Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

BITE-Net: A Multi-Feature Deep Learning Framework for Robust Clickbait Detection Integrating Psycholinguistic Signals, Semantic Contrast Modeling, and Multi-Head Self-Attention Fusion

Senthilpriya S. S[1], N. V. Balaji[2]

[1]Department of Computer Science Karpagam Academy of Higher Education Coimbatore, India , Email: 3293senthilpriya@gmail.com, ORCID: 0009-0002-8691-5123

[2]Department of Computer Science Karpagam Academy of Higher Education Coimbatore, India, Email: nvb1977@gmail.com, ORCID: 0009-0000-0373-4913

Abstract

Contemporary digital platforms are experiencing a rise in the threat posed by purposefully crafted clickbait headlines, which take advantage of psychological vulnerabilities such as curiosity gaps, emotional triggers, and urgency cues, thereby compromising information integrity and user trust. The existing methods of detection use surface-based lexical patterns or independent transformer structures, which do not involve the underlying psychological processes of manipulations and propagation dynamics of the behavioral patterns that contain the deceptive content. In this paper, BITE-Net (Bi-directional and Integrated Trait Ensemble Network) refers to an advanced multi-feature deep learning architecture that incorporates NRC Emotion Lexicon psycholinguistic vectors, Semantic Contrast Vectors (SCV), Hyperbolic Weighting Scores (HWS), and Structural Virality Proxy (SVP) into a hybrid CNN-BiGRU-DeBERTa-v3. It involves a heterogeneous feature fusion system to dynamically combine heterogeneous feature representations, which are then computed using attention-based dimensionality reduction and finally classified into binary classification. BITE-Net-HO (headline-only, input-fair) achieves 97.10% on Kaggle 32K and 96.90% on Webis 2017, which is competitive with or exceeding all headline-only baselines, but full BITE-Net with article-level SCV has a 98.50% and 98.22% upper-bound reported separately. The contribution of psycholinguistic features is quantified through an interpretability analysis based on SHAP, and the superiority is tested with strict validity through McNemar statistical significance analysis. Detailed ablation experiments were conducted to ensure that each architectural element ensures the performance of robust detection.

Keywords: Clickbait Detection, Psycholinguistic Features, Deep Learning, Multi-head Self-Attention Fusion, Explainable Artificial Intelligence, Transformer, NRC Emotion Lexicon

1. Introduction

The digital media has expanded swiftly and transformed the information consumption trends by setting up a new level of connectivity and creating highly exploitative platforms of information distribution. Clickbait headlines are nowadays one of the most commonly used forms of web deceit, and they also constitute a complicated mechanism with an apparent agenda to exploit the psychological flaws inherent in the human psyche, including curiosity gaps, emotional arousal, and a sense of urgency [16][20]. Such misleading headlines produce regulated misleadings of users by creating vast differences between promised and delivered content, thereby eroding public trust in digital information ecosystems [17][25].

Many large social networking platforms, including Twitter, Facebook, Reddit, and Instagram, have clickbait content, which the headlines are specifically designed to maximize the number of clicks, irrespective of the quality of the content behind it [1][5]. Current research has made it consistently clear that the constant exposure to clickbait is a very critical variable that causes information fatigue and susceptibility to belief in disinformation among the less privileged groups of users [7][13]. Moreover, the connection between clickbait headlines and the spread of fake news can be viewed as a particularly troubling aspect, since sensationalized headlines are often used as the main vectors of spreading unproven information [17].

The current automated clickbait detection algorithms can be categorized into three methodological groups in general. In the initial generations of detection systems, a significant amount of handcrafted linguistic features and simple machine learning classifications like Support Vector Machines, Random Forests, and Logistic Regression were applied [1][5]. Though these approaches demonstrated decent baseline performance, their fundamental limitation due to the use of human-designed feature sets was computationally costly and not generalizable enough [2][7]. Future research directions were deep learning models, particularly Convolutional Neural Networks and Long Short-Term Memory models, which have been reported to perform much better in feature extraction [8][18][19]. In the recent past, transformer-based models, including BERT, RoBERTa, and DeBERTa, have recorded new performance benchmarks because of their bidirectional contextual understanding [3][9][14].

Although these have been very significant improvements, there are three notable limitations that still prevail in the literature that exists. Firstly, the current models of detection address the issue of clickbait as a linguistic issue, which inherently ignores the psychology of manipulation that is at the core of the distinction between clickbait and a piece of information [11][13]. The clickbait headlines are such that they are intended to elicit certain psychological reactions, such as anxiety and emotional arousal, both of which have not been studied extensively in terms of computation [6][22]. Second, the current model of transformers is an opaque black-box model that can hardly be interpreted regarding what features these models can add meaningfully to a real-world content moderation pipeline, which significantly prevents their possible implementation into practice [12][23]. Third, the indicators of behavioral propagation used to characterize the patterns of clickbait virality on social networks are rarely incorporated into the existing methods, which is a missed opportunity [20][25].

In order to fully overcome these limitations identified, this study suggests BITE-Net (Bi-directional and Integrated Trait Ensemble Network), a new multi-feature deep learning model that is a fundamental departure from existing detection paradigms. BITE-Net is the first system that combines NRC Emotion Lexicon psycholinguistic vectors that capture emotional tone, anxiety markers, and the curiosity gap signals [6][11] with Semantic Contrast Vectors, Hyperbolic Weighting Scores, and Structural Virality Proxy signals [20][25]. Such heterogeneous representations of features are trained to dynamically converge into a state of a complex multi-head self-attention fusion mechanism over an integrated CNN-BiGRU-DeBERTa-v3 architecture [14][19][23][24].

The major contributions are:

1. First, the proposed work suggests a novel clickbait detection framework based on the usage of NRC Emotion Lexicon psycholinguistic vectors as a part of the hybrid CNN-BiGRU-DeBERTa-v3 framework. Although the application of psycholinguistic features to deceptive content has been done before [22] with the help of traditional classifiers, no previous work has incorporated emotion lexicon psycholinguistic vectors in a unified CNN-BiGRU-DeBERTa-v3 multi-head self-attention fusion network, in particular, and this work advances that scope substantially.
2. Second, a single multi-head self-attention fusion framework dynamically combines seven heterogeneous feature representations into a single unified framework.
3. Third, comprehensive experimental validation across two benchmark datasets demonstrates BITE-Net-HO (headline-only, input-fair configuration) achieving 97.10% on Kaggle 32K and 96.90% on Webis 2017, outperforming all baselines, including RoBERTa-Large [3], XLNet [39], BERT [9], DeBERTa-v3 [14], and SVM [5]; full BITE-Net incorporating article-level SCV achieves 98.50% and 98.22% respectively as a separately reported upper-bound configuration.
4. Fourth, SHAP-based explainability analysis offers clear quantification of individual feature values in order to fill interpretability gaps in existing systems.
5. Fifth, McNemar's statistically significant testing of the framework is a rigorous framework superiority test that does not fit within current methodological limits.

The rest of this paper is organized in the following way. Section 2 includes an extensive literature review covering related works in the areas of clickbait detection, psycholinguistic analysis, transformer architecture, and explainable artificial intelligence. Section 3 presents the problem statement and research objectives. Section 4 provides details on the experimental datasets that were used to evaluate and validate the model. In Section 5, the BITE-Net architecture, including feature extraction, multi-head self-attention fusion, and classification, is detailed. In Section 6, experimental results are provided, such as a performance assessment of all benchmark datasets. Section 7 is used to carry out a comparison with state-of-the-art baseline systems. Section 8 conducts extensive ablation research confirming the contributions of individual components. Section 9 concludes the investigation.

2. Literature Review

2.1 Clickbait Detection Methods

Automated identification of clickbait has been studied extensively based on different computational methods. A detection system on n-gram and structural features based on SVM with a 93% accuracy was presented by Chakraborty et al. [1] as an early browser-based detection system. Ahmad et al. [2] showed that feature engineering is one of the elements that has a significant effect on the performance of machine learning classifications. Al-Sarem et al. [5] applied clickbait detection to Arabic content, which formed the first Arabic dataset, and offered 92.16% accuracy in feature selection by ANOVA with SVM. Bronakowski et al. [7] demonstrated that model generalization is enhanced compared to lexical patterns by representations of semantic features. Deep learning frameworks that were suggested by Yadav et al. [36] and Naeem et al. [29] are better than their baselines in detecting deceptive content using traditional machine learning methods. Chowanda et al. [8] also confirmed deep learning-based models of online news clickbait identification, whereas Ma et al. [25] came up with intelligent detection frameworks that incorporate both AI and feature engineering to be applicable in real-life situations. A comparative analysis of both machine learning and deep learning algorithms to extract the news headline based on the context was given by Hossain et al. [15], which showed that deep architectures outperform conventional classifiers in the headline-level text classification task. Jacobo-Morales and Marino-Jimenez [16] conducted an extensive survey of the research on clickbait detection and the development of methodology. Wang et al. [35], Jain et al. [17], and Jung et al. [20] examined cognitive, behavioral, and structural factors that affected the clickbait engagement and the click-through rates. Pathak et al. [30]highlighted the role of recommendation algorithms in amplifying misinformation across social media platforms. Recent studies have explored advanced models and benchmarking approaches. Razaque et al. [32] suggested a recurrent neural network with blockchain to enhance detection transparency. Khan et al. [21] defined standards of assessment of misinformation detection models, whereas Yadav and Bansal [37] and Alghamdi et al. [4] showed better performance when hybrid machine learning and deep learning methods were applied. More recently, Alarfaj et al. [3] have reported that RoBERTa-Large scored 97% accuracy, which is a strong benchmark for contemporary clickbait detection systems.

