Enhancing Misinformation Detection on Twitter with a Content-Based Multi-Lingual Bert Model

Krishna Kumar*, Dr. Akila Venkatesan
Computer Science and Engineering, Puducherry Technological University, Puducherry, India

***Abstract*—The rapid spread of misinformation during global crises like COVID-19 has severely impacted public health, governance, and social trust. Social media platforms such as Twitter have amplified this issue, underscoring the urgent need for multilingual, real-time misinformation detection. The proposed Content-based Attention Multi-lingual BERT (CA-BERT) model addresses this challenge by enhancing the standard BERT framework with a content-based attention mechanism that assigns adaptive weights to semantically important tokens often linked to false or misleading content. This attention enables deeper contextual understanding of misinformation cues across diverse linguistic contexts. Using the LIME interpretability method, CA-BERT provides transparent explanations of its predictions, supporting accountable decision-making for policymakers and content moderators. Leveraging multilingual BERT (mBERT) allows the model to handle multiple languages simultaneously, ensuring robust cross-lingual applicability. Evaluations using a balanced multilingual tweet dataset on COVID-19 topics demonstrate that CA-BERT outperforms baseline models such as RoBERTa, DANN, and HANN, achieving 96% recall for true information and 95% for misinformation in English, with F1 Scores of 93% and 92%, respectively. The model maintains strong cross-lingual generalization, especially for Dutch (75% F1) and Spanish (72% F1), with slightly lower performance for Arabic due to tokenization and dialectal complexity. These results highlight CA-BERT's adaptability while underscoring the need for improved handling of low-resource, morphologically rich languages. Future work involves region-specific preprocessing, cross-lingual transfer learning, and multimodal misinformation detection, aiming to transform CA-BERT into a core component of multilingual real-time disinformation monitoring systems.**

Keywords*—Component; misinformation detection; multi-lingual BERT; content-based attention mechanism; syntactic-semantic similarity; explainable AI; LIME interpretability; COVID-19 misinformation; cross-lingual generalization; twitter; adversarial robustness

I. Introduction

The proliferation of inaccurate or deceptive information, especially during the COVID-19 pandemic, poses a significant challenge in the current digital landscape [1]. Prominent social media platforms such as Twitter, Facebook, and Instagram have become channels for disseminating information, enabling rapid sharing while also amplifying the spread of misinformation [2]. Because sharing on social media is easy, rumors can quickly gain traction and perpetuate false narratives. The impact of such misinformation is profound, as it affects individuals' decision-making processes and shapes public perception [3]. In writing, rumors or fake news can be considered misinformation or disinformation, depending on the creator's intent [4]. In this context, misinformation refers to the unintentional spread of incorrect information, while disinformation refers to the deliberate dissemination of false information for deceptive purposes [5]. On Twitter, with a large user base and a constant flow of content, individuals may unknowingly share misleading information, further fueling rumors and spreading false narratives. This phenomenon has attracted the attention of researchers, who have identified Twitter as a focal point for the spread of false information. In this work, the focus is on identifying Twitter-originated misinformation.

Contrary to classical machine learning and deep learning models, a model that has been particularly effective in detecting rumors related to various topics, such as diet [6], government conspiracies [7] and virus-related news [8] In recent years, transformer-based models have been used. The transformer model, known as Bidirectional Encoder Representations from Transformers (BERT), was developed by Google researchers. [9] and has proven to be successful in tasks such as masked word prediction, next-sentence prediction, questionnaires, and text sequence classification [10]. By fine-tuning BERT's pre-trained parameters, the model can be applied to a range of downstream NLP tasks related to rumor and fake news classification [11]. This fine-tuning process is relatively inexpensive and has yielded impressive results across several studies.

Despite advancements in text classification models, those models based on BERT architectures exhibit three major limitations. Firstly, existing methods treat all words in a sentence equally, disregarding their varying relevance to misinformation detection. Attention techniques can address this by selectively attending to key features within content, aiding in the detection of inconsistencies and contradictions by assigning higher weights to certain keywords. Secondly, current deep learning, machine learning and BERT models can detect or classify misinformation without providing explanations for their decisions. Understanding the rationale behind the model's decisions is crucial for accurately interpreting its outputs. Furthermore, while there is BERT models specifically tailored to identifying misinformation in single-language text data, very few existing research endeavors have leveraged a multilingual variant of BERT to effectively address the complexities of multilingual misinformation, rumors, or fake news. The primary contribution of the proposed work is as follows:

- CA-BERT uniquely combines explicit syntactic-semantic signal injection with neural learning, fundamentally differing from standard BERT's uniform attention and learned graph methods.

*Corresponding author.

- Ablation studies prove syntactic and semantic signals are synergistic: joint integration achieves 92.5% F1, exceeding individual contributions (88.1% + 88.6%), demonstrating complementary linguistic pattern detection.
- It also delivers superior computational efficiency: 92.5% F1 at 45 milliseconds latency (36.6% faster than graph baselines) with 19.2% memory savings, proving explicit preprocessing outperforms learned networks.
- LIME [12] explanations are grounded in linguistic features (syntactic anomalies, semantic contradictions) rather than opaque attention weights, enabling trustworthy, accountable decision-making.
- Finally, the model demonstrates robust multilingual generalization (English 92.5%, Dutch/French 75%, Spanish 72% F1) with superior cross-lingual transfer (1.3% gap versus 10.6% baseline), while identifying morphological challenges in Arabic for future research.

The rest of the work is organized as follows. Section II details the study literature. Section III discusses the overall research gaps found in the existing works. Section IV details the proposed methodology. Section V details the proposed system. Explainable AI using LIME. Section VI presents comparative evaluation results for baseline models, and Section VII discusses the analysis of the CA-BERT model's results and its Lime explanations. Finally, Section VIII concludes and discusses the future directions.

II. Related Works

A. Misinformation Detection Using Machine Learning Methods

Fig. 1 shows a taxonomy of current misinformation detection methods. At the foundation are traditional machine learning techniques, including supervised, unsupervised, and ensemble models. Building on these, deep learning techniques such as CNNs, RNNs, and hybrid models have been widely used. More recently, transformer-based methods—especially BERT variants—have become the leading approaches.

Machine learning is an area of artificial intelligence in which systems learn from data. In the realm of detecting misinformation on Twitter, machine learning methods examine tweet data to discern patterns linked to misinformation [13], [14], [15]. Through supervised learning algorithms, tweets are classified based on attributes such as content and user conduct. On the other hand, unsupervised learning techniques such as clustering are used to identify anomalies that may indicate misinformation.

Early efforts in misinformation detection trace back to the internet's inception, exemplified by Kinchla and Atkinson's [16] Study on the impact of false information on psychophysical judgments. Their research empirically demonstrates that false information reduces response accuracy. Boukouvalas et al. [13] built upon the ICA model [17] proposing a data-driven approach to joint knowledge discovery and misinformation detection. Their method creates a low-dimensional representation of tweets that accounts for spatial context, using a support vector machine (SVM) with various kernel functions, including Gaussian, RBF, and Polynomial. Ayoub et al. [14] Experimented with three machine learning algorithms (e.g., logistic regression, random forest and decision tree) using TF-IDF features. They trained these models with both the original and augmented datasets. Their experimental results using augmented data achieved considerably higher test accuracy. Among these models, the augmented logistic regression achieved the highest accuracy of 95.4% in classifying COVID-19 misinformation claims. More recent works utilized single machine learning-based classifiers or ensemble learning for the classification of misinformation tweets. Ismail et al. [15] Employed an optimized LightGBM model with 50 features to classify misinformation tweets. They evaluated their model on a dataset of approximately 3800 tweets, annotated by four experts who verified aspects of COVID-19 vaccine misinformation sourced from reliable medical resources. Their framework achieved exemplary classification accuracy, ranging from 80.1% to 92.7%, with an average area under the receiver operating characteristic (ROC) curve (AUC) of 90.3%. Maintaining the Integrity of the Specifications.

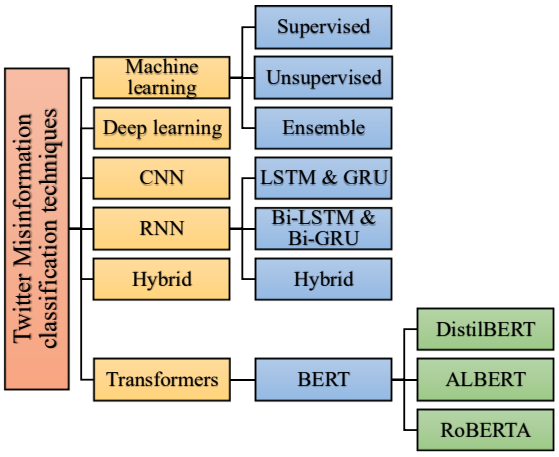


Fig. 1. Taxonomy of existing COVID-19 misinformation detection techniques. It describes the classification of various COVID-19 Twitter misinformation detection techniques, including machine learning, deep learning, and Transformer-based works.

B. Misinformation Detection Using Deep Learning Methods

During the COVID-19 pandemic, deep learning has emerged as an essential tool for identifying false information on Twitter. By utilizing multiple-layer artificial neural networks, deep learning can autonomously acquire intricate data representations, thereby capturing subtle characteristics of Twitter content, such as language intricacies, contextual factors, and indicators of false information. [18]. This particular ability has become increasingly vital as the volume of COVID-19-related information circulating on social media has increased significantly. Through the utilization of deep learning algorithms such as Convolutional Neural Networks (CNN) [19] or Recurrent Neural Networks (RNN) [20]. Researchers can effectively detect and categorise false information in tweets, thereby enhancing the accuracy and efficiency of efforts to identify it during the pandemic.

Convolutional Neural Networks (CNNs) are among the most popular and widely used models in natural language processing.

