuracy of the proposed model

Fig. 11. demonstrates that the proposed model's detection accuracy improves with more training epochs and that including emojis with text data significantly enhances sentiment detection accuracy. At 50 epochs, the model achieves near-optimal performance with detection accuracies of about 97.2% for without emoji data and slightly below 98% for with text and emoji data. By reducing misclassifications and improving understanding of the relationships within comments, Navie Bayes classifier enhance detection accuracy.

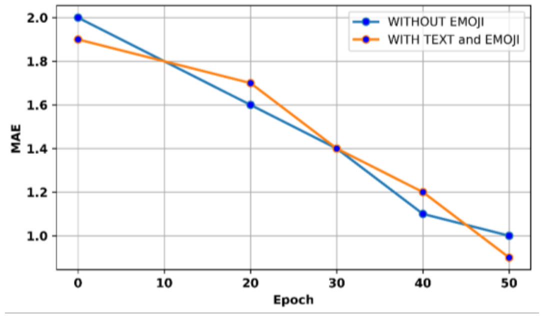

Fig. 12. Mean average error of the proposed model

The mean average error of the proposed model is illustrated in the fig. 12. From the figure it is observed that when the epoch is high like 50 the proposed model achieves a low MAE of 1 and 0.5 without emoji and with text and emoji. More accurate sentiment predictions in fake news detection are produced by combining enhanced tokenization, efficient multimodal representation, and precise semantic comprehension. This results in a decrease in the MAE between predicted and real sentiment values.

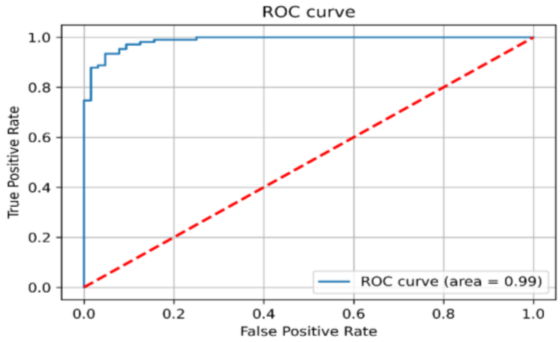

Fig. 13. ROC of the proposed model

Fig. 13. displays the ROC curve of the proposed model, showing the true positive rate versus the false positive rate. The curve demonstrates high performance with an area under the curve (AUC) of 0.99. The SemantiXpert Probabilistic Classifier's improved semantic comprehension has the potential to improve the ROC curve by more effectively distinguishing between instances of positive and negative sentiment.

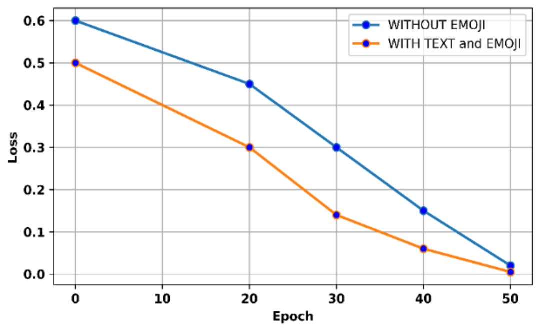

Fig. 14. Loss of the proposed model

Fig. 14. depicts the loss of the proposed model. It has been noted that when the epoch is large, such as 50, the proposed approach achieves a low MAE of 0.1 and 0.09, respectively,

without and with text and emoji. Concat-ViLT lowers loss by introducing visual information into the sentiment analysis process by combining text and emoji representations in the same embedding space.

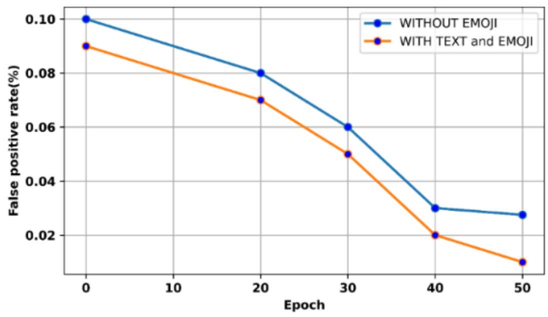

Fig. 15. False positive rate of the proposed model

The proposed model's false positive rate is demonstrated in Fig. 15. The proposed approach obtains a low false positive rate of 0.01% and 0.03% with text and emoji and without emoji, respectively, when the epoch value is as high as 50. At low epochs, it also reaches a maximum false positive rate value of 0.09% with text and emoji, and 0.1% without emoji. The contrastive semantic clause filter network's capacity to understand actual links between domain and user intent aids in the reduction of false positives by effectively capturing the semantics of user comments