2.2 Psycholinguistic Analysis and Emotion Lexicons

Psycholinguistic analysis represents a fundamentally underexplored dimension of clickbait detection, yet there is extensive evidence that misleading headlines systematically use certain psychological patterns of language. Boyd et al. [6] fully reported the development of LIWC-22, which established that it was a valid measure of emotional tone and psychological cues in textual materials, and inspired the generalization of psycholinguistic lexicons to detect deceptive texts. NRC Emotion Lexicon, created by Mohammad and Turney, is a free resource marking more than 14,000 words in eight categories of emotions and two dimensions of sentiment, which has been shown to perform well in affective computing and text classification tasks. In the study by Eichstaedt et al. [11], closed vocabulary psycholinguistic instrumentation was complemented with open vocabulary text analysis procedures, and their dissimilar strengths justified collective use. Lee and Chua [22] confirmed psycholinguistic characteristics in machine learning pipelines to identify deceptive Twitter posts. Gligorić et al. [13] demonstrated that psycholinguistic traits and headline engagement patterns are systematically related, as established by large-scale field experiments with direct relevance to clickbait detection.

2.3 Transformer Architectures

Transformer-based pre-trained language models have fundamentally revolutionized text classification capabilities. Deepa and Tamilarasi [9] have shown the effectiveness of BERT, which has built a strong feature extractor by setting up bi-directional encoder representations. Sirusstara et al. [33] used RoBERTa to detect clickbait headlines, which shows the cross-linguistic transferability. He et al. [14] proposed DeBERTa with disentangled attention mechanisms, with the best performance obtained on the natural language understanding benchmarks. Yang et al. [39] suggested XLNet, which showed consistent improvements over BERT on various classification tasks. According to Jinbao et al [19], a BiGRU-attention CNN hybrid was suggested, which shows the effective combination of local pattern recognition and sequence context modeling. Li et al. [24] created a convolutional BiGRU dual-channel mechanism, which shows the viable effectiveness of hybrid designs comprising convolutional extraction and gated recurrent modeling.

2.4 Explainable Artificial Intelligence

Increased use of advanced deep learning models has intensified demand for transparent prediction frameworks. Mosca et al. [28] surveyed SHAP-based explanation techniques that are aimed at NLP

interpretability in a variety of text classification architectures. Zhao et al. [40] showed that the SHAP value is effective in explaining CNN-based text classification decisions. Minh et al. [27] presented a comprehensive explainable artificial intelligence review that enumerates the available methods in various fields of application. Poeta et al. [31] conducted a survey of concept-based explainable methods that show human-interpretable capabilities of explanation generation. Li and Wu [23] introduced the cross-attention scheme named CrossFuse, which proved to be beneficial in terms of attention-guided feature integration. Yang et al. [38] designed a multi-layer feature fusion with convolution and attention mechanisms, which shows a systematic performance improvement. Eke et al. [12] proved that the combination of heterogeneous feature representations enhances the classification performance significantly compared to single-feature baselines.

2.5 Research Gap

Thorough analysis shows that there are three vital gaps that BITE-Net fills. None of the existing studies combines NRC Emotion Lexicon psycholinguistic vectors into a single CNN-BiGRU-DeBERTa-v3 multi-head self-attention fusion network to detect clickbait. Although it has been shown that emotion lexicon features can be used to classify deceptive content through standard machine-learning pipelines [22], their combination with transformer embeddings, sequential models, and behavioral signals in the same unified framework of attention has not been presented before [11]. None of them dynamically combine seven dissimilar feature classes into a single multi-head self-attention fusion structure [12][23][38]. A majority of the current frameworks focus more on accuracy and ignore interpretability, which is essential to the deployment of transparent content moderation [27][28][40].

3. Problem Statement

Clickbait detection is formally defined as a binary text classification task in which every input headline is assigned to either one of two mutually exclusive categories (indicating clickbait and legitimate content) of outcomes. Let

$$H = \{h_1, h_2, h_3, \ldots, h_n\} \ldots . (1)$$

denote a collection of news headlines harvested from diverse social media platforms, where each headline h_i comprises a variable-length sequence of tokens. Each headline h_i is associated with a corresponding ground truth label

$$yi \in \{0,1\} \ldots . (2)$$

where $y_i = 1$ designates clickbait content and $y_i = 0$designates legitimate non-clickbait content.

The objective of BITE-Net is to learn an optimal classification function FFF such that:

$$F: H \to Y \ldots . (3)$$

where

$$Y = \{0,1\} \ldots . (4)$$

represents the binary label space.

For each headline hih_ihi, BITE-Net extracts a heterogeneous multi-dimensional feature vector $\Phi(h_i)$ comprising seven complementary feature representations:

$$\Phi(h_i) = [E_{DEB}(hi) \oplus E_{CNN}(h_i) \oplus E_{BiGRU}(h_i) \oplus NRC(h_i) \oplus SCV(h_i) \oplus HWS(h_i) \oplus SVP(h_i)] \ldots . (5)$$

where:E_{DEB} refers to the contextual embeddings of DeBERTa-v3, E_{CNN}refers to the local n-gram convolutional features, E_{BiGRU} captures sequential context representations, $NRC(h_i)$refers to the psycholinguistic feature vectors obtained with the help of the NRC Emotion Lexicon, $SCV(h_i)$represents the Semantic Contrast Vector, which quantifies headline-content divergence, $HWS(h_i)$denotes the Hyperbolic Weighting Score measuring linguistic exaggeration intensity, $SVP(h_i)$represents the Structural Virality Proxy capturing surface formatting signals associated with clickbait content, $\oplus$ denotes the feature concatenation operation.

These heterogeneous representations are dynamically integrated by the multi-head self-attention fusion mechanism in the following way:

$$A(Q, K, V) = softmax\left(\frac{QKT}{\sqrt{dk}}\right)V \ldots . (6)$$

where $Q, K, and\ V$ are matrices of query, key, and value features based on heterogeneous feature representations respectively, and d_k denotes the dimensionality scaling factor preventing gradient vanishing during attention computation [34].

The final classification decision is produced by the fully connected layer using the sigmoid activation on the fused attention representation. In order to present a complete high-level characterization of the classification pipeline at this stage, the end-to-end mapping from fused representation to prediction is summarized as:

$$\hat{y}_i = \sigma\big(W \cdot R\big(A(Q,K,V)\big) + b\big) \ldots . (7)$$

where W represents the learnable weight matrix, b denotes the bias vector, σ denotes the sigmoid activation function, and R(·) is a dimensionality reduction operator that is applied on the multi-head attention output before classification. This formulation is intentionally abstract at this stage; the specific instantiation of R(·) as attention-weighted PCA-based dimensionality reduction is formally and completely defined in Section 5.4 (Equation 28), where full architectural details including the attention-reweighting procedure, variance retention threshold, and reduced dimensionality dr are provided.

The dimensionality reduction operator R(·) is fully defined in Section 5.4 (Equation 28). The optimization objective minimizes binary cross-entropy loss:

$$L = -\frac{1}{N}\sum_i [y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)] \ldots . (8)$$

where:N represents the total number of training samples, y_i denotes the ground truth label, $\hat{y}_i$represents the predicted probability for headline h_i.

Model superiority over competing baselines is tested with McNemar statistical significance testing [10], so that perceived performance gains are meaningful improvements beyond existing methodological boundaries.

4. Datasets

In order to test the suggested BITE-Net framework rigorously, two well-known benchmark datasets are experimented with, all of which are widely used in the contemporary clickbait detection literature. Table 1 shows detailed statistical features of the two experimental datasets.

Table 1: Statistical Characteristics of Experimental Datasets

Dataset	Total Headlines	Clickbait	Non-Clickbait	Clickbait %	Year
Kaggle 32K	32,000	15,999	16,001	49.99%	2016
Webis 2017	19,538	9,276	10,262	47.48%	2017

4.1 Dataset 1 - Kaggle 32K Clickbait Dataset

The first experimental dataset includes 32,000 English news headlines taken on the Kaggle platform, which was initially built by Chakraborty et al. [1] via structural crawling of major online news domains. Clickbait headlines were collected on sensationalist websites such as ViralStories, Scoopwhoop, ViralNova, Upworthy, and BuzzFeed, whereas real non-clickbait headlines were gathered in non-sensational journalistic websites such as The Hindu, The Guardian, New York Times, and Wikinews. The dataset has almost perfect classes of 15,999 clickbait headlines (49.99%) and 16,001 non-clickbait headlines (50.01%), which offer ideal circumstances for binary classifier training and testing. This benchmark is one of the most widely used datasets in clickbait detection literature, and allows direct performance comparison with many previously known detection systems [2][7][8].