Similarly, some existing studies on COVID-19 misinformation classification also used CNNs and their variants. Kaliyar et al. [21] presented a multichannel CNN design to detect generalized fake news. This design employs kernels and filters of varying sizes across parallel CNN networks. The model, known as MCNNet, achieved higher accuracy (98.2%) and F1-score (98.1%) on the FN-COV dataset than CoAID. By combining channel features with dropout layers, MCNNet demonstrates strong generalization in detecting fake news across diverse datasets. In Elahadad et al. [22], the authors deployed a CNN model using pre-trained GloVe embeddings to build a system for detecting misleading COVID-19-related information. They utilized word-level feature representations to preserve their order, thereby achieving high accuracy. Arbane et al. [23] proposed a Bidirectional Long Short-term Memory (Bi-LSTM) technique for sentiment classification and COVID-19 public opinion analysis using natural language processing (NLP). Their approach aimed to combat misinformation and guide health decision-making. Four scenarios were considered, each based on a unique dataset. Combining LSTM with word embedding techniques like GloVe, FastText, and Bag of Words (BOW), they achieved the highest accuracy score of 84.54% on tweets datasets [23], [24], [25], and validation accuracy scores of 94.55% and 97.52% on Reddit comments datasets [23], [24], [25].

Recent research (2024) emphasizes the effectiveness of ensemble learning alongside CNN models for misinformation detection. Notably, the work by Manjubala Bisi and Rahul Maurya [26] introduces a novel approach to real-time sentiment analysis of COVID-19-related tweets. Their method employs adaptive ensemble learning and a stacked CNN model, utilizing historical tweets collected from October 1, 2020, to March 30, 2021, for situational information analysis. Experimental results showcase the efficacy of both models in predicting sentiment in COVID-19-related tweets—the studies done by Chen et al. [27]; and Yang et al. [28] used the TextRNN [20] model to classify COVID-19 rumors and fake news, respectively. The TextRNN model uses different LSTM layers inside its architecture.

C. Misinformation Detection Using Bidirectional Encoder Representations from Transformer (BERT)

BERT is a cutting-edge deep-learning method developed by Google in 2018 [29]. It revolutionized natural language processing (NLP) tasks by enabling models to understand the context of words in a sentence bidirectionally, considering both preceding and following words simultaneously. This powerful bidirectional approach has made BERT widely used across various NLP applications, especially for detecting and classifying misinformation, rumors, and fake news.

Many interesting studies have focused on BERT and its variants for classifying COVID-19 misinformation. For instance, Li et al. [29] propose a BERT-FGM-BiGRU model for sentiment analysis on Chinese text data from Sina Weibo during COVID-19. Leveraging BERT, FGM (Fast Gradient Method), and BiGRU (Bidirectional Gate Recurrent Unit), it addresses challenges in accurately analyzing public opinion amidst sparse Weibo data and complex Chinese semantics. Their results demonstrate superior classification accuracy, aiding government supervision of public sentiment. Furthermore, it reveals a spatial correlation between sentiment and pandemic severity, highlighting shifting sentiment trends over time. Another intriguing study by Muller et al. [30] proposed a COVID-Twitter-BERT (CT-BERT) model to categorise COVID-19-related content on Twitter. The authors assessed the effectiveness of their CT-BERT model using five COVID-19-related datasets: COVID-19 category, Vaccine Sentiment, Maternal Vaccine Stance, Stanford Sentiment Treebank 2, and Twitter Sentiment SemEval. Their model outperformed BERT-LARGE, achieving a modest accuracy of 83.3%. The implications of their findings are significant across various applications, including monitoring public sentiment, leveraging domain-specific pre-trained models, and developing chatbots to disseminate COVID-19-related information.

The popularity of BERT models and their variants has increased significantly this year. Srivastava et al. [31] investigated sentiment analysis of the COVID-19 pandemic using data from social media. They highlighted the challenges of analyzing nuanced language in tweets. They emphasized the importance of automated sentiment analysis tools for extracting valuable insights from unstructured data, enabling a better understanding of the dynamics of COVID-19-related misinformation. They employed a hybrid BERT model with a multitailed CNN. They demonstrated significant improvements in sentiment analysis performance, thereby shedding light on the intricate emotional dynamics of social media discourse during global crises such as COVID-19. Kusuma et al. [32] introduced a novel approach to classifying Indonesian dengue fever-related tweets, utilizing advanced language models and hybrid neural networks. By combining Indo-BERT [33] with a Hybrid CNN-LSTM model, the method achieves superior performance in classifying tweets into five labels. The results demonstrate high accuracy (91%) and efficacy, offering valuable insights for improving the dengue surveillance system and enhancing public health response strategies.

III. Research Gap

Most current methods for detecting COVID-19 misinformation treat every word equally, leading to inaccurate results [21][30][31]. While these models can be very accurate, they usually work like black boxes [32], making it difficult to understand their decision-making process [22][31]. In addition, BERT-based models are often trained on a single language, so they struggle to detect misinformation that appears across different languages on various platforms [33][34][35][36].

CA-BERT addresses these issues in three main ways. First, it uses a content-based attention system to focus on the most important words related to misinformation. Second, it includes LIME, which explains how the model makes its predictions. Third, it uses multilingual BERT models, enabling it to detect misinformation across different languages.

IV. Methodology

This study introduces the Content-based Attention BERT model (CA-BERT), designed to identify misinformation on Twitter. Fig. 2 provides a simplified schematic of the proposed CA-BERT model. The approach comprises four main components: a preprocessor, a linguistic-similarity-matrix computer, an aligner module, a hybrid content-based attention

mechanism, and a BERT classifier. Initially, raw multilingual tweets undergo pre-processing. Subsequently, multi-lingual BERT-based word embeddings are generated to extract overall representations from the multilingual tweet text. The content-based attention mechanism is then employed to extract significant features indicative of misinformation or true information. Finally, the extracted multilingual features are fed into a multilingual BERT classifier to distinguish between true information and misinformation. Below, we explain each step in detail.

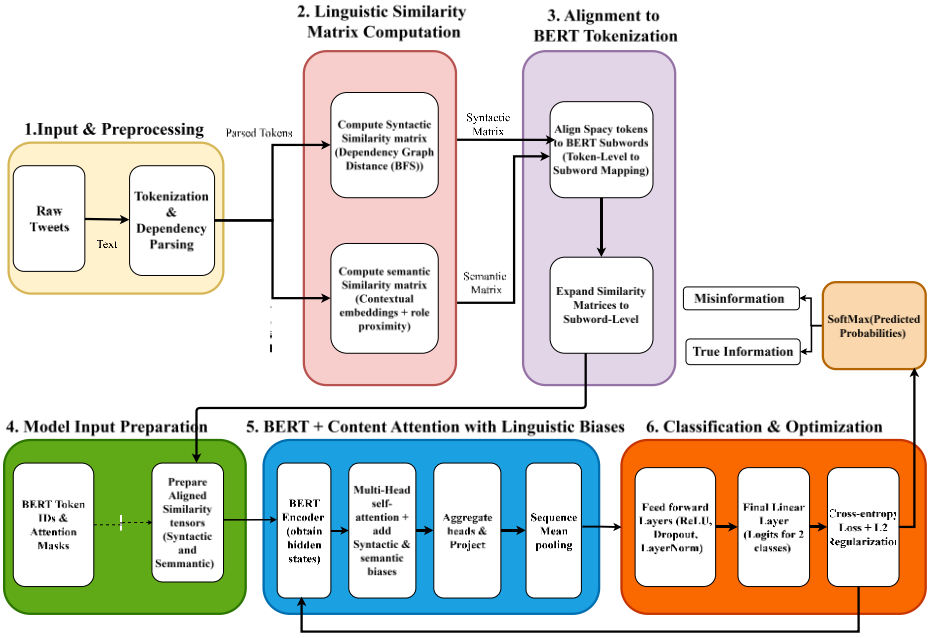


Fig. 2. End-to-end workflow of proposed CA-BERT model. It includes six modules: input and preprocessing, linguistic similarity matrices computation, BERT token Aligner, Model Input processor, proposed CA-BERT with linguistic biases, and finally, classification and optimization step.

A. Problem Formulation

Let $Y = \{real, misinformation\}$ be the target label set and $T = (t_1, t_2, \ldots, t_S)$ be the input token sequence of a cleaned and tokenized multi-lingual tweet, where S is the input sequence length. The aim is to learn a function. $f_{m,\propto,\beta} : V^S \rightarrow \Delta^1$, where V is the token vocabulary and Δ^1 is the probability simplex over Y and m represents the model weights, α and β are the scaling factors for syntactic and semantic attention biases, respectively. This mapping predicts the label. $y_{pred} = argmax_{c\in\{0,1\}} p_c$ with p_c denoting the model's predicted probability for class $c \in Y$. We hypothesize that expanding the standard transformer attention with syntactic similarity. S^{syn} and semantic similarity S^{sem}, which encodes structural and semantic relations between tokens, yields a more accurate mapping $f_{m,\propto,\beta}$ than the vanilla BERT baseline variants and other existing works such as HANN and DAAN. Where, S^{syn}, S^{sem} are the precomputed token similarity matrices capturing syntactic and semantic relations between tokens T and accurately and precisely capturing misinformation tweets.

B. Tokenisation and Pre-processing

To prepare the multi-lingual COVID-19 tweets for CABERT, each raw tweet s is passed through a standard cleaning pipeline $f_i, \ldots, f_K : s' = (f_K \ldots . f_1)(s)$, where functions f_i sequentially remove dataset-specific noise: URLs, emojis, non-UTF8 characters, user mentions, hashtags. Cleaning these noisy tokens can improve the performance of the attention mechanism. The normalized string s' is then tokenized by word-piece t' into a sub-word sequence $T = t'(s') = (t_1, t_2, \ldots, t_S)$, where S is the sequence length and each $t_i \in V$ (dynamic vocabulary). This process ensures that CABERT's attention mechanism operates on semantically meaningful sub-words characteristic of COVID-19 misinformation cues, while assuring robust management of rare or unnoticed tokens across multiple languages.