4.4. Comparative study of the proposed model

This section highlights the proposed Multi-Token Concatenated Embedding and SemantiXpert Probabilistic Classifier social media sentiment analysis for fake news detection model with the traditional models and the achieved result was clarified in detail in this section by comparing it with ANN [26], CNN [26], DNN [26], and LRA-DNN [26] and existing classifiers such as decision tree, Naive Bayes, Logistic Regression, and KNN showing their results based on various metrics such as accuracy, recall, precision, false positive rate, sensitivity, specificity, error rate, ROC-AUC, and Mean Squared Error (MSE). and F1-score.

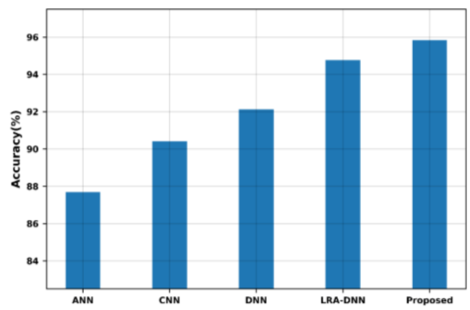

Fig. 16. Comparison of accuracy of the proposed model

Fig. 16. illustrates the comparison of the accuracy of the proposed model with existing models. The existing models such as ANN, CNN, DNN, and LRA-DNN achieve an accuracy value of 87.69%, 90.42%, 92.13%, and 94.77%. Compared with existing models the proposed model attains the highest accuracy value of 95.83%. This shows that the suggested method is more effective at accurately identifying sentiment.

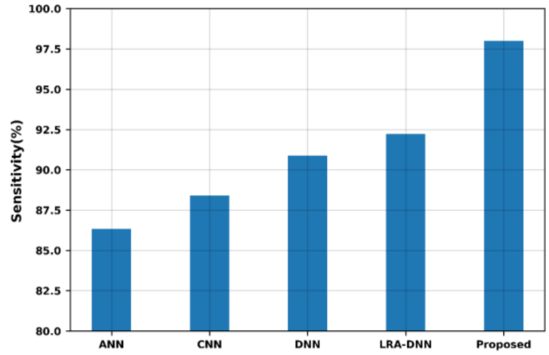

Fig. 17. Comparison of the sensitivity of the proposed model.

Fig. 17. compares the sensitivity of the suggested model with existing models. The existing models such as ANN, CNN, DNN, and LRA-DNN attain a sensitivity of 86.33%, 88.41%, 90.89%, and 92.23%. Compared with existing models the proposed model achieves the highest sensitivity of 98%.

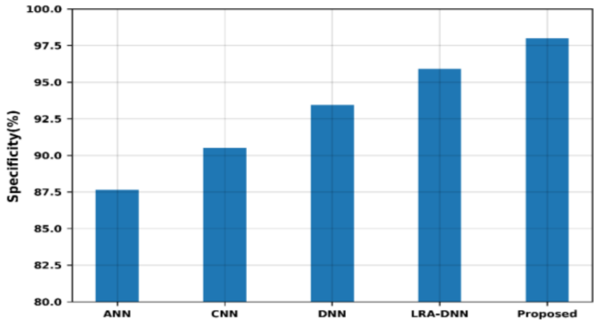

Fig. 18. Comparison of specificity

Fig. 18. displays a comparison of the proposed model's specificity to existing models. Existing models, such as ANN CNN DNN and LRA-DNN, obtain specificity values of 87.64%, 90.52%, 93.45%, and 95.91%, respectively, however, the suggested model has a better specificity value of 98%. Compared with conventional models, the proposed one has a greater specificity value.