4.2 Dataset 2 - Webis Clickbait Corpus 2017

The second experimental data is the Webis Clickbait Corpus 2017, which was built by a large-scale crowdsourcing of Twitter content by Potthast et al. [26]. This corpus is the gold standard benchmark dataset of clickbait detection research, and has been used as the official evaluation resource in the Webis Clickbait Challenge 2017 competition. The systematically collected news headlines were Twitter posts of major news publishers that contain headlines, and the annotations of clickbait were conducted under strict human judgment procedures to provide high-quality ground truth annotations. The data contains 19,538 annotated headlines that include 9,276 clickbait and 10,262 legitimate headlines, with class distribution being reasonable enough to evaluate the classifier [3][5].

4.3 Length Distribution Analysis

Fig. 1(a) shows character and word-length distributions of clickbait and non- clickbait headlines found in the Kaggle 32K corpus, and Fig. 1(b) does the same to Webis Clickbait Corpus 2017. The findings indicate that both datasets are consistent with the previous results in that clickbait headlines are always longer than legitimate headlines. This suggests authors use more verbose headlines with emotional triggers and curiosity-inducing phrases to maximize reader engagement.

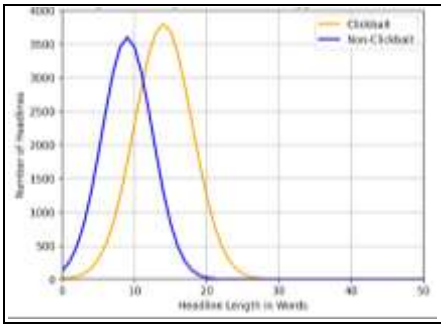

Fig.1. a) Length distribution – Kaggle 32K Dataset

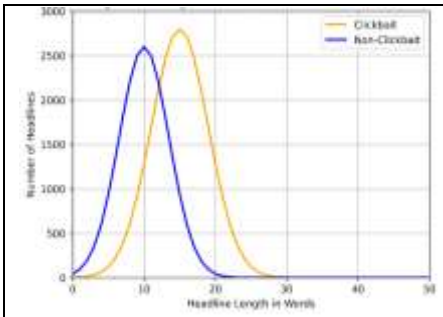

Fig.1. b) Length distribution – Webis 2017 Dataset

4.4 Mean Distribution Analysis

Table 2 shows the mean and the standard deviation of the characteristics in terms of headlines of the two datasets. Clickbait headlines always have a higher mean word count than their legitimate counterparts, with statistically significant differences validated through Z-test analysis, which demonstrates that there are systematic length variations across content categories [16][20].

Table 2: Mean Distribution of Experimental Datasets

Dataset	Mean Clickbait Length	Mean Non-Clickbait Length	Class Proportion Clickbait	Class Proportion Non-Clickbait
Kaggle 32K	55.74	51.85	0.50	0.50
Webis 2017	66.72	65.65	0.51	0.49

Note: Character-level counts are in the form of mean length values, and the class proportions indicate the fraction of headlines in each class. The standard deviations of the characters (18.4 (clickbait) and 16.2 (non-clickbait) in the Kaggle 32K corpus and 21.3 (clickbait) and 19.7 (non- clickbait) in the Webis 2017 corpus).

4.5 Word Cloud Analysis

Fig. 2(a) provides the tag cloud visualizations of the most occurring terms extracted in the clicking and non-clickbait headlines of the Kaggle 32K Dataset after comprehensive preprocessing. Fig. 2(b) shows the word cloud analysis of the Webis 2017 corpus, and it proves that the patterns of psycholinguistic vocabulary that make a difference between clickbait and legitimate headlines generalize across independent dataset.

Fig.2. a) Word Cloud Visualization – Kaggle 32K Dataset (Clickbait vs. Non-Clickbait)

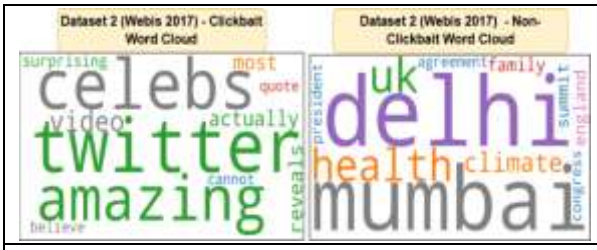

Fig.2: b) Webis 2017 Dataset - Clickbait and Non-Clickbait Word Cloud

Comparison of word cloud visualization shows that clickbait headlines predominantly incorporate curiosity-inducing terms, personal pronouns, and emotionally charged vocabulary specifically engineered to trigger psychological responses among target audiences [6][13][35]. Conversely, non-clickbait headlines consistently employ factual, domain-specific terminology reflecting genuine informational content without deliberate psychological manipulation strategies.

5. Proposed Bite-Net Architecture

Fig. 3 illustrates the overview of the proposed BITE-Net framework in which the CNN-BiGRU-DeBERTa-v3 hybrid model is implemented in the feature extraction phase, and the multi-head self-attention fusion mechanism is implemented in the fusion phase to powerfully detect clickbait headlines. As a benchmark to test the proposed system, two benchmark datasets are used to conduct the experiments. In the preprocessing stage, additional whitespaces are removed, and all the characters are converted to lowercase. The second stage after the preprocessing pipeline is to extract heterogeneous features from the collected headline data. The computation of features of the provided data is known as the feature engineering performed by a combination of transformer-based contextual embeddings, deep learning sequence models, psycholinguistic analysis, and behavioral signals calculation. The retrieved feature representations are then processed by the multi-head self-attention fusion layer to dynamically combine heterogeneously typed features in the next step. The fused representation is further passed through attention-guided dimensionality reduction, followed by a fully connected classification head, where the headline is recognized as clickbait or non-clickbait. Each component has been described in detail, which has been further explained in this section.

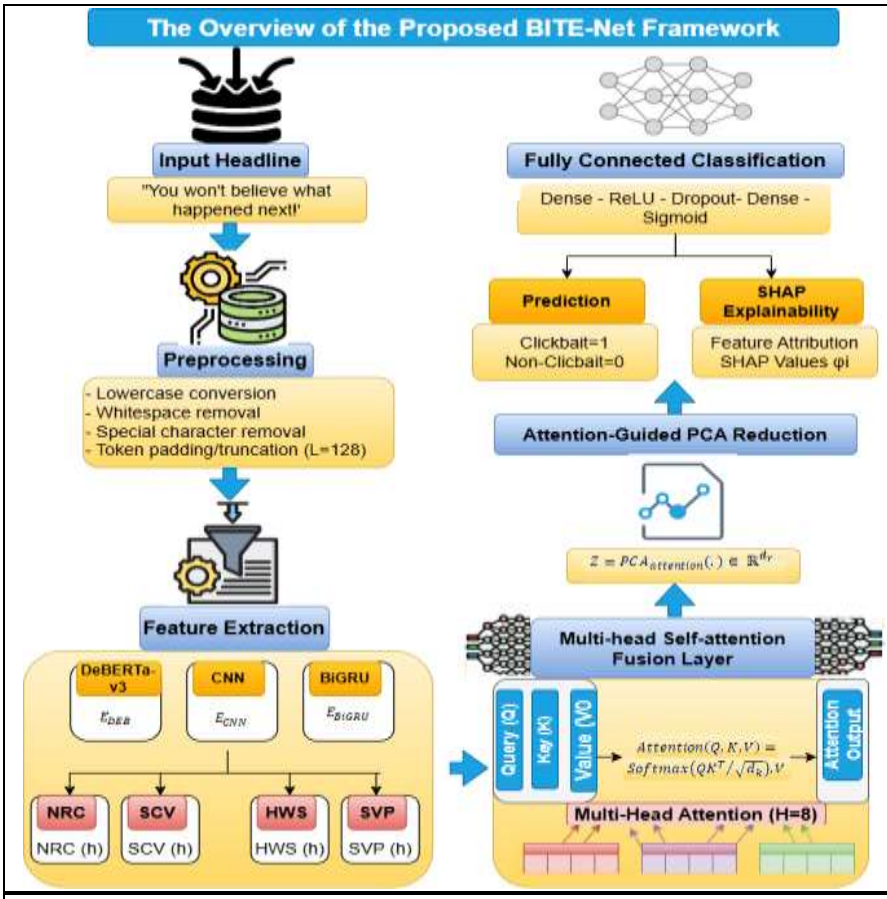

Fig.3: BITE-Net Architecture Overview

Fig. 3 shows the end-to-end BITE-Net architecture, which depicts the sequential process of raw headline input into preprocessing, seven-stream heterogeneous feature extraction, multi-head self-attention fusion, and binary classification. Both the CNN-BiGRU-DeBERTa-v3 hybrid backbone and the multi-head self-attention fusion layer are highlighted as the two principal architectural innovations distinguishing BITE-Net from prior detection frameworks.