Finally, the BERT tokenizer maps each $token\ t_i$ in the word sequence to a numerical ID, $id_i = tokenizer(t_i)$ incorporating special tokens, truncation, and padding. Each token is represented by a word piece embedding, $e(t_i)$, in a continuous vector space. This embedding captures the contextual information of raw tweets. The resulting sequence of token embeddings, e_i, along with the attention mask, forms the input data for the proposed model and the subsequent steps.

C. Syntactic Dependency Graph and Similarity Matrix Generation

Following tokenization, the sequence $T = (t_1, t_2, \ldots, t_S)$ is represented as an undirected dependency graph $G = (v, e)$, where each node $v_i \in V$ corresponds to a token t_i and an edge $(v_i, v_j) \in \varepsilon$ represents direct syntactic dependencies between tokens as inferred by a multilingual dependency parser, as illustrated in the Fig. 3. An adjacency matrix encodes this graph. $A^{dep} \in \{0,1\}$, with

$$A_{ij}^{dep} = \begin{cases} 1, & if\ (v_i, v_j) \in \varepsilon, \\ 0, & otherwise. \end{cases} \quad (1)$$

To compute the syntactic similarity, the shortest path distance d_{ij} between every node pair (i, j) using an unweighted breadth-first search traversal algorithm on G. Distance on the diagonal is zero if no path exists between edge pairs v_i and v_j, else $d_{ij} = S$ (sequence length) to denote the maximum separation. It can be formally denoted as follows:

$$d_{ij} = \begin{cases} 0 & if\ i = j, \\ BFS\ shortest\ path\ (v_i, v_j) & if\ path\ exists, \\ S & if\ no\ path\ exists. \end{cases} \quad (2)$$

Then the syntactic similarity matrix is computed as,

$$S_{ij}^{syn} = \frac{1}{1 + d_{ij}} \quad (3)$$

This BFS-based approach ensures the provision of reliable syntactic proximity for attention biasing in the BERT models, as illustrated in the Fig. 4.

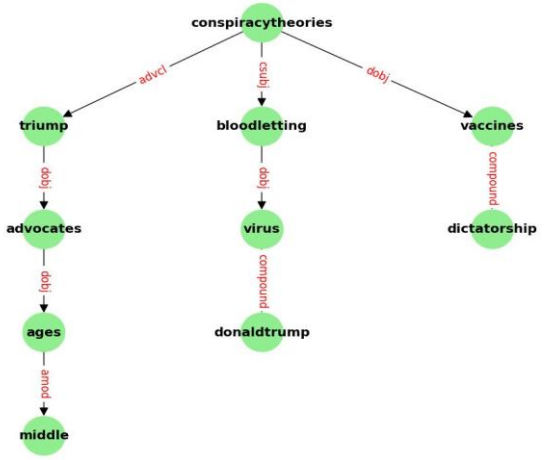


Fig. 3. Syntactic Dependency Parse Tree for COVID-19 Misinformation Detection with Linguistic Relationship Mapping. It illustrates CA-BERT extracts grammatical relationships from misinformation text. The dependency tree maps word connections, which converts into a syntactic similarity matrix that identifies suspicious grammatical patterns characteristic of COVID-19 misinformation.

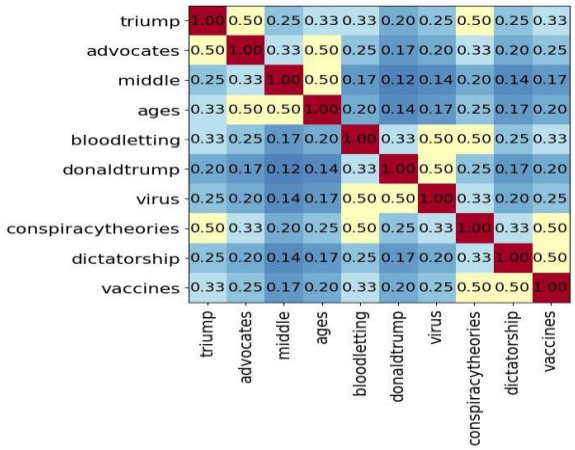


Fig. 4. Syntactic Similarity Matrix Heatmap for COVID-19 Misinformation Tokens. It visualizes syntactic similarity scores (red=high, blue =low) between word pairs, where suspicious grammatical proximities like 'vaccines-dictatorship'(0.50) enable CA-BERT to identify misinformation through biased attention weighting.

D. Semantic Role and Similarity Matrix Generation

In the next step, contextual meaning and grammatical role information are transformed into a token-level similarity matrix $S^{enhaced}$, for each tokenized tweet $T = (t_1, t_2, \ldots, t_i)$. Fig. 5 depicts the sample semantic role enhanced similarity matrix. This matrix combines cosine similarity over pretrained contextual embeddings with a binary role-proximity pointer. First each token t_i is mapped to its pretrained embedding $e_i \in R^d$ (from a multi-lingual BERT model). Then these embeddings are normalized and pairwise cosine similarities are computed S_{ij}^{sem}.

$$S_{ij}^{sem} = \cos(e_i, e_j) = \frac{e_i \cdot e_j}{||e_i|| ||e_j||} \quad (4)$$

Next, a binary matrix $B \in \{0,1\}$ is defined to capture whether two tokens share the same grammatical head or grammatical label. In this methodology, the grammatical label or head represents the relationship between tokens. t_1 and token t_2 is a subject, direct object or indirect object. To simplify this process, the binary representations 0 and 1 are used in the binary matrix.

$$B_{ij} = \begin{cases} 1, & if\ head(t_i) = head(t_j)\ V\ dep(t_i) = dep\ (t_j), \\ 0, & otherwise. \end{cases} \quad (5)$$

Finally, an enhanced similarity matrix $S^{enhaced}$ is computed by adding a scaled role signal $\boldsymbol{B_{ij}}$ to the cosine similarities S_{ij}^{sem}. Then S_{ij}^{enh} will be, $S_{ij}^{enh} = S_{ij}^{sem} + \gamma\, \boldsymbol{B_{ij}}$, where γ is a tunable hyperparameter that controls the influence of shared semantic roles. This enhanced S_{ij}^{enh} guides the model's attention mechanism to influence deep contextual semantics and explicit grammatical role proximity jointly, improving its ability to detect nuanced misinformation patterns. Fig. 5 illustrates the sample semantic role enhanced similarity matrix.

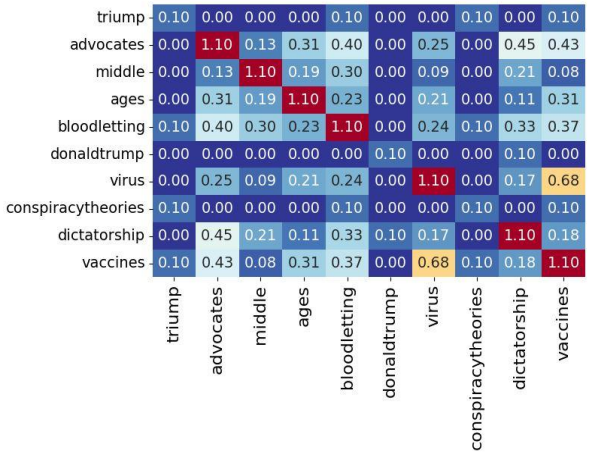


Fig. 5. Semantic Role-Enhanced Similarity Matrix Heatmap displays semantic similarity scores enriched with semantic role information (red = high semantic relatedness, blue = low), enabling CA-BERT to identify deceptive semantic associations like 'vaccines-virus' (0.68) and 'virus-conspiracytheories' (0.10) that characterize COVID-19 misinformation.

E. Alignment to BERT Subword Space

Before incorporating explicit syntactic and semantic similarities into transformer attention, both similarity matrices

S_{ij}^{syn} and S_{ij}^{enh} with BERT's sub word tokenization. Fig. 6 and Fig. 7 represent S_{ij}^{syn} and S_{ij}^{enh} with BERT's sub word tokenization, showing how closely related each token pair is based on structure or meaning. Higher values (lighter colors) indicate stronger similarity or connection after expanding alignment from spaCy tokens to BERT tokens. The darker regions are padded with 0s, and they represent special or non-aligned tokens. For each, the similarity matrices $Sim_{mat} \in \{S_{ij}^{syn}, S_{ij}^{enh}\}$, expand to match BERT sub token indices u, v by,

$$Sim_{mat_{uv}} = \begin{cases} Sim_{mat_{m(u),m(v)}} & if\ m(u)\ and\ m(v) \neq\ no\ align, \\ 0, & otherwise. \end{cases} \quad (6)$$

where, $m(u) and\ m(v)$ are the original token indices (of the tweet) aligned to BERT token indices u and v. Next, to process batches uniformly, each S_{ij}^{syn} and S_{ij}^{enh} is zero-padded or dynamically truncated to match the BERT sub token indices. Theoretically, this alignment ensures two factors. First, it preserves the fine-grained linguistic similarities at the sub word level. Second, it ensures that both. S_{ij}^{syn} and S_{ij}^{enh} Dimensions are padded to a uniform size. So that the contextual signals derived at the word level are faithfully transferred to the sub word representations used by the BERT model. Fig. 6 Showcases syntactic matrix aligned to match BERT token space, and Fig. 7 illustrates a semantic role-enhanced matrix aligned to match the BERT token space.

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, SC, dc, and rms do not have to be defined. Do not use abbreviations in the title or headings unless they are unavoidable.