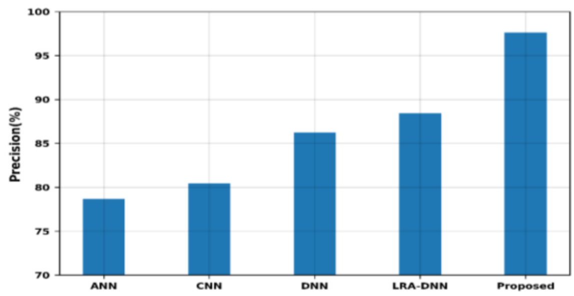

Fig. 19. Comparison of precision

A comparison of the proposed models' precision value with existing models is shown in Fig. 19. compared with existing models the proposed model achieves the highest precision value of 97.64%, whereas existing models such as ANN, CNN, DNN, and LRA-DNN achieve a precision value of 78.67%, 80.44%, 86.24%, and 88.45% respectively. The proposed model's higher precision indicates that it has a lower false positive rate, making it more reliable in correctly identifying positive sentiments.

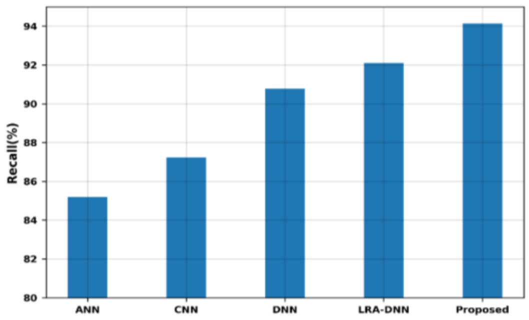

Fig. 20. Comparison of recall of the proposed model

The recall value of the proposed model is compared with existing models, which is shown in fig. 20. The proposed model achieves a high recall value of 94.14%, whereas existing models such as ANN, CNN, DNN, and LRA-DNN attain a recall value of 85.195, 87.24%, 90.78%, and 92.11% respectively. Compared with existing models the proposed model achieves the highest recall value.

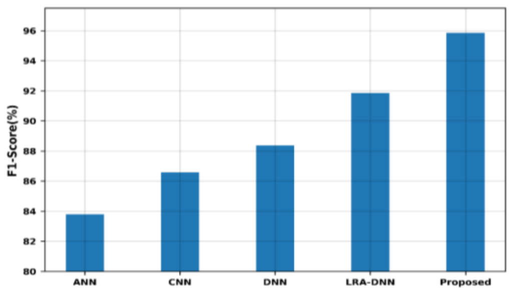

Fig. 21. Comparison of F1-score

The F1-score of the proposed model is compared with the existing models, which is depicted in Fig. 21. Various existing models such as ANN, CNN, DNN, and LRA-DNN achieve an F1-score value of 83.7%, 86.8%, 88.4%, and 91.9%, whereas the proposed model achieves an F1-score value of 95.85%. Compared to all existing approaches, the proposed one has the greatest F1-score value.

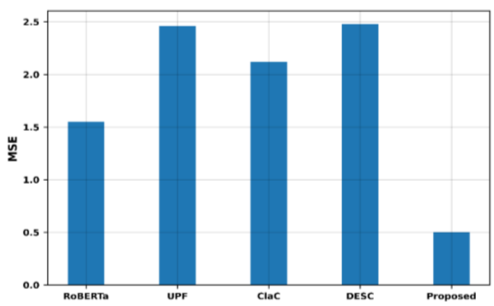

Fig. 22. Comparison of MSE of the proposed model with existing models

The above fig. 22. depicts the MSE comparison of the proposed model with existing models [27]. the existing models such as RoBERTa, UPF, CloC, and DESC achieve a MSE value of 1.6, 2.4, 2.17, and 2.4 respectively. Lower MSE values indicate better model performance. The proposed model achieves a low MSE value of 0.5, demonstrating its ability to accurately predict sentiment polarity.

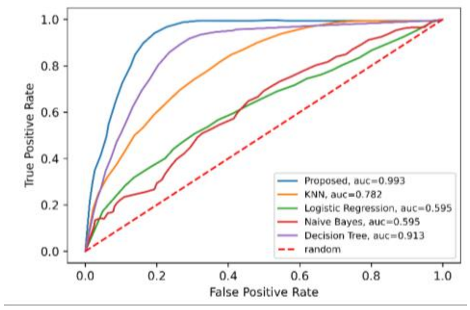

Fig. 23. ROC comparison of the proposed model with existing classifiers

ROC comparison of the proposed model with existing classifiers is illustrated in fig. 23. [28]. The existing classifiers such as decision tree, Naive Bayes, Logistic Regression, and KNN achieve an AUC value of 0.913, 0.595, 0.595, and 0.785. The proposed classifier achieves a significantly higher AUC value of 0.993 compared to existing classifiers, indicating its superior ability to distinguish between positive and negative sentiments.