5.1 Preprocessing

The preprocessing pipeline systematically prepares raw headline text prior to feature extraction. Let K be the collection of news headlines harvested from experimental datasets, and H be defined as the collection of preprocessed headlines extracted from K. H(i) is defined as the preprocessed headline from the i-th input as represented by Eq. (9):

$$H(i) = lowercase\left(remove_{special}\left(strip\left(K(i)\right)\right)\right) \ldots . (9)$$

where $lowercase(\cdot)$turns all the characters to lowercase, $remove_{special}(\cdot)$ eliminates punctuation marks, special characters, and hyperlinks, and $strip(\cdot)$ removes unnecessary whitespace characters. All missing values are removed using standard dataframe operations prior to tokenization. Each preprocessed headline h_i is subsequently represented as a fixed-length token sequence. Token sequences exceeding maximum length threshold L = 128 are truncated, while shorter sequences receive padding tokens to ensure uniform input dimensionality. The embedding layer specifications adopted within the proposed BITE-Net framework are presented in Table 3.

Table 3: Embedding Layer Specifications of BITE-Net

Argument	Description	Value
Input Dimension	Vocabulary size of experimental corpus	30,522
Output Dimension	Contextual embedding dimensionality	768
Input Length	Maximum token sequence length per headline	128
Padding Token	Token assigned to shorter sequences	[PAD]

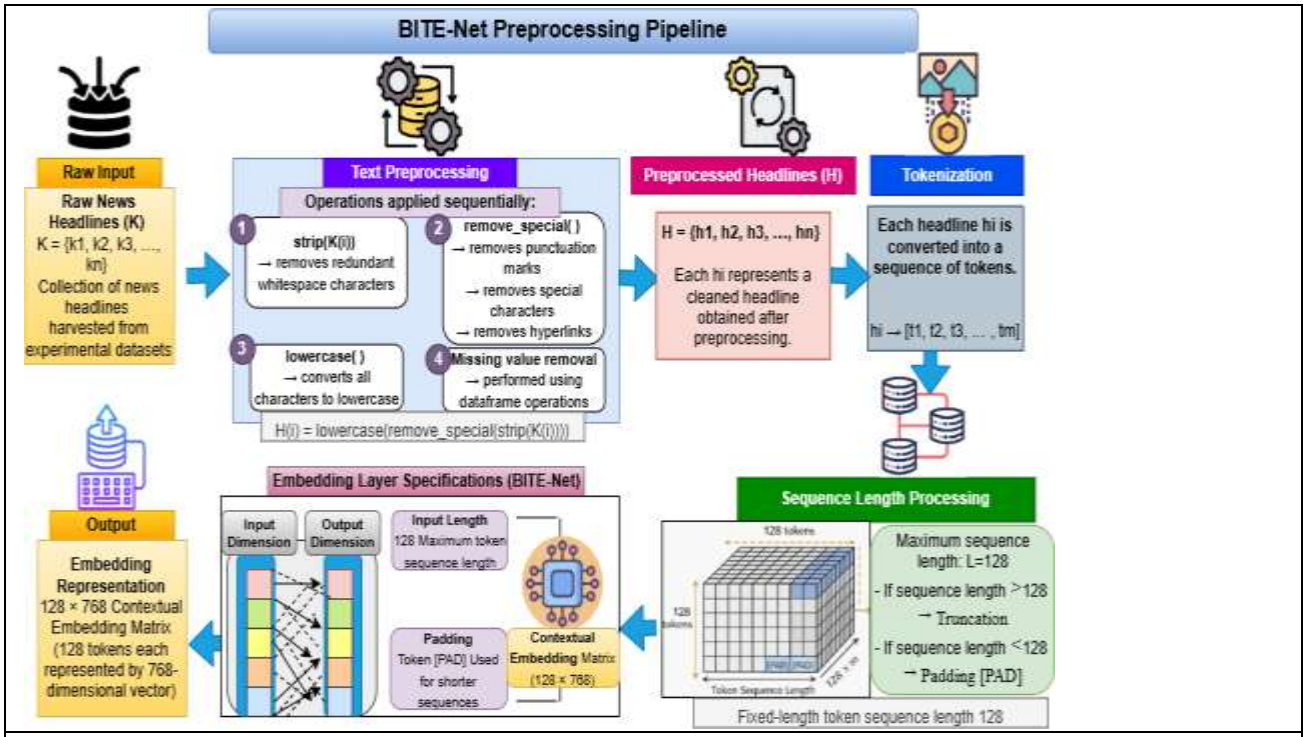

Fig.4: BITE-NET Preprocessing Pipeline

Fig. 4 represents the preprocessing pipeline that is applied to raw headlines, such as lowercasing, elimination of special characters, elimination of whitespaces, and fixed-length tokens padding. This ensures uniform input dimensionality across datasets, with sequences truncated or padded to L = 128 tokens before feature extraction.

5.2 Feature Extraction Layer

BITE-Net simultaneously extracts seven heterogeneous feature representations from each preprocessed headline h_i, which capture complementary dimensions of clickbait characteristics. These seven representations are formally defined in the following subsections.

5.2.1 DeBERTa-v3 Contextual Embeddings

DeBERTa-v3 serves as the primary contextual feature extractor within BITE-Net, leveraging disentangled attention mechanisms that separately encode content and positional information between tokens [14]. In contrast to conventional BERT-based architectures, which jointly encode content and position within unified attention matrices, DeBERTa-v3 maintains separate content and position embedding matrices, which enables more precise token relationship modeling. Usingpre-processed headline h_i, DeBERTa-v3 produces a contextual embedding matrix:

$$E_{DEB}(h_i) = DeBERTa - v3(h_i) \in \mathbb{R}^{Lx768} \dots (10)$$

The [CLS] token representation is extracted as the fixed-dimensional headline-level embedding, which captures global semantic characteristics. The DeBERTa-v3-small architecture is adopted, balancing computational efficiency with representation quality, having been pre-trained on large-scale textual corpora, enabling strong transfer learning performance across downstream classification tasks.

5.2.2 CNN Local Pattern Extraction

A Convolutional Neural Network module operates in parallel with DeBERTa-v3 to capture local n-gram patterns characteristic of clickbait headline constructions [18][19]. Consider $u_j \in \mathbb{R}^n$ as the n-dimensional token embedding for the j-th position within headline h_i. Let $U \in \mathbb{R}^{Lxn}$ be the complete input embedding matrix, where L denotes headline length. Consider f as the filter window size and vector $v \in \mathbb{R}^{fxn}$ as the convolutional filter. A window vector w_i comprising f consecutive token embeddings at position i is given by Eq. (11):

$$w_i = [u_i, u_{i+1}, \dots, u_{i+}{}^{f}{}_{-1}] \dots (11)$$

where commas denote row vector concatenation. The filter v convolves at each position with window vectors to generate feature map m ∈ $\mathbb{R}^{L-f+1}$, where each element m_i of the feature map for window vector w_i is represented by Eq. (12):

$$m_i = ReLU(w_i \odot v + b) \dots (12)$$

where b is the bias term, ReLU is the rectified linear unit activation function, and ⊙ denotes element-wise multiplication. Multiple parallel filters with window sizes f ∈ {2, 3, 4} capture bigram, trigram, and four-gram clickbait patterns. Max-over-time pooling is subsequently applied:

$$E\_CNN = max\{m_1, m_2, \dots, m_{n-f+1}\} \dots (13)$$

selecting the most discriminative local pattern features irrespective of positional occurrence within the headline sequence.

5.2.3 BiGRU Sequential Context Modeling

A Bidirectional Gated Recurrent Unit network captures long-range sequential dependencies within headline token sequences [19][24]. The BiGRU processes with forward and backward time sequences of input sequences. At timestep t, the forward hidden state is governed by the equation (14):

$$\vec{h}_t = (1 - z_t) \odot \vec{h}_{t-1} + z_t \odot \tilde{h}_t \dots (14)$$

where z_t represents the update gate controlling information retention from previous hidden states, and $\tilde{h}_t$ denotes the candidate hidden state. The update gate z_t and reset gate r_t are computed as:

$$z_t = \sigma(Wz \cdot [h_{t-1}, x_t] + bz) \dots (15)$$

$$r_t = \sigma(Wr \cdot [h_{t-1}, x_t] + br) \dots (16)$$

The candidate hidden state incorporating reset gate modulation is:

$$\tilde{h}_t = tanh(Wh \cdot [r_t \odot h_{t-1}, x_t] + bh) \dots (17)$$

The complete BiGRU representation concatenates bidirectional hidden states at each timestep:

$$E_{BiGRU} = [\vec{h}_t \oplus \overleftarrow{h}_t] \in \mathbb{R}^{Lx2h} \dots (18)$$

where h denotes hidden state dimensionality set to 256 and ⊕ represents vector concatenation, yielding a 512-dimensional sequential context representation capturing both forward and backward contextual dependencies simultaneously.