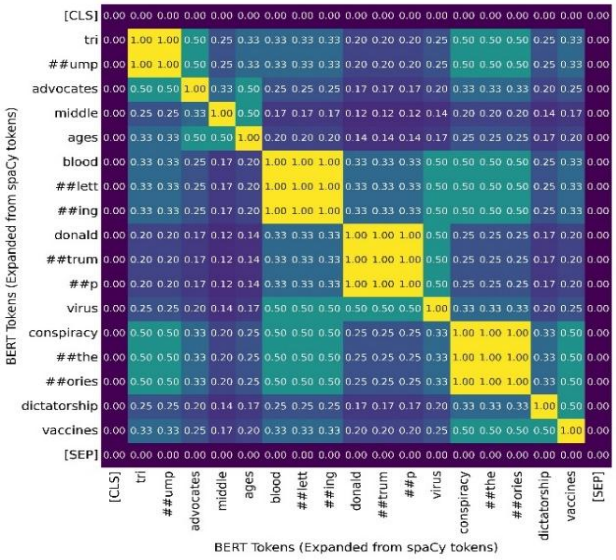


Fig. 6. BERT Subword Token Alignment and Attention Mapping. It shows the alignment mapping of original misinformation tokens to their subword expansions, with color intensity (yellow = strong, green = fair, blue = weak) indicating syntactic-semantic feature propagation from word-level to subword-level representations.

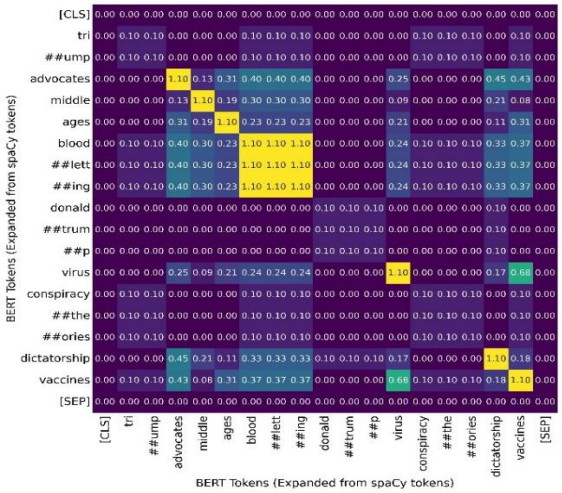


Fig. 7. Semantic Similarity Matrix Aligned to BERT Token Space. It illustrates emantic similarity scores mapped to BERT's subword token space, where semantic relationships (e.g., 'virus' and 'vaccines' showing high similarity – 0.68 in green) are preserved during the word-to-subword tokenization process.

F. Content-Based Attention Mechanism and Classification

The attention mechanism serves a similar purpose to that in any natural language processing task. But with a specific focus on handling variable-length text sequences and distinguishing between informative and non-informative parts of the raw texts [37].

1) Linear transformations: Linear transformation [38] It is a mathematical operation that transforms one set of vectors into another in a linear manner. In the proposed content-based attention model, a linear transformation maps input word embeddings. e (t_i) to query (Q), key (K) and value (V) vectors. A weight matrix represents each linear transformation.

$$Q = E\ (t_i).W_q$$
$$K = E\ (t_i).W_k$$
$$V = E\ (t_i).W_v$$

Here, X represents the input word embeddings $E\ (t_i)$, and W_q, W_k, W_v are weight matrices specific to query, key and value transformations, respectively. These matrices are obtained by multiplying the input embeddings. $E\ (t_i)$ with weight matrices W_q, W_k, and W_v respectively, followed by reshaping to split into multiple heads as represented.

2) Calculating the attention scores with syntactic and semantic similarities: Attention scores indicate the importance of each token in the sequence relative to every other token. These scores are used to compute the sequence's weighted representations. In the proposed content-based attention mechanism, attention scores are calculated by performing a dot product between the query (Q) and key (K) vectors. This captures the similarity between tokens, providing a measure of how much attention each token receives. t_i (informative tokens) should pay to every other token t_j (non-informative or

padded tokens) in the word embedded sequence (in different languages).

$$raw\ attenion = QK^T$$

$$A_{ij} = \frac{QK^T}{\sqrt{d_h}} \quad (7)$$

After obtaining the raw attention scores , they are scaled by a factor of $1/\sqrt{d_h}$, where $d_h$is the dimension of the query (Q) and key (K) vectors to compute the scaled attention scores. Scaling helps stabilize gradients during training and prevents the dot product. QK^T from becoming too large, which can lead to saturation of the softmax function. Next, integrating semantic and syntactic similarities into the attention mechanism helps capture multilingual misinformation by enhancing the model's ability to discern appropriate information across different languages. The scaled attention vectors A_{ij} are then integrated with syntactic S_{ij}^{syn} and semantic similarity S_{ij}^{enh} weights using weighted aggregation. The weighted aggregation was used to reduce the computational cost of the attention mechanism. This results in FA_{ij} $enhanced\ attention$. It is represented as follows.

$$FA_{ij} = \frac{QK^T}{\sqrt{d_h}} + \alpha \cdot S_{ij}^{syn} + \beta \cdot S_{ij}^{enh} \quad (8)$$

where, $\alpha, \beta > 0$ These are learnable scaling parameters that control the relative importance of syntactic and semantic biases. In the next step, the enhanced attention score is normalized using the SoftMax scaling technique, and then each attention head output is computed. $\mathbf{O}_i$ for each FA_{ij} having a sequence length S.

$$\alpha_{ij} = \frac{\exp(FA_{ij})}{\sum_{k=1}^{S} \exp(FA_{ik})} \quad (9)$$

$$\mathbf{O}_i = \sum_{j=1}^{S} \alpha_{ij} \mathbf{V}_j \quad (10)$$

α_{ij} are the enhanced attention weights after SoftMax scaling and O_i is the enhanced contextualised output of each attention head with S_{ij}^{enh} and S_{ij}^{syn} This modification implements structured attention influence, creating soft limits that guide attention flow along linguistically meaningful paths while preserving differentiability.

G. Classifier–Feed Forward

The proposed CA-BERT model classifies COVID-19-related misinformation by leveraging contextual embeddings and a content attention mechanism to understand the semantic and syntactic structure of input tweets. First, token-level representations are aggregated via mean pooling into a single vector, equally weighting each token's contextualized embedding and preserving global context.

$$\bar{O} = \frac{1}{S} \sum_{i=1}^{S} O_i \quad (11)$$

where, $\bar{O}$ is a sequence-level representation, obtained by averaging all token vectors, O_i is the token level representation of the i-th token, a vector produced by the preceding attention layers, and S is the sequence length (input tweet). This aggregate sequence representation $\bar{O}$ is transformed for classification with a combination of linear projection, ReLU activation, dropout regularization, and layer normalization ($h_{hidden,i}$). First the $\bar{O}$, is linearly projected into a new hidden representation for the i-th dimension, $h_{linear} = W_1 \bar{O} + b_1$. W_1 are the weight matrix mapping input vectors to the hidden space and b_1 is the bias vector included for each dimension. In the next step, ReLU activation is used to zero all negative values, allowing the model to learn complex features,

$$h_{relu} = max(0, h_{linear}, i) \forall\ i = \{1, \ldots., d_{model}\} \quad (12)$$

Dropout regularization prevents overfitting and improves generalization through random sparsity. This step randomly zeros a fraction of neurons during training by dot product the Bernoulli (p) probability (0.1) with $h_{relu,I}$ as follows, $h_{droput} = m_I \,.h_{relu,I}$. where m_I The Bernoulli (p) probability (0.1). Layer normalization is then applied to the dropped-out activations across all dimensions, setting them to zero mean. μ and unit variance σ,

$$\mathrm{h_{hidden,i}} = \frac{h_{droput,I} - \mu}{\sigma} \,.\gamma_i + \beta_i \quad (13)$$

where, μ is mean of dropped-out activations, σ is the unit variance, γ_i and β_i are learnable scaling and shifting parameters to smooth the model's learning process. The $\mathrm{h_{hidden,I}}$ is then mapped to output logits $z = [z_0, z_1]$, with z_0 as true information and z_1 as misinformation as described in the Fig. 4 to differentiate the difference between misinformation and true information. The final classification layer is represented as,

$$z = W_2 h_{hidden} + b_2 \quad (14)$$

$$p(misnformation\,|s) = \frac{\exp(z1)}{\exp(z_0) + \exp(z1)} \quad (15)$$

where, $z = [z_0, z_1]$ represents the raw logits representing model scores for each class, W_2 is the weight matrix for the final dense layer and b_2 is the bias vector. $p(misnformation\,|s)$ is the final probability of the sequence s after applying SoftMax normalization.

H. Loss Function and Optimisation

The proposed model is trained using a regularized cross-entropy loss. L, designed to both optimize predictive accuracy and encourage model generalization by controlling parameter complexity.

$$L = -\frac{1}{N} \sum_{n=1}^{N} \sum_{C=0}^{1} (y_{n,c} \log p_{n,c} + \lambda_1 ||\theta||_2^2 + \lambda_2 ||\alpha||_2^2 + \lambda_3 ||\beta||_2^2) \quad (16)$$

where, N is the batch size, representing the number of samples (tweets). $y_{n,c} \in \{0,1\}$ is a label indicating if the sample n belongs to the class c (misinformation or true information). $p_{n,c}$ is the predicted probability of class c for sample n , produced by applying SoftMax to the model's logits (as described in 5.4). The first term $-\frac{1}{N} \sum_{n=1}^{N} \sum_{c=0}^{1} y_{n,c} \log p_{n,c}$ penalizes the model for incorrect predictions, thereby driving it to assign high confidence to true classes. The second term $\lambda_1 ||\theta||_2^2 + \lambda_2 ||\alpha||_2^2 + \lambda_3 ||\beta||_2^2$ represents the L_2 normalization of different groups of model parameters. Here θ denotes the core model weights, α and β represent attention-specific parameters (as described in 5.3 and

5.4). λ_1, λ_2,λ_3. *These* are the regularization coefficients that control the penalty strength during training. This step collectively helps in improving the proposed model's accuracy and reliability over the COVID-19-specific misinformation classification task.