The proposed model significantly outperforms existing models across various metrics, including accuracy, precision, recall, F1-score, specificity, MSE, and ROC-AUC. The novel pre-processing techniques and advanced semantic classification contribute to these improvements, making the proposed model a robust solution for sentiment analysis, particularly in handling the complexities of sarcasm and emoji usage in social media data for fake news detection.

5. Conclusion

This research concludes by presenting a thorough strategy for detecting fake news on social media platforms that combines advanced pre-processing and classification techniques with sentiment analysis approaches. The novel pre-processing model, which includes MR-BPE tokenization and Concat-ViLT, effectively captures the multimodal context of text-emoji pairs, thereby improving the recognition of sarcasm or deception. The introduction of the SemantiXpert Probabilistic Classifier further refines sentiment analysis by incorporating semantic relationships and user intentions. It combines CLIP and a Clause Separator Algorithm to break down user comments into meaningful clauses, allowing for more precise semantic analysis and fewer misclassifications. The accuracy of polarity prediction was increased by using the Polarised

Probabilistic classifier. Here, the total accuracy of sentiment prediction was increased by reducing misclassifications with the use of Naive Bayes and ELECTRA classifiers. Comparative analysis demonstrates the superior performance of the proposed model over existing methods such as ANN, CNN, DNN, and LRA-DNN, achieving the highest accuracy of 95.83%, sensitivity, and AUC values of 0.993 while also minimizing the MAE value of 0.5. The inclusion of emojis provides additional sentiment cues, leading to more accurate detection of fake news than without emojis. Overall, the proposed approach significantly improves sentiment analysis in social media by offering a more complex comprehension of text-emoji interactions and enhancing the ability to detect fake news by identifying polarity and sarcasm.

Reference

[1] M. Swartz and A. Crooks, "Comparison of emoji use in names, profiles, and tweets," *In 2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pp. 375-380, 2020, February. IEEE.

[2] Alrumaih, A. Al-Sabbagh, R. Alsabah, H. Kharrufa and J. Baldwin, "Sentiment analysis of comments in social media," International Journal of Electrical & Computer Engineering, vol. 10, no. 6, pp. 2088-8708, 2020.

[3] Debnath, N. Pinnaparaju, M. Shrivastava, V. Varma and I. Augenstein, "Semantic textual similarity of sentences with emojis," In Companion Proceedings of the Web Conference 2020, pp. 426-430, 2020, April.

[4] Surikov and E. Egorova, "Alternative method sentiment analysis using emojis and emoticons," Procedia Computer Science, vol. 178, pp. 182-193, 2020.

[5] S. Gupta, O. Garg, R. Mehrotra and A. Singh, "Social media anatomy of text and emoji in expressions," *In Smart Computing Techniques and Applications: Proceedings of the Fourth International Conference on Smart Computing and Informatics*, vol. 2, pp. 41-49, 2021. Springer Singapore.

[6] M. Krommyda, A. Rigos, K. Bouklas and A. Amditis, "An experimental analysis of data annotation methodologies for emotion detection in short text posted on social media," *In Informatics*, vol. 8, no. 1, pp. 19, 2021, March. MDPI.

[7] F. Benrouba and R. Boudour, "Emotional sentiment analysis of social media content for mental health safety," *Social Network Analysis and Mining*, vol. 13, no. 1, pp. 17, 2023.

[8] S. Al-Azani and E.S.M. El-Alfy, "Early and late fusion of emojis and text to enhance opinion mining," *IEEE Access*, vol. 9, pp. 121031-121045, 2021.

[9] 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, pp. 1-20, 2022.

[10] T.P. Kumar and B.V. Vardhan, "A Pragmatic Approach to Emoji based Multimodal Sentiment Analysis using Deep Neural Networks," *Journal of Algebraic Statistics*, vol. 13, no. 1, pp. 473-482, 2022.