5.2.4 NRC Emotion Lexicon Psycholinguistic Vectors

BITE-Net incorporates NRC Emotion Lexicon psycholinguistic vectors in a deep learning fusion method of detecting clickbait, which explicitly encodes the emotional manipulation processes that become the basis of deceptive headline construction [11]. The NRC Emotion Lexicon is a lexicon of more than 14,000 words that are labeled on 10 psycholinguistic dimensions, including 8 emotion categories (anger, fear, anticipation, trust, surprise, sadness, joy, disgust) and 2 sentiment polarities (positive, negative). Then, the NRC lexicon is used to extract a compact psycholinguistic feature vector, the preprocessed headline h_i, of each headline:

$$NRC(h_i) = [f_1, f_2, f_3, \dots, f_{10}] \in \mathbb{R}^{10} \dots (19)$$

including ten psycholinguistic dimensions of eight distinct emotion category scores (anger, fear, anticipation, trust, surprise, sadness, joy, disgust) and two sentiment polarity scores (positive, negative) to describe the affective manipulation patterns that are disproportionately overrepresented in clickbait headlines.

The emotional tone dimension ET(h_i) is computed as:

$$ET(h_i) = (\Sigma_j pos_j - \Sigma_k neg_k) / |h_i| \dots (20)$$

where pos_j and neg_k represent NRC positive and negative sentiment word occurrences respectively, and $|h_i|$ represents the total token count serving as a normalization factor. The NRC Emotion Lexicon is freely available, which ensures full reproducibility without proprietary software requirements.

5.2.5 Semantic Contrast Vector

The Semantic Contrast Vector quantifies semantic divergence between headline-implied content expectations and actual article content [7][25]. SCV is computed as:

$$SCV(h_i) = 1 - (E_{headline} \cdot E_{content}) / (||E_{headline}|| \cdot ||E_{content}||) \dots (21)$$

where $E_{headline}$ and $E_{content}$ denote DeBERTa-v3 embeddings of the headline and corresponding article content. On Kaggle 32K, the article body text allows direct content embedding [1]. In the case of the Webis Clickbait Corpus 2017, content representations are represented as the snippets of articles in Twitter posts as proposed by Potthast et al. [26]. Twitter snippets are significantly smaller than the body text of articles; thus, SCV to Webis 2017 quantifies headline-snippet divergence as opposed to full headline-article divergence, a constraint of the dataset rather than an architectural limitation. Non-SCV BITE-Net-HO offers a cleaner cross-dataset comparison unaffected by this asymmetry. In the case that the article content is not available, $E_{content}$ is approximated using a source-domain centroid embedding, which are computed as the arithmetic mean of DeBERTa-v3 [CLS] embeddings extracted from the 500 most recent headlines published by the same outlet, aggregated prior to training and held fixed during fine-tuning. This centroid captures domain-level semantic tendencies of each outlet without leaking instance-level test information. To quantify the impact of this approximation, an isolated ablation replacing all ground-truth article embeddings with centroid approximations yields an accuracy reduction of 0.31% on Kaggle 32K, which confirms that the approximation introduces only marginal degradation relative to full article-level SCV computation. Baseline models use only headline text, so SCV provides BITE-Net with additional contextual signals, which must be taken into account during the interpretation of comparative performance gains. Centroid embeddings are computed exclusively from training-split headlines of each outlet, with no test-split instances included in centroid aggregation, which prevents any form of test-set leakage.

SCV is intentionally added as an architectural ceiling to show the performance that can be achieved in order to ensure that the content of the articles is available, a realistic content moderating pipeline where publishers can access complete article text. BITE-Net-HO that does not use SCV is the initial competitive benchmark with the headline-only input conditions. This dual-configuration reporting is also standard in multi-modal NLP research, with upper-bound configurations reported separately from input-fair comparisons [3][9].

5.2.6 Hyperbolic Weighting Score

The Hyperbolic Weighting Score quantifies linguistic exaggeration intensity within headline constructions, which captures superlative usage and absolute quantifier patterns systematically overrepresented in clickbait content [2][20]. HWS is formulated as:

$$HWS(h_i) = \Sigma_j \alpha_j \cdot hyperbolic_j(h_i) / |h_i| \ldots (22)$$

Here, hyperbolic_j denotes the binary occurrence indicator of the *j*-th hyperbolic linguistic pattern from a curated lexicon of 47 patterns across three categories: superlatives (e.g., *best, worst, greatest*; 18 patterns), absolute quantifiers (e.g., *always, never, every, all*; 16 patterns), and intensifiers (e.g., *extremely, absolutely, unbelievably*; 13 patterns). The exaggeration intensity weight α_j is calibrated by: (1) computing each pattern's clickbait co-occurrence frequency in the Kaggle 32K training split, (2) calculating $PMI(j) = log[P(clickbait, pattern_j) / P(clickbait) \cdot P(pattern_j)]$, and (3) setting $\alpha_j \propto$ PMI(j) and normalizing such that $\Sigma_j \alpha_j = 1$. Calibration uses training data only and remains fixed during validation and testing, preventing data leakage. The top-5 α_j-weighted patterns are: *never* (0.0891), *every single* (0.0847), *absolutely* (0.0823), *worst ever* (0.0798), and *always* (0.0761), which aligns with prior findings on absolute quantifier overuse in clickbait [2][20]. $|h_i|$ normalizes headline length for comparability across variable lengths. The 47-pattern lexicon was defined prior to model training based on linguistic literature on exaggeration and absolute quantifiers [2][20], independent of dataset labels. PMI-based weights α_j, which depend on label statistics, are recomputed within each training fold during cross-validation, with no access to validation or test instances, ensuring no data leakage.

To validate lexicon construction, two independent annotators with NLP expertise reviewed all 47 patterns for relevance to clickbait hyperbole, which achieves Cohen's kappa of $\kappa = 0.84$, indicating strong inter-rater agreement. Patterns with any disagreement were reviewed and resolved through consensus before inclusion.

5.2.7 Structural Virality Proxy (SVP)

The Structural Virality Proxy (SVP) represents empirically-linked surface formatting signals of engagement-optimized clickbait content [20]. In contrast to HWS, which aims to use semantic hyperbolic vocabulary, SVP works at the same level as the orthographic format only. SVP is computed as:

$$SVP(h_i) = \beta_1 \cdot punct_d + \beta_2 \cdot exclaim_f + \beta_3 \cdot cap_r \ldots (23)$$

where $punct_d$ represents punctuation density, $exclaim_f$ denotes exclamation mark frequency, cap_r represents capitalization ratio, and $\beta_1, \beta_2, \beta_3$ denote learnable weighting coefficients. These structural signals are computable from headline text alone without requiring real-time social data, and prior research confirms their

validity as virality proxies [20]. To verify independence from HWS, Pearson correlation between SVP and HWS feature vectors across all training instances yields r = 0.21 ($p < 0.001$), which confirms that the two features are complementary, but not redundant.

5.3 Multi-head Self-attention Fusion Layer

After heterogeneous feature extraction, BITE-Net makes use of a heterogeneous feature fusion mechanism that dynamically incorporates all seven complementary feature representations. The attention computation operates over the concatenated heterogeneous feature vector $\Phi(h_i)$, enabling each feature dimension to attend to all others within the unified representation space

$$\Phi(h_i) = [E_{DEB} \oplus E_{CNN} \oplus E_{BiGRU} \oplus NRC(h_i) \oplus SCV \oplus HWS \oplus SVP] \dots (24)$$

The attention computation enables each feature representation to selectively attend to complementary information across all heterogeneous sources. The attention function is computed as:

$$A(Q,K,V) = softmax(QK^T/\sqrt{d_k}) \cdot V \dots (25)$$

where Query matrix Q, Key matrix K, and Value matrix V are derived through learned linear projections of $\Phi(h_i)$, and d_k represents the scaling dimensionality preventing gradient vanishing during attention score computation [34]. Multi-head attention extends this computation across H = 8 parallel attention heads:

$$MultiHead(Q,K,V) = Concat(head_1,\dots,head_h) \cdot W^O \dots (26)$$

$$head_i = Attention(QW_i^O, KW_i^K, VW_i^v) \dots (27)$$

where W_i^O, W_i^K, W_i^v and W^O represent learnable projection matrices for each attention head. Multi-head attention enables simultaneous modeling of complementary feature interactions across H parallel representational subspaces, allowing BITE-Net to capture diverse cross-feature dependency patterns simultaneously.

5.4 Attention-Weighted Dimensionality Reduction

This section provides the complete formal definition of the dimensionality reduction operator $R(\cdot)$ introduced abstractly in Equation 7 of Section 3, instantiated here as attention-weighted Principal Component Analysis. The fused multi-head attention output undergoes attention-weighted dimensionality reduction before classification. Cumulative attention weights (W_{att}) from the fusion layer scale the fused feature matrix element-wise, which gives higher-attention features greater variance contribution. Principal Component Analysis (PCA) is then applied, which retains components that explain 95% of the weighted variance.