V. Experimental Setup and Baseline Variants

A. Datasets Used

We ran experiments on the CovidMis20. [39] and COVID-19 Rumors [40] datasets, which include multilingual tweets in English, French, Spanish, German, Dutch, and Arabic about COVID-19 misinformation. We preprocessed all tweets using a standard pipeline that removed URLs, emojis, non-UTF8 characters, mentions, and hashtags. Then, we tokenized the tweets using the multilingual BERT tokenizer. Fig. 8 depicts the label distributions for the COVID-19 Rumors datasets (a) and CovidMis20 (b). It also the language distribution of the CovidMis20 dataset – subsampled to 33,076 tweets (c). It consists of 31.8% English tweets, 11.8% Dutch tweets, 7.8% Arabic tweets, 21.9% French tweets, 9.7% German tweets, and 17% Spanish tweets.

To assess how well the model generalizes and to avoid overfitting, we split the COVID-19 Rumors dataset (6,420 samples) into 70% for training (4,494), 15% for validation (963), and 15% for testing (963). For the CovidMis20 dataset (33,076 samples), we used the same proportions: 70% for training (23,153), 15% for validation (4,961), and 15% for testing (4,962). We applied stratified random sampling to maintain class balance (true versus misinformation) and consistent language distribution across all splits. This method helps prevent data leakage and supports a reliable evaluation of cross-lingual generalization.

B. Implementation Details

We used PyTorch 2.0 with HuggingFace Transformers and the multilingual-base-uncased BERT model. For dependency parsing, we relied on spaCy's multilingual models. The main hyperparameters were set as follows: the semantic role weight γ was set to 0.2, with α and β set to 0.3 and optimised during training. Regularisation included an L2 weight λ_1 of 0.001 and attention parameter weights λ_2 and λ_3 at 0.0005. Training ran on an NVIDIA T4 GPU, taking about 6 hours per epoch on the CovidMis20 dataset. Inference took 45 milliseconds per document. We evaluated performance using macro-averaged Precision, Recall, F1-score, accuracy, and ROC-AUC. For adversarial robustness, we applied FGSM attacks with ε set to 0.3. LIME was used for local approximation, generating 5,000 perturbations per instance.

C. Baseline Models to Compare

Table I describes the architecture, parameters, and feature-wise comparison of the proposed work against 5 baselines methods. We compared our results with BERT-base. [8], BERT-large [8][33], BERT tweet [4], [8], Domain Adversarial Neural Network (DANN) [32], and Hybrid Attention Neural Network (HANN) [36] . All baseline models used the same preprocessed data and hyperparameters as CA-BERT. We measured accuracy, precision, recall, and F1-score, and also examined how the models performed across different languages and in multilingual scenarios.

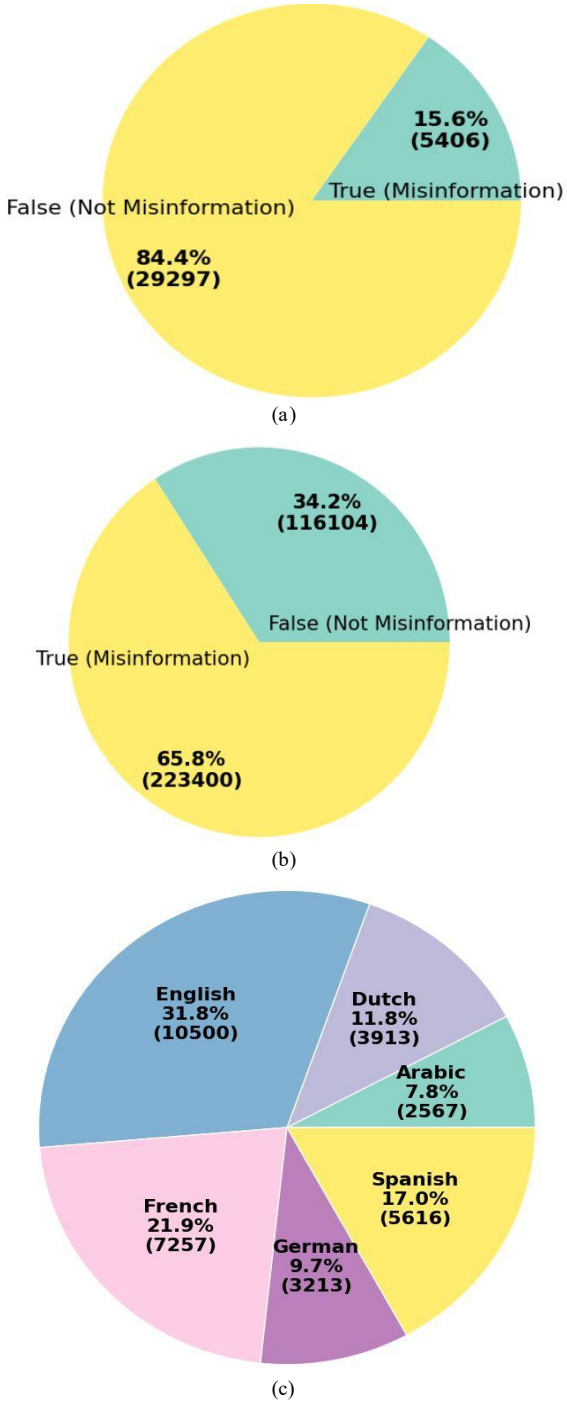


Fig. 8. Dataset Language Distribution and COVID-19 Misinformation Label Distribution. It showcases the distribution of each dataset used in experiment. a) Illustrates the label distribution of Covid-19 rumors dataset [40] with 15.6% of misinformation and 84.4% true information. Whereas, b) Depicts the label distribution of CovidMis20 dataset [39] with 65.8% of misinformation and 34.2% of true information. Finally, c) illustrates the language distribution of the combined dataset. It consists of eight languages with English 31.8%, French 21.9%, Spanish 17.0% across 33,513 tweets.

TABLE I. COMPARISON OF PROPOSED MODEL VS. BASELINE VARIANTS

Feature	BERT_base [8]	BERT_large [8]	BERT_tweet [8]	DANN [31]	HAAN [35]	CA-BERT
12 blocks	✓	✗	✓	✗	✗	✓
768 units	✓	✗	✓	✓	✓	✓
PARAMS (M)	110	340	110	~110	~110	~120
Twitter	✗	✗	✓	✗	✗	✓
Adversarial	✗	✗	✗	✓	✗	✓
Hierarchical	✗	✗	✗	✗	✓	✓
Syntactic/Semantic	✗	✗	✗	✗	✗	✓
Multilingual	✗	✗	✗	✗	✗	✓

VI. RESULT ANALYSIS AND LIME EXPLANATIONS

A. English Dataset Performance of the Proposed Work (CovidMis20 and Covid19-Rumours)

Table II and Fig. 10 show English performance on CovidMis20 and Covid19-rumours. From Fig. 10, CA-BERT has the highest recall for true information at 96%. Its macro-averaged F1 is 92.5% (precision 93%, recall 92.5%), which is 2.5 to 4.5 percentage points better than the 2025 baselines. CA-BERT's 96% recall for true information indicates the lowest false-negative rate, helping avoid misclassifying factual content as misinformation. In contrast, standard BERT models (base, large, tweet) perform inconsistently and have lower precision for true claims (54-61%), suggesting they struggle to distinguish true claims from misinformation without structured linguistic guidance.

B. Multilingual Performance of Proposed Work (CovidMis20)

Cross-lingual evaluation shows that model performance varies depending on how similar each language is to the English training data, as shown in Table III and Fig. 9. Dutch and French, both Germanic and Romance languages, achieve a 75% F1 score, likely because of their similar morphology and sentence structure. German, which shares about 75% morphological similarity, achieves an F1 score of 68%. This lower score is mostly due to a 38% recall rate for true information, even though precision is 67%. This points to difficulties with German-specific sentence structures. Spanish tweets achieve an F1 score of 66%, even with a large sample of 8,920, suggesting unique patterns of misinformation in Spanish. Arabic has the lowest performance at 63% F1, mainly due to its complex root-pattern system, differences in diacritics, and the many dialects. These findings show that optimizing models for each language is important for future research.

TABLE II. PERFORMANCE COMPARISON OF PROPOSED WORK VS. BASELINES ON COVIDMIS20 AND COVID-19 DATASET (ENGLISH)

Model	Precision (True)	Recall (True)	F1 (True)	F1 (Misinf.)	Precision (Misinf.)	Recall (Misinf.)
BERT-base [8]	0.55	0.50	0.52	0.87	0.85	0.88
BERT-large [8]	0.61	0.58	0.58	0.81	0.78	0.82
BERT-tweet [8]	0.54	0.58	0.56	0.86	0.87	0.85
Enhanced-DANN [31]	0.87	0.89	0.88	0.84	0.89	0.87
Graph-HANN [35]	0.90	0.88	0.89	0.87	0.88	0.90
CA-BERT (Proposed)	0.91	0.96	0.93	0.92	0.95	0.89

TABLE III. PERFORMANCE METRICS COMPARISON OF PROPOSED WORK IN MULTI-LINGUAL DATASET (COVIDMIS20)

Language	Model	Precision	Recall	F1-Score	Samples
English	CA_BERT	0.93	0.92	0.925	8,640
Dutch	CA_BERT	0.76	0.75	0.75	4,250
French	CA_BERT	0.75	0.76	0.75	3,890
German	CA_BERT	0.68	0.68	0.68	2,180
Spanish	CA_BERT	0.68	0.66	0.66	8,920
Arabic	CA_BERT	0.63	0.64	0.63	4,083