[11] Z. Ahanin and M.A. Ismail, "Feature extraction based on fuzzy clustering and emoji embeddings for emotion classification," *International Journal of Technology Management and Information System*, vol. 2, no. 1, pp. 102-112, 2020.

A. Yoo and J.T. Rayz, "Understanding emojis for sentiment analysis," *In The International FLAIRS Conference Proceedings*, vol. 34, 2021, April.

[12] M.A. Ullah, S.M. Marium, S.A. Begum and N.S. Dipa, "An algorithm and method for sentiment analysis using the text and emoticon," *ICT Express*, vol. 6, no. 4, pp. 357-360, 2020.

[13] L. Holthoff, "The Emoji Sentiment Lexicon: Analysing Consumer Emotions in Social Media Communication," *In Proceedings of the 49th European Marketing Academy (EMAC) Annual Conference*, 2020, May.

[14] S. Gupta, R. Singh and V. Singla, "Emoticon and text sarcasm detection in sentiment analysis," *In First International Conference on Sustainable Technologies for Computational Intelligence: Proceedings of ICTSCI 2019*, pp. 1-10, 2020. Springer Singapore.

[15] A.S. Talaat, "Sentiment analysis classification system using hybrid BERT models," *Journal of Big Data*, vol. 10, no. 1, pp. 1-18, 2023.

[16] L. Bryan-Smith, J. Godsall, F. George, K. Egode, N. Dethlefs and D. Parsons, "Real-time social media sentiment analysis for rapid impact assessment of floods," *Computers & Geosciences*, vol. 178, pp. 105405, 2023.

[17] Liu, F. Fang, X. Lin, T. Cai, X. Tan, J. Liu and X. Lu, "Improving sentiment analysis accuracy with emoji embedding," Journal of Safety Science and Resilience, vol. 2, no. 4, pp. 246-252, 2021.

[18] M. Shahzad, C. Freeman, M. Rahimi and H. Alhoori, "Predicting Facebook sentiments towards research," *Natural Language Processing Journal*, vol. 3, pp. 100010, 2023.

[19] S. Gupta, A. Singh and V. Kumar, "Emoji, Text, and Sentiment Polarity Detection Using Natural Language Processing," *Information*, vol. 14, no. 4, pp. 222, 2023.

[20] Alsayat, "Improving sentiment analysis for social media applications using an ensemble deep learning language model," Arabian Journal for Science and Engineering, vol. 47, no. 2, pp. 2499-2511, 2022.

[21] A.R. Pathak, M. Pandey and S. Rautaray, "Topic-level sentiment analysis of social media data using deep

learning," *Applied Soft Computing*, vol. 108, pp. 107440, 2021.

[22] B.S.W. Bagus Satria Wiguna, C.V.H. Cinthia Vairra Hudiyanti, A.A.R. Alqis Alqis Rausanfita and A.Z.A. Agus Zainal Arifin, "Sarcasm detection engine for twitter sentiment analysis using textual and emoji feature," *Jurnal Ilmu Komputer dan Informasi*, vol. 14, no. 1, pp. 1-8, 2021.

[23] Q.A. Xu, C. Jayne and V. Chang, "An emoji feature-incorporated multi-view deep learning for explainable sentiment classification of social media reviews," *Technological Forecasting and Social Change*, vol. 202, pp. 123326, 2024.

[24] K.E. Naresh Kumar and V. Uma, "Intelligent sentinet-based lexicon for context-aware sentiment analysis: optimized neural network for sentiment classification on social media," *The Journal of Supercomputing*, vol. 77, no. 11, pp. 12801-12825, 2021.

[25] N. Shelke, S. Chaudhury, S. Chakrabarti, S.L. Bangare, G. Yogapriya and P. Pandey, "An efficient way of text-based emotion analysis from social media using LRA-DNN," *Neuroscience Informatics*, vol. 2, no. 3, pp. 100048, 2022.

[26] R.A. Potamias, G. Siolas and A.G. Stafylopatis, "A transformer-based approach to irony and sarcasm detection," *Neural Computing and Applications*, vol. 32, no. 23, pp. 17309-17320, 2020.

[27] M. Asif, A. Ishtiaq, H. Ahmad, H. Aljuaid and J. Shah, "Sentiment analysis of extremism in social media from textual information," *Telematics and Informatics*, vol. 48, pp. 101345, 2020.