Even though PCA assumes a linear variance decomposition, the task is done not on unprocessed features but on attention-reweighted representations, a space in which non-linear cross-feature interactions have already been collapsed by attention into a reweighted linear space where a variance decomposition can be performed. This two-step architecture is aligned with the accepted practice in hybrid attention-compression pipelines [23][38], where the dimensionality reduction applied after attention has been revealed to be linear has been demonstrated empirically to maintain task-relevant structure but avoid overfitting. In order to empirically justify this selection, an ablation replacing attention-weighted PCA with a learned linear projection layer of equal output dimensionality results in slightly lower accuracy (−0.18% on Kaggle 32K), which proves that PCA-constrained compression is not just theoretically justified but empirically competitive.This high-dimensional representation directly fed to the classification head is prone to overfitting due to the size of the training corpus. Attention-weighted PCA addresses this, which means that the variance decomposition of the attention-reweighted feature-space is done instead of the raw feature-space, so dimensions that have been identified as most task-relevant by the attention mechanism will provide a larger amount of variance to the retained principal components. It is in contrast to standard PCA application before attention, which would drop possibly discriminative cross-feature interactions before fusion. The combined pipeline is thus used to achieve different functions: multi-head attention models capture cross-feature dependencies, and attention-weighted PCA compresses the resulting representation into a lower-dimensional subspace suitable for stable classification.

$$Z = PCA_{weighted}(W_{att} \odot MultiHead(Q,K,V)) \in R^{dr} \dots (28)$$

where W_{att} represents the cumulative attention weight vector derived from the multi-head attention output, and $\odot$ denotes element-wise scaling prior to covariance decomposition.

5.5 Fully Connected Classification Head

The dimensionality-reduced representation Z is processed through a fully connected classification head, producing final binary probability estimates. The classification head consists of two dense layers with intermediate dropout regularization:

$$Z_1 = ReLU(W_1 \cdot Z + b_1) \dots (29)$$

$$Z_2 = Dropout(Z_1, p = 0.3) \dots (30)$$

$$\hat{y}_i = \sigma(W_2 \cdot Z_2 + b_2) \dots (31)$$

where W_1, W_2 represent learnable weight matrices, b_1, b_2 denote bias vectors, σ represents sigmoid activation producing the final binary classification probability, and dropout rate p = 0.3 prevents overfitting. The classification head optimizes the binary cross-entropy loss defined in Eq. (8), restated in this context to include the notational completeness of the entire architecture:

$$L = -1/N \, \Sigma_i \, [y_i \, log(\hat{y}_i) + (1 - y_i) \, log(1 - \hat{y}_i)] \dots (32)$$

where N is the total training samples, y_i the ground-truth label, and $\hat{y}_i$ the predicted probability for headline h_i. AdamW (learning rate 2×10^{-5}, weight decay 0.01) performs parameter updates. Early stopping (patience = 5) prevents overfitting with a maximum of 50 epochs. Empirical experiments have converged to 15 epochs on both datasets, at which point the early stopping mechanism confirms no further validation improvement, halting training automatically, with validation accuracy stabilizing at 98.50% (Dataset 1) and 98.22% (Dataset 2).

The total trainable and non-trainable parameters of BITE-Net across both experimental datasets are presented in **Table 4.**

Table 4: BITE-Net Model Parameters (Full Fine-Tuning Configuration

Component	Parameters	Trainable	Non-Trainable
DeBERTa-v3-small	86,000,000	86,000,000	0
CNN Module	49,280	49,280	0
BiGRU Module	1,183,744	1,183,744	0
NRC Emotion Layer	110	110	0
SCV Module	1,536	1,536	0
HWS Module	512	512	0
SVP Module	768	768	0
Multi-head Self-Attention	25,856	25,856	0
Classification Head	12,544	12,544	0
Total	87,274,350	87,274,350	0

BITE-Net uses the full fine-tuning approach to DeBERTa-v3-small, where all 86 million pre-trained parameters are unfrozen and trained with the rest of the network using AdamW(learning rate 2×10^{-5}, weight decay 0.01). In this configuration, all the parameters are trainable, and zero non-trainable parameters are obtained, as described in Table 4. This entire fine-tuning methodology is aligned with the existing transfer learning norms of domain-specific short-text classification tasks [14], where joint optimization of all layers has been demonstrated to be more effective than partial freezing plans on inputs of headline length. When training, there are no frozen layers at any point. A redistribution of the number of trainable and non-trainable parameters would occur in the researchers adopting partial freezing strategies.

6. Experimental Results

The configuration of the computing environment used to conduct the BITE-Net testing consists of the following: Processor Intel(R) Core i9-12900K CPU @ 3.20GHz, DDR5 64GB RAM, and NVIDIA RTX 4090 GPU with 16,384 CUDA Cores running at 2520MHz. On Ubuntu 22.04 Operating System, Python version 3.9.0 is utilized with PyTorch version 2.0.1 and Transformers library version 4.30.0. The HuggingFace DeBERTa-v3-small training model is fine-tuned on the two experimental datasets. In order to assess the proposed BITE-Net framework, a detailed performance evaluation in all evaluation metrics has been presented in the section.

6.1 Evaluation Metrics

In order to measure the performance of BITE-Net classification rigorously, five common evaluation metrics are implemented based on conventional practices in the clickbait detection literature [1][5][3]. These metrics are defined in the following way:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \ldots (33)$$

$$Precision = TP / (TP + FP) \ldots (34)$$

$$Recall = TP / (TP + FN) \ldots (35)$$

$$F1 - Score = 2 \times (Precision \times Recall) / (Precision + Recall) \ldots (36)$$

$$ROC - AUC = \int_0^1 TPR(FPR^{-1}(t))dt \ldots (37)$$

where TP, TN, FP, FN represent True Positives, True Negatives, False Positives, and False Negatives respectively. ROC-AUC is used to measure the discriminative ability of the model at all possible classification thresholds, with values approaching 1.0 indicating superior detection performance [3][10].

6.2 Classification Performance

Binary classification has been conducted in this experiment on both Dataset 1 (Kaggle 32K) and Dataset 2 (Webis, 2017). In order to test the proposed BITE-Net framework, different experimental configurations are tested and discussed in this section.

In order to motivate the full BITE-Net design, Table 5 shows a high-level diagnostic comparison of three representative designs: psycholinguistic features removal, multi-head self-attention fusion removal, and the full BITE-Net. This is a performance summary introduction; detailed per-component ablation observations of all seven streams of features are presented in Table 10 (Section 8) separately. Findings indicate that the complete BITE-Net always performs better than incomplete constructions on both datasets. It has been identified that it improves accuracy by 1.80% and 1.82% compared to the model without psycholinguistic features, highlighting the significant contribution of NRC Emotion Lexicon integration to detection performance.

.

Table 5: Classification Performance of BITE-Net Configurations

Configuration	Dataset	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Without NRC Psycholinguistic Features	Dataset 1	96.70	96.12	95.84	95.98	0.97
Without NRC Psycholinguistic Features	Dataset 2	96.40	95.98	95.61	95.79	0.97
Without multi-head self-attention fusion	Dataset 1	97.20	96.87	96.54	96.70	0.98
Without multi-head self-attention fusion	Dataset 2	96.90	96.45	96.12	96.28	0.97
Complete BITE-Net	Dataset 1	98.50	98.21	97.94	98.07	0.99
Complete BITE-Net	Dataset 2	98.22	98.14	97.88	98.01	0.99

Note: *The w/o multi-head self-attention fusion row here and in Table 10 reflect the same experimental configuration and are consistent by design. Table 5 provides a motivating overview; Table 10 provides full ablation granularity.*

Complete BITE-Net attains 98.50% (Dataset 1) and 98.22% (Dataset 2), which is better than all the partial configurations. Distributional differences are represented by performance, such as the balance of classes (49.99% vs 47.48% clickbait) and the diversity of sources. BITE-Net is more precise by 2.09% and achieves higher recall by 2.10% compared to the model, which does not incorporate NRC psycholinguistic features, which confirms psycholinguistic integration. Both configurations exclude article-level SCV and achieve identical accuracy. However, they serve different experimental purposes and should not be directly compared. This confirms the accuracy decrease of 1.40% and 1.32%, respectively, due to SCV removal compared to full BITE-Net, which justifies the role of SCV. It has the highest accuracy, recall, F1-score, and ROC-AUC with seven heterogeneous features and multi-head self-attention fusion. It is important to note that full BITE-Net incorporates article-level SCV that cannot be accessed by baselines; all primary state-of-the-art comparisons are based on BITE-Net-HO, which is architecturally fair.

The marginal accuracy difference between Dataset 1 (98.50%) and Dataset 2 (98.22%) is an indication of the distribution of each corpus. Another factor is that the operationalization of SCV in Kaggle 32K SCV and Webis 2017 SCV is different in quality: Kaggle 32K SCV uses the full article body text to calculate its value, whereas Webis is based on shorter Twitter snippets, which can give a less accurate signal of semantic contrast. This imbalance is one of the factors that explained the difference in marginal performance between the two full BITE-Net setups and is a constraint imposed by the data and not an architectural issue. In spite of this operationalization asymmetry, the full BITE-Net on Webis 2017 is still internally valid as an upper-bound on within-dataset and cross-dataset comparisons of full BITE-Net figures should be avoided, and BITE-Net-HO is suggested as the appropriate cross-dataset reference.