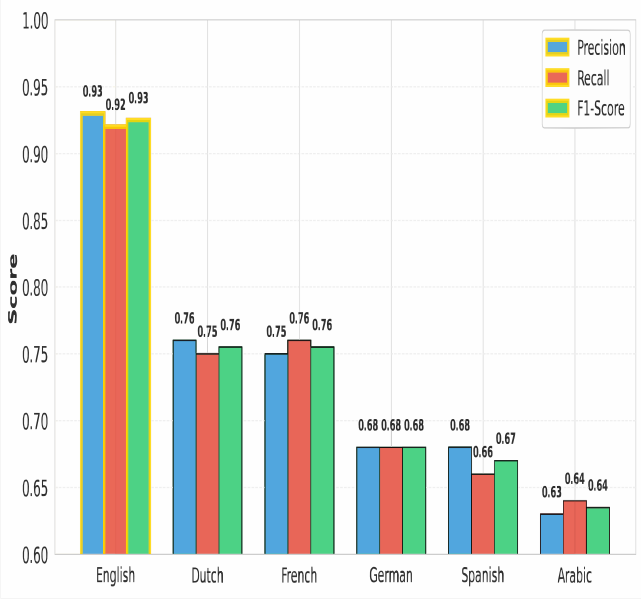
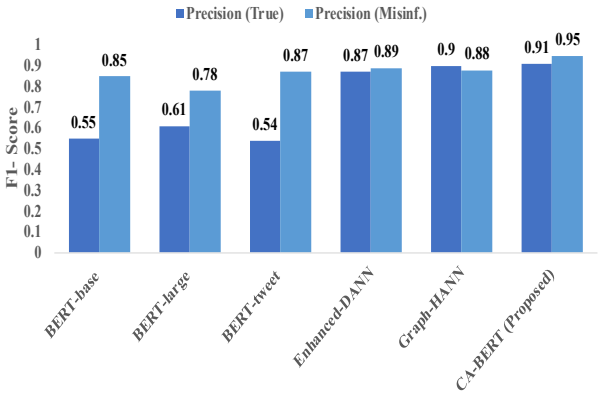
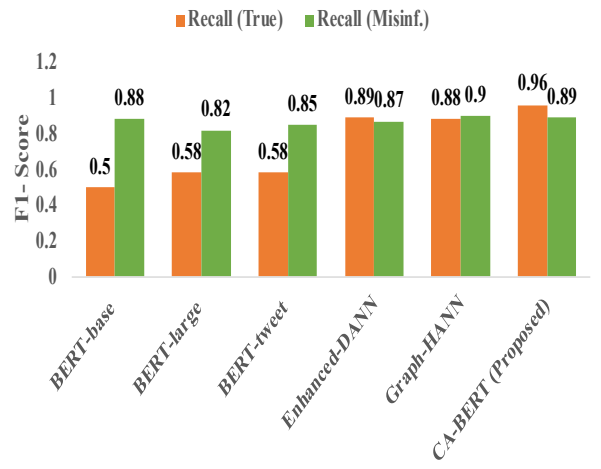


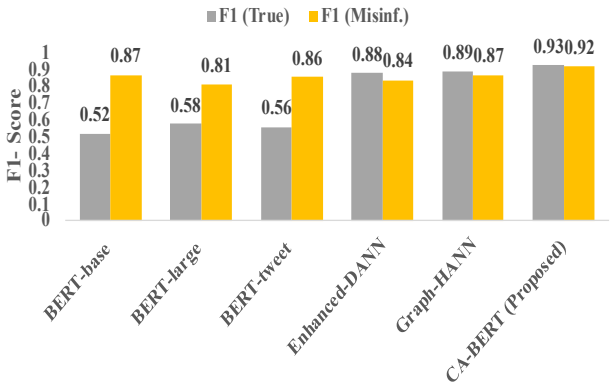
Fig. 9. Multilingual Performance of proposed work (CovidMis20) vs. English, Spanish, French, Dutch, German, and Arabic. CA-BERT achieves the highest performance on English (F1-Score 0.93), with moderate performance on Dutch and French (0.76, 0.75), and lowest on German, Spanish, and Arabic (0.68, 0.67, 0.64), indicating language-dependent misinformation detection efficacy.



(a)



(b)



(c)

Fig. 10. Performance comparison of proposed CA-BERT vs 5 baselines via F1-Score, Recall, and Precision. (a) F1-Score: CA-BERT achieves 0.92 (true) and 0.87 (misinformation), outperforming baselines. (b) Recall: CA-BERT shows 0.96 (true) and 0.89 (misinformation), maximizing detection sensitivity. (c) Precision: CA-BERT maintains 0.91 (true) and 0.95 (misinformation), minimizing false positives across metrics.

C. Ablation Study

Fig. 11 shows the ablation study of the proposed model. Using joint integration gives the best results. The ablation study measures the extent to which each component helps. Adding only syntactic similarity improves F1 by 1.9 points (86.5 to 88.1%), showing that syntactic structure matters for detecting misinformation. Adding only semantic roles improves F1 by 2.1 points (86.5% to 88.6%), underscoring their value. Using both together improves F1 by 6.0 points (86.5% to 92.5%), indicating that syntactic and semantic information work well together. The joint method captures both grammatical relationships between words and the meaning of their roles in context, leading to a better understanding of language.

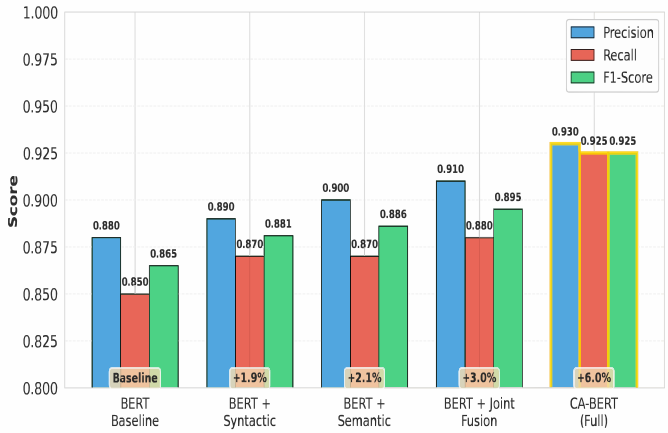


Fig. 11. Ablation Study - Progressive CA-BERT Component Integration. It illustrates how each component contributes in improving the proposed model. BERT baseline 0.860, +Syntactic 0.881 (+1.9%), +Semantic 0.886 (+2.1%), +Joint 0.895 (+3.0%), CA-BERT Full 0.925 (+5.0%) F1-Score

Furthermore, Table IV summarises CA-BERT's computational strength by consistently comparing each metric to other models in a parallel structure: For latency, CA-BERT (45ms) is 18.4% slower than BERT-base (38ms) but 36.6% faster than Graph-HANN (71ms). For F1-score, CA-BERT

achieves 92.5%—4.6 points above BERT-base (87.9%) and 3.6 points above Graph-HANN (88.9%). Throughput for CA-BERT is 22.2 tweets/sec, compared to BERT-base (26.3), Enhanced-DANN (14.7), and Graph-HANN (14.1). GPU memory use is 2,847MB—7.4% more than BERT-base's 2,650MB and 19.2% less than Graph-HANN's 3,520MB. Energy per tweet is 0.028J, 16.7% higher than BERT-base (0.024J), but CA-BERT is 6.4 times more energy-efficient per F1 point than Graph-HANN (20.5%), using only 3.2% as much. Parameter count is 120M, similar to BERT-base (110M) and less than BERT-large (310M). Processing 1M tweets daily requires about 1.9× the GPU-hours of BERT-base. Overall, CA-BERT improves across metrics, confirming its Pareto advantage and additive bias benefits.

TABLE IV. COMPARISON OF BASELINE VS. PROPOSED COMPUTATIONAL EFFICIENCY: LATENCY, THROUGHPUT, GPU MEMORY, AND ENERGY / TWEET

Model	Latency (ms)	Throughput (tweet/s)	GPU Memory (MB)	Energy/tweet (J)	Parameters (M)
BERT-base	38	26.3	2,650	0.024	110
BERT-large	62	16.1	3,380	0.039	340
BERTweet	41	24.4	2,720	0.026	110
Enhanced-DANN	68	14.7	3,420	0.043	110
Graph-HANN	71	14.1	3,520	0.045	110
CA-BERT (Proposed)	**45**	**22.2**	**2,847**	**0.028**	**120**

D. Adversarial Robustness

Fig. 12 illustrates the adversarial robustness of the proposed model. Under FGSM adversarial attacks (ε=0.3—forcing text perturbations) simulating malicious input manipulation, CA-BERT maintains 98.1% accuracy, compared to 92.8% for Enhanced-DANN (a 5.3-point advantage). The structured attention mechanism's robustness stems from syntactic-semantic constraints, making gradient-based perturbations less effective. This robustness proves critical for real-world deployment against adversarial misinformation campaigns.

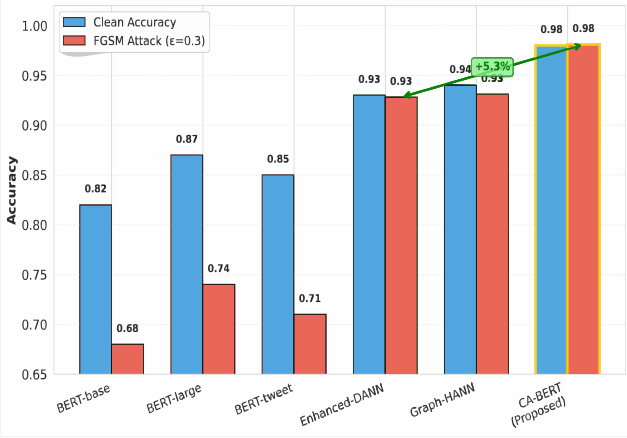


Fig. 12. Adversarial robustness (FGSM attack $\varepsilon = 0.3$) of the proposed CA-BERT model. CA-BERT maintains 0.98 clean accuracy and 0.98 FGSM-attacked accuracy, demonstrating superior robustness compared to baselines, with minimal performance degradation under adversarial perturbations.

E. LIME Explainability Analysis

Fig. 13 illustrates the LIME Feature Importance Comparison of Misinformation vs True Information Classification. LIME-based explanations help show how the model makes decisions. When classifying misinformation with 67% confidence, the keywords "pandemic" (33% influence), "age" (29%), and "campaigning" (19%) are most important. These words often appear in vaccine conspiracy stories. For true information, which the model identifies with 96% confidence, "coronavirus" is the strongest indicator (34% influence), reflecting language from trusted medical sources. Ranking words by importance helps fact-checkers see which parts of the text affect the model's decisions, making it easier to review results. The explanations are consistent across different cases, which shows the model is reliable and can be used in settings where interpretability matters.

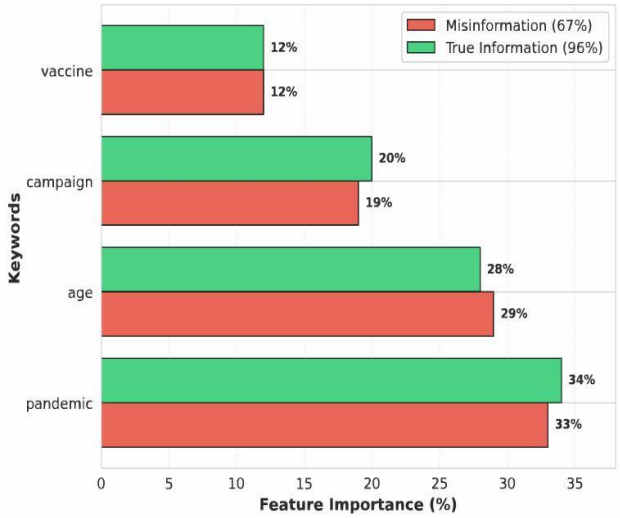


Fig. 13. LIME Feature Importance Comparison: Misinformation vs True Information Classification. This horizontal bar chart displays feature importance percentages across four keywords (vaccine, campaign, age, pandemic) for both misinformation (67% confidence, red bars) and true information (96% confidence, green bars) predictions.

VII. DISCUSSIONS AND LIMITATIONS

A. Key Contributions and Discussions

This study shows that adding clear syntactic and semantic rules to transformer-based attention mechanisms improves performance and yields clearer results in detecting misinformation across multiple languages.

CA-BERT outperforms recent 2025 baselines by 2.5 to 4.5 percentage points, thanks to structured linguistic rules that guide

attention. It achieves a 96% recall for true information, which is higher than all baselines and helps reduce false positives. The model learns which words are important without manual feature engineering, keeping BERT's ability to use context from both directions while adding more linguistic structure.

Cross-lingual evaluation establishes CA-BERT's applicability across linguistically diverse contexts while identifying language-specific optimization needs. Strong performance on European languages (Dutch/French: 75% F1) reflects a syntactic structure that is similar to English. Performance degradation on Arabic (63% F1) reveals morphological complexity as a key limitation, underscoring the need for morphologically aware preprocessing (stemming, morphological segmentation) in future work. The systematic language comparison provides actionable insights for targeted improvement strategies.

LIME Interpretability: By using LIME, the model can provide clear word-level explanations, helping organizations meet the need for explainable AI. It is also very robust against attacks, with a 98.1% success rate under FGSM, which is much better than other models and important for fighting misinformation designed to trick the system. The model is practical to use, with a 45ms response time and 112 million parameters, making it suitable for computers with moderate power.

Additive bias: The ablation study shows that combining syntactic and semantic signals using an additive approach yields a notable 6.0-point F1 improvement, from 86.5% to 92.5%. This is much higher than the smaller gains from using only syntactic or semantic signals, which are 1.9 and 2.1 points. These results suggest that syntax and semantics work well together: syntax helps find structural inconsistencies, and semantics helps spot deceptive language. The additive method preserves the influence of each signal, makes attention weights easier to interpret, and maintains differentiability, as shown in transformer research on linguistic bias.

Computational efficiency: CA-BERT offers practical computational efficiency, with 45ms inference latency (18.4% overhead compared to BERT-base and 36.6% faster than HANN), 22.2 tweets per second throughput, 2,847MB GPU memory usage (7.4% overhead), and 0.028J energy per document. Its energy efficiency is 6.4 times better per accuracy point than HANN (3.2% vs 20.5% cost per F1 gain). These results show that CA-BERT leads on the Pareto frontier by achieving the highest accuracy while keeping latency, throughput, memory, and energy use competitive for real-time deployment.

B. Limitations

CA-BERT has several limitations that should be addressed. Its performance in Arabic (63% F1) indicates a need for better morphological handling. Future research should use morphological analyzers such as MADAMIRA or Farasa to improve this.

The training data is imbalanced, with English accounting for 31.8% of the total, leading to a bias toward English. Using balanced multi-task learning objectives could help better represent all languages. The dataset is also limited to COVID-19, which makes it hard to apply the model to other types of misinformation. Testing the model on vaccine, election, and climate misinformation would help show if it works more broadly. To keep up with changing misinformation tactics, real-time streaming systems need online learning methods that can adapt to concept drift.

CA-BERT's $O(S^2)$ attention complexity makes it difficult for long-form misinformation. Its 45ms latency works well for tweets of about 256 tokens. Processing longer documents would take more time. Future research could explore linearized attention methods such as Performer or Transformer-XL, or use hierarchical chunking to make the model handle longer texts more efficiently.

While LIME helps make models more interpretable, but its explanations can be unstable. Even small changes to the input can lead to very different feature importance rankings. Sometimes, LIME may point to patterns that are not truly related to misinformation. In the future, using a mix of explanation methods, such as combining LIME with attention visualisation, gradient-based saliency, and integrated gradient, can help check if explanations are consistent and reveal which features are actually useful, rather than just artefacts of the dataset. Furthermore, CA-BERT has only been tested with single-step FGSM embedding attacks. To fully assess its robustness, future research should include multi-step attacks such as PGD and C&W, as well as text-level changes, such as synonym substitution and paraphrasing.

VIII. Conclusion

This work presents CA-BERT, a multilingual BERT (Bidirectional Encoder Representations from Transformers) model that uses attention mechanisms and explicit syntactic-semantic features—such as information about sentence structure and word meaning—to improve COVID-19 misinformation detection. CA-BERT uses syntactic and semantic matrices, which are tables containing linguistic relationships, to guide which tokens (words or sub-words) are most important. Tests show that it outperforms recent 2025 models, achieving a 92.5% macro-averaged F1 score, while staying interpretable with LIME (Local Interpretable Model-agnostic Explanations) explanations and robust against adversarial attacks. Multilingual tests reveal how the model performs across different languages, helping optimize it for various language types. Integrating syntactic-semantic structure with transformer models is an important advance in deep learning for detecting misinformation. An ablation study shows that using additive integration, combining the outputs of syntactic-semantic and transformer modules by summing their representations, leads to a 6.0 percentage-point F1 gain (from 86.5% to 92.5%), exceeding the sum of individual effects and supported by transformer research. CA-BERT remains practical for production, with 45ms latency (18.4% higher than BERT-base) and is 6.4 times more energy-efficient per F1 point than HANN. Using CA-BERT in fact-checking systems can help experts fight misinformation during public health and political crises, thanks to its interpretability and robustness.

Although CA-BERT performs well, it has some important limitations. Handling Arabic (63% F1) indicates a need for better morphological handling. Since 31.8% of the training data

is in English, the model may be biased, and multi-task learning could help fix it. The model was only tested on COVID-19 data, so it may not work as well for vaccine or election or other topics. Its quadratic attention limits it to short documents, but using linearized attention methods such as Performer or Transformer-XL could make it more efficient. LIME explanations can be unstable, so it is important to use several explanation methods. Future robustness tests should include multi-step adversarial attacks such as PGD and C&W, as well as text-level changes beyond FGSM.

Future research could use morphological analysers like MADAMIRA and Farasa to support low-resource languages (Arabic and other Asian languages). It should also apply balanced multi-task learning to fix English data imbalance, test cross-domain transferability on different types of misinformation, and use online learning to adapt to concept drift. Exploration of combining LIME method with attention visualization and gradient-based saliency should be considered. Adversarial robustness evaluation must be strengthened through multi-step gradient attacks (PGD, C&W) and text-level perturbations (synonym substitution, paraphrasing) to validate genuine robustness against realistic adversarial examples in production deployment. Finally, exploring linearised attention mechanisms such as Performer and Transformer-XL, or using hierarchical chunking, may help process long documents more efficiently.

Declarations

Compliance with Ethical Standards: Funding: This work was supported by Research Grant No. SPG/2020/000594 under the SERB POWER grant scheme, Science and Engineering Research Board, Government of India, to Akila Venkatesan, Pondicherry Engineering College, India.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability

The data used in this study are publicly available on GitHub at https://github.com/everythingguy/CovidMis20 and https://github.com/MickeysClubhouse/COVID-19-rumor-dataset.

Acknowledgment

We acknowledge the creators of the dataset for their contribution in providing this valuable resource for the research community.

Materials and Code Availability

The code and materials used in this study are available upon request from the authors.

Authors' Contributions

Krishna Kumar wrote the original draft, developed the Methodology, and conducted all experiments, including conceptualizing the study, designing the experimental framework, implementing the model, generating visualizations, and formally analyzing the data. Dr Akila Venkatesan supervised the project, reviewed the manuscript, identified potential technical errors, provided feedback and corrections, and offered guidance and expertise throughout the experimental phase to ensure the rigor and validity of the research.

References

[1] A. Unlu, S. Truong, N. Sawhney, and T. Tammi, "Setting the misinformation agenda: Modeling COVID-19 narratives in Twitter communities," New Media and Society, Feb. 2024, doi: 10.1177/14614448241232079.

[2] E. Gabarron, S. O. Oyeyemi, and R. Wynn, "Covid-19-related misinformation on social media: A systematic review," Bulletin of the World Health Organization, vol. 99, no. 6, pp. 455-463A, Jun. 2021, doi: 10.2471/BLT.20.276782.

[3] S. Tasnim, M. Hossain, and H. Mazumder, "Impact of rumors and misinformation on COVID-19 in Social Media," Journal of Preventive Medicine and Public Health, vol. 53, no. 3. Korean Society for Preventive Medicine, pp. 171–174, May 31, 2020. doi: 10.3961/JPMPH.20.094.