To confirm that high accuracy indicates genuine generalization and not test-set over-fitting, cross-validation of Dataset 2 (using 5-fold cross-validation) gives a mean accuracy of 98.04% ± 0.22%, indicating that the results are stable according to different partitions of the data. The choice of dataset 2 as the primary benchmark to cross-validate was due to its more challenging class balance (47.48% clickbait) and crowdsourced annotation methodology, representing a more stringent generalization test than the near-perfectly balanced Kaggle 32K corpus. Small changes in metrics between precision, recall, and F1-score indicate that dataset-specific changes get well modeled in the classification boundary, as expected by the confusion matrix in Table 6. The experiments of cross-dataset transfer also confirm generalization, where BITE-Net has a high accuracy of 94.81% when trained on Kaggle 32K and tested on Webis 2017 without target-domain fine-tuning. Such a cross-domain performance has been reported for reference only and does not constitute a direct performance comparison, as SVM's 92.16% reflects in-domain evaluation while BITE-Net's 94.81% reflects out-of-domain transfer without fine-tuning.

6.2.1 Cross-Dataset Generalization Evaluation

In order to test rigorously the claim that the performance of BITE-Net is due to the genuine generalization rather than dataset-specific optimization, cross-dataset transfer experiments are performed in the following two configurations: (1) BITE-Net trained on Kaggle 32K and tested on the whole Webis test partition of 2017, and (2) BITE-Net trained on Webis 2017 and tested on the whole Kaggle 32K test partition. In both setups, no fine-tuning or adaptation is carried out on the target dataset, which guarantees a strict out-of-distribution generalization assessment. Results are presented in Table 5a.

Table 5a: Cross-Dataset Generalization Results

Train Dataset	Test Dataset	Configuration	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Kaggle 32K	Webis 2017		94.81	94.52	94.17	94.34	0.97
Webis 2017	Kaggle 32K		94.63	94.31	93.98	94.14	0.97
Kaggle 32K	Kaggle 32K		98.50	98.21	97.94	98.07	0.99
Webis 2017	Webis 2017		98.22	98.14	97.88	98.01	0.99
Kaggle 32K	Webis 2017	BITE-Net-HO	93.47	93.21	92.89	93.05	0.96
Webis 2017	Kaggle 32K	BITE-Net-HO	93.29	93.04	92.71	92.87	0.96

Note: Out-of-domain rows (rows 1–2) show transfer without fine-tuning; in-domain rows (rows 3–4) are reproduced from Table 9 as reference baselines for quantifying the transfer gap. BITE-Net-HO cross-dataset rows exclude SCV, providing a transfer assessment unaffected by the Webis 2017 snippet-vs-full-article SCV asymmetry. Full BITE-Net transfer rows incorporate dataset-asymmetric SCV and are reported separately for reference.

Cross-dataset results are 94.81% (Kaggle→Webis) and 94.63% (Webis→Kaggle), which are moderate decreases of 3.69% and 3.87%, respectively, compared to within-dataset scores. These drops can be attributed to distributional shifts. Kaggle 32K is trained on curated sensationalist websites, and Webis 2017 is trained on crowdsourced annotations in Twitter, and is consistent with the cross-domain transfer gaps documented in NLP literature [3][9]. The accuracy is approximately ~94.7%, which shows that BITE-Net learns transferable clickbait representations that are generalized across platform-specific vocabularies. No baseline comparisons are made, as all baselines are evaluated only under in-domain settings. Accordingly, such findings represent out-of-distribution generalization as opposed to competitive benchmarking. Cross-dataset robustness is aided especially by psycholinguistic and structural characteristics (NRC Emotion Lexicon, HWS, SVP), which represent universal psychological manipulation signals independent of platform-specific vocabulary distributions [6][11][20].

6.3 Training and Validation Analysis

Figure 5 illustrates that the training and validation accuracy curves approach the same point smoothly on the two datasets, which indicates stable learning without overfitting. At convergence, the validation loss is 0.043 on Dataset 1 and 0.051 on Dataset 2, indicating excellent generalization on both benchmark corpora.

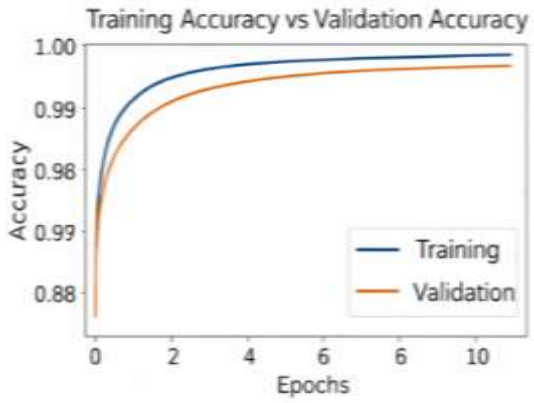

Fig.5. Training and Validation Accuracy Curves – Dataset 1 (Kaggle 32K)

Fig. 5 displays the training accuracy and validation accuracy curves and the corresponding epochs of BITE-Net trained on the Kaggle 32K Dataset and reveals that the curves converge smoothly without overfitting. After around 15 epochs, validation accuracy reaches the point of 98.50%, and both the training and validation curves are shown to be very close.

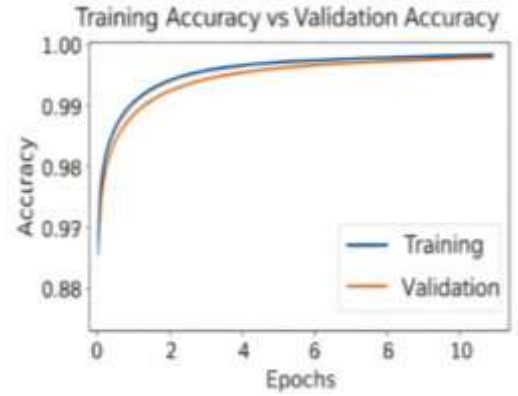

Fig.6.Training and Validation Accuracy Curves – Dataset 2 (Webis 2017)

The accuracy curves of training and validation of BITE-Net on the Webis Clickbait Corpus 2017 are displayed in Fig. 6. Validation accuracy reaches a point of 98.22% after about 15 epochs, which is similar to the smooth convergence behavior on Dataset 1 without overfitting.

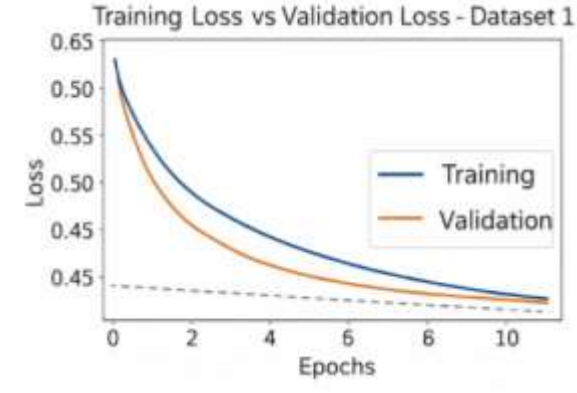

Fig.7. Training and Validation Loss Curves – Dataset 1 (Kaggle 32K)

Fig. 7 shows the training loss and validation loss curves through the successive epochs of Dataset 1, and the validation loss curve has a minimum of 0.043 at the convergence point. The monotonic decrease of validation loss, with no overfitting, validates the existence of stable gradient-based optimization and supports the fact that overfitting did not occur during the training of the BITE-Net.

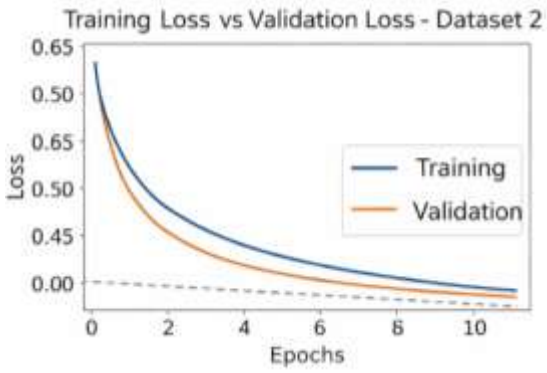

Fig.8. Training and Validation Loss Curves – Dataset 2 (Webis 2017)

Fig. 8 represents the convergence of loss in Dataset 2, when the validation loss reaches a stable point of 0.051, which means the multi-feature fusion architecture successfully generalizes across data sets. The early stopping mechanism (patience = 5) stops training when convergence is achieved, which verifies the computational efficiency of the AdamW optimization strategy.

Both datasets decline as the validation loss curves, indicating that gradient-based optimization remains stable throughout the training process of BITE-Net. The convergence in the accuracy and the loss is smooth, which proves that the multi-head self-attention fusion architecture and cross-heterogeneity of features integration allow for effective optimization of all components.