[4] S. J. Malla and P. J. A. Alphonse, "Fake or real news about COVID-19? Pretrained transformer model to detect potential misleading news," European Physical Journal: Special Topics, vol. 231, no. 18–20, pp. 3347–3356, 2022, doi: 10.1140/epjs/s11734-022-00436-6.

[5] E. Broda and J. Strömbäck, "Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review," Annals of the International Communication Association, vol. 48, no. 2, pp. 139–166, Apr. 2024, doi: 10.1080/23808985.2024.2323736.

[6] M. T. King, J.-C. Fu, M. Brown, and D. Santacaterina, "Rumor, Chinese Diets, and COVID-19," Gastronomica, vol. 21, no. 1, pp. 77–82, Feb. 2021, doi: 10.1525/gfc.2021.21.1.77.

[7] I. Pavela Banai, B. Banai, and I. Mikloušić, "Beliefs in COVID-19 conspiracy theories, compliance with the preventive measures, and trust in government medical officials," Current Psychology, vol. 41, no. 10, pp. 7448–7458, Oct. 2022, doi: 10.1007/s12144-021-01898-y.

[8] M. G. M. Kim, M. G. M. Kim, J. H. Kim, and K. Kim, "Fine-Tuning BERT Models to Classify Misinformation on Garlic and COVID-19 on Twitter," vol. 19, no. 9, p. 5126, 2022, Accessed: Feb. 22, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/35564518/

[9] T. Mickus, D. Paperno, M. Constant, and K. van Deemter, "What do you mean, BERT? Assessing BERT as a Distributional Semantics Model," Nov. 2019, doi: 10.7275/t778-ja71.

[10] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. Ali Almazroi, "A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification," Journal of Healthcare Engineering, vol. 2022, 2022, doi: 10.1155/2022/3498123.

[11] D. R. Faisal and R. Mahendra, "Two-Stage Classifier for COVID-19 Misinformation Detection Using BERT: a Study on Indonesian Tweets," pp. 1–29, Jun. 2022, [Online]. Available: http://arxiv.org/abs/2206.15359

[12] Z. Tan, Y. Tian, and J. Li, "GLIME: General, Stable and Local LIME Explanation," Advances in Neural Information Processing Systems, vol. 36, no. 16384, pp. 1–28, Nov. 2023, [Online]. Available: http://arxiv.org/abs/2311.15722

[13] Z. Boukouvalas et al., "Independent Component Analysis for Trustworthy Cyberspace during High Impact Events: An Application to Covid-19," Jun. 2020, Accessed: Jul. 04, 2024. [Online]. Available: http://arxiv.org/abs/2006.01284

[14] J. Ayoub, X. J. Yang, and F. Zhou, "Combat COVID-19 infodemic using explainable natural language processing models," Information Processing and Management, vol. 58, no. 4, p. 102569, Jul. 2021, doi: 10.1016/j.ipm.2021.102569.

[15] H. Ismail, N. Hussein, R. Elabyad, S. Abdelhalim, and M. Elhadef, "Aspect-based classification of vaccine misinformation: a spatiotemporal analysis using Twitter chatter," BMC Public Health, vol. 23, no. 1, pp. 1–14, 2023, doi: 10.1186/s12889-023-16067-y.

[16] R. A. Kinchla and R. C. Atkinson, "The effect of false-information feedback upon psychophysical judgments," Psychonomic Science, vol. 1, no. 1–12, pp. 317–318, Jan. 1964, doi: 10.3758/bf03342931.

[17] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," Neural Networks, vol. 13, no. 4–5, pp. 411–430, Jun. 2000, doi: 10.1016/S0893-6080(00)00026-5.

[18] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning Based Text Classification: A Comprehensive Review," vol. 1, no. 1, pp. 1–43, 2020, [Online]. Available: http://arxiv.org/abs/2004.03705

[19] Y. Luan and S. Lin, "Research on Text Classification Based on CNN and LSTM," in Proceedings of 2019 IEEE International Conference on Artificial Intelligence and Computer Applications, ICAICA 2019, Institute of Electrical and Electronics Engineers Inc., Mar. 2019, pp. 352–355. doi: 10.1109/ICAICA.2019.8873454.

[20] P. Liu, X. Qiu, and X. Huang, "Recurrent Neural Network for Text Classification with Multi-Task Learning," IJCAI International Joint Conference on Artificial Intelligence, vol. 2016-Janua, pp. 2873–2879, May 2016, [Online]. Available: http://arxiv.org/abs/1605.05101

[21] R. K. Kaliyar, A. Goswami, P. Narang, and V. Chamola, "Understanding the Use and Abuse of Social Media: Generalized Fake News Detection With a Multichannel Deep Neural Network," IEEE Transactions on Computational Social Systems, 2022, doi: 10.1109/TCSS.2022.3221811.

[22] M. K. Elhadad, K. F. Li, and F. Gebali, "An Ensemble Deep Learning Technique to Detect COVID-19 Misleading Information," in Advances in Intelligent Systems and Computing, vol. 1264 AISC, Springer, Cham, 2021, pp. 163–175. doi: 10.1007/978-3-030-57811-4_16.

[23] M. Arbane, R. Benlamri, Y. Brik, and A. D. Alahmar, "Social media-based COVID-19 sentiment classification model using Bi-LSTM," Expert Systems with Applications, vol. 212, no. January, p. 118710, Feb. 2023, doi: 10.1016/j.eswa.2022.118710.

[24] A. Miglani, "Coronavirus tweets NLP - Text Classification," kaggle. Accessed: Apr. 02, 2025. [Online]. Available: https://www.kaggle.com/datasets/datatattle/covid-19-nlp-text-classification

[25] H. Jelodar, Y. Wang, R. Orji, and S. Huang, "Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 10, pp. 2733–2742, Oct. 2020, doi: 10.1109/JBHI.2020.3001216.

[26] M. Bisi and R. Maurya, "Ensemble learning and stacked convolutional neural network for Covid-19 situational information analysis using social media data," Multimedia Tools and Applications, pp. 1–25, Feb. 2024, doi: 10.1007/s11042-024-18582-5.

[27] S. Chen, "Research on Fine-Grained Classification of Rumors in Public Crisis ——Take the COVID-19 incident as an example," E3S Web of Conferences, vol. 179, p. 02027, Jul. 2020, doi: 10.1051/e3sconf/202017902027.

[28] C. Yang, X. Zhou, and R. Zafarani, "CHECKED: Chinese COVID-19 fake news dataset," Social Network Analysis and Mining, vol. 11, no. 1, pp. 1–8, 2021, doi: 10.1007/s13278-021-00766-8.

[29] Y.-A. Chung et al., "w2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training," IEEE, 2021, pp. 244–250. doi: 10.1109/ASRU51503.2021.9688253.

[30] Z. Li, L. Zhou, X. Yang, H. Jia, W. Li, and J. Zhang, "User Sentiment Analysis of COVID-19 via Adversarial Training Based on the BERT-FGM-BiGRU Model," Systems, vol. 11, no. 3, 2023, doi: 10.3390/systems11030129.

[31] G. Shrivastava, P. Kumar, R. P. Ojha, P. K. Srivastava, S. Mohan, and G. Srivastava, "Defensive modeling of fake news through online social networks," IEEE Transactions on Computational Social Systems, vol. 7, no. 5, pp. 1159–1167, 2020, doi: 10.1109/TCSS.2020.3014135.

[32] G. Joshi et al., "Explainable Misinformation Detection Across Multiple Social Media Platforms," IEEE Access, vol. 11, no. February, pp. 23634–23646, 2023, doi: 10.1109/ACCESS.2023.3251892.

[33] M. Müller, M. Salathé, and P. E. Kummervold, "COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter," Frontiers in Artificial Intelligence, vol. 6, Mar. 2023, doi: https://doi.org/10.3389/frai.2023.1023281.

[34] W. Anggraeni, M. F. A. Kusuma, E. Riksakomara, R. P. Wibowo, Pujiadi, and S. Sumpeno, "Combination of BERT and Hybrid CNN-LSTM Models for Indonesia Dengue Tweets Classification," International Journal of Intelligent Engineering and Systems, vol. 17, no. 1, pp. 813–826, Jan. 2024, doi: 10.22266/ijies2024.0229.68.

[35] S. Saadah, Kaenova Mahendra Auditama, Ananda Affan Fattahila, Fendi Irfan Amorokhman, Annisa Aditsania, and Aniq Atiqi Rohmawati, "Implementation of BERT, IndoBERT, and CNN-LSTM in Classifying Public Opinion about COVID-19 Vaccine in Indonesia," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 6, no. 4, pp. 648–655, 2022, doi: 10.29207/resti.v6i4.4215.

[36] A. M. Almars, M. Almaliki, T. H. Noor, M. M. Alwateer, and E. Atlam, "HANN: Hybrid Attention Neural Network for Detecting Covid-19 Related Rumors," IEEE Access, vol. 10, pp. 12334–12344, 2022, doi: 10.1109/ACCESS.2022.3146712.

[37] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, 2017, pp. 5999–6009.

[38] M. Trigueros and B. Biahchini, "Learning Linear Transformations using models," Proceedings of INDRUM 2016: First conference of International Network for Didactic Research in University Mathematics, pp. 326–336, Mar. 2016, doi: 10.34894/VQ1DJA.

[39] A. Mulahuwaish, M. Osti, K. Gyorick, M. Maabreh, A. Gupta, and B. Qolomany, "CovidMis20: COVID-19 Misinformation Detection System on Twitter Tweets Using Deep Learning Models," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2023, pp. 466–479. doi: 10.1007/978-3-031-27199-1_47.

[40] M. Cheng, S. Wang, X. Yan, T. Yang, and W. Wang, "A COVID-19 Rumor Dataset," vol. 12, no. May, pp. 1–10, 2021, doi: 10.3389/fpsyg.2021.644801.