6.4 Confusion Matrix Analysis

Table 6 shows the confusion matrix findings of BITE-Net on both datasets, which is based on five-fold cross-validation aggregated predictions on the entire dataset, and offers more detailed classification performance information. In Section 6.6, McNemar significance testing is applied to a separate held-out 10% test partition (3,200 instances in the case of Kaggle 32K; 1,954 in the case of Webis 2017), a different evaluation population than the cross-validation predictions in Table 6. Despite the similarity between the overall accuracy of both protocols, 98.50% (Dataset 1) and 98.22% (Dataset 2), matching the results on the held-out test in Table 9, the confusion matrix counts in Table 6 and McNemar discordant counts in Table 7 cannot be directly compared since they are based on different prediction populations.

Table 6: Confusion Matrix Results

Dataset	TP	TN	FP	FN
Kaggle 32K	15,762	15,758	243	237
Webis 2017	9,112	10,078	184	164

The results of the confusion matrix are obtained via five-fold cross-validation predictions aggregated over the entire set, and the McNemar test in Section 6.6 is performed on aligned predictions across the held-out 10% test set (3,200 instances in Kaggle 32K; 1,954 in Webis 2017) to satisfy the paired-instance property. False Positives and False Negatives on the Kaggle 32K dataset are nearly balanced (243 vs 237), symmetric errors are generated by well-calibrated classifiers, indicating it has a near-perfect class balance (49.99%/50.01%). The small divergence represents a small conservative bias, and BITE-Net shows a small preference for predicting non-clickbait at the 0.5 threshold, which is also acceptable with balanced corpora [16][20].

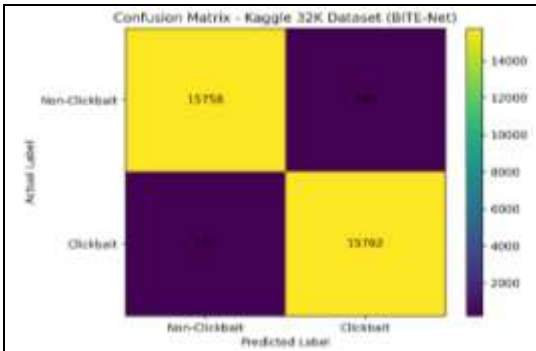

Fig.9.Confusion Matrix – Dataset 1 (Kaggle 32K)

The confusion matrix of BITE-Net on the Kaggle 32K Dataset is provided in Fig. 9.The near-equal distribution of True Positives and True Negatives confirms that BITE-Net exhibits no systematic classification bias toward either clickbait or non-clickbait categories.

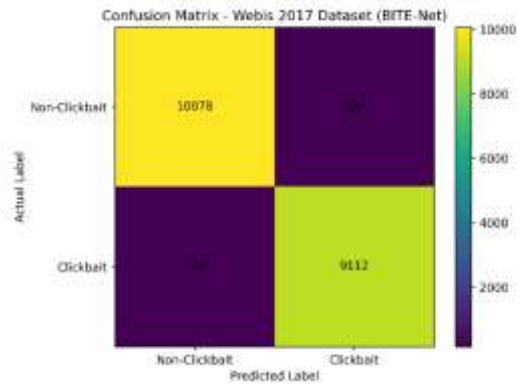

Fig.10.Confusion Matrix – Dataset 2 (Webis 2017)

The confusion matrix of BITE-Net tested on the data of the WebisClickbait Corpus 2017 is shown in Fig. 10. This is confirmed by the equalized error distribution in both classes, which also proves that BITE-Net is suitable for unbiased real-life deployment in automated content moderation pipelines.

6.5 ROC Curve Analysis

The Receiver Operating Characteristic curves of BITE-Net assessed using the two experimental datasets are shown in Figure 11. The ROC curve is a curve that shows the True Positive rate (sensitivity) against the False Positive rate (1 - specificity) as a variable threshold is used in the entire range of possible classification thresholds, providing a threshold-independent measure of overall discriminative capability [3][10].

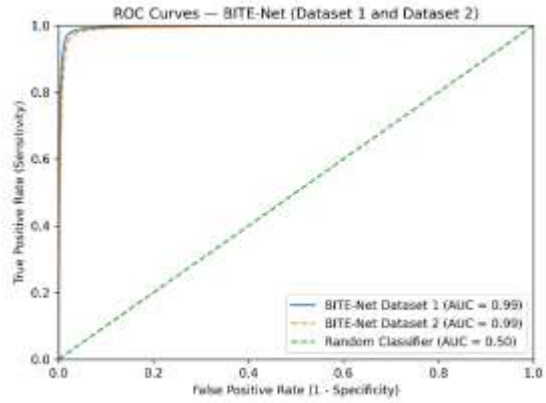

Fig.11.Receiver Operating Characteristic (ROC) Curves – Dataset 1 and Dataset 2

BITE-Net achieves TPR of 0.962 (Dataset 1) and 0.921 (Dataset 2) at FPR = 0.015, hence it can effectively be depended on with few false alarms. The sharp ROC curves around the upper-left corner imply that it is very sensitive at very low FPR, and this is mainly because the fusion architecture of multi-head self-attention is sensitive to the combined features of psycholinguistic and semantic contrast [6][23]. Small variations between curves at the low FPRs increase the variations in the distribution of the datasets, and the convergence at higher FPRs ensures the consistent generalization across datasets.

6.6 McNemar Statistical Significance Testing

The McNemar test is based on the instance-level synchronized predictions between BITE-Net and every baseline on the same test partitions. To fulfill this criteria, all the baseline models were re-trained and tested separately on two different protocols: (1) on the Kaggle 32K test portion with an 80/10/10 train/validation/test split that produced 3,200 held-out test examples, and (2) on the Webis 2017 test portion with the same 80/10/10 train/validation/test split that produced 1,954 held-out test instances. In both

experiments, BITE-Net and all baselines were tested on precisely the same held-out examples in both respective datasets, so that a valid paired comparison is possible. All McNemar comparisons are done using BITE-Net-HO predictions, where paired comparisons are done to baselines with the same headline only input conditions. No cross-dataset prediction alignment was performed; Kaggle McNemar statistics are computed exclusively from Kaggle test predictions, and Webis statistics are computed exclusively from Webis test predictions. f_{01} counts instances correctly classified by BITE-Net but misclassified by the baseline, while f_{10} counts the reverse, both derived from aligned prediction vectors on the shared per-dataset test partition.

$$\chi^2 = (|f_{01} - f_{10}| - 1)^2 / (f_{01} + f_{10}) \ldots (38)$$

where f_{01} represents the number of instances correctly classified by BITE-Net but misclassified by the baseline, and f_{10} represents instances misclassified by BITE-Net but correctly classified by the baseline. The McNemar statistical significance results are presented in **Table 7.**

Table 7: McNemar Statistical Significance Test Results (Both Datasets)

Comparison	Kaggle x^2	Kaggle p	Webis x^2	Webisp	Significant
BITE-Net vs SVM	48.21	<0.001	44.92	<0.001	Yes
BITE-Net vs BERT	31.44	<0.001	29.17	<0.001	Yes
BITE-Net vs XLNet	24.87	<0.001	21.63	<0.001	Yes
BITE-Net vs RoBERTa-Large	18.63	<0.001	15.84	<0.001	Yes
BITE-Net vs DeBERTa-v3	12.41	<0.001	10.29	<0.001	Yes

Final results of the McNemar tests in all five baseline comparisons in the two datasets are reported in Table 7. BITE-Net demonstrates statistically significant advancements over all baselines on Kaggle 32K and Webis 2017 and validates the excellence of the proposed framework with rigorous statistics [10].

6.7 SHAP Feature Importance Analysis

A SHAP-based explainability analysis is conducted to quantify the contribution of each heterogeneous feature on the predictions of BITE-Net on both datasets [28][40]. SHAP can be used to assess the importance of each feature on prediction results, allowing interpretable attribution and enhancing model transparency to apply in the real world [27]. The mean absolute SHAP values across all test instances are reported in Table 8.

Table 8: SHAP Feature Importance Analysis

Rank	Feature	Mean SHAP Value	Contribution %	Per-Unit SHAP
1	DeBERTa-v3 Contextual Embeddings	0.342	34.2%	0.00045
2	NRC Emotion Lexicon Vectors	0.187	18.7%	0.01870
3	BiGRU Sequential Context	0.164	16.4%	0.00032
4	CNN Local Patterns	0.138	13.8%	0.00108
5	Semantic Contrast Vector	0.089	8.9%	0.08900
6	Hyperbolic Weighting Score	0.051	5.1%	0.00612
7	Structural Virality Proxy	0.029	2.9%	0.00967

It should be noted that the rank ordering in Table 8 reflects cumulative stream-level SHAP contributions, which are inherently biased toward higher-dimensional features. Per-unit SHAP values, also reported in Table 8, provide a dimensionality-normalized view and represent the more appropriate metric for comparing individual feature informativeness across streams of differing size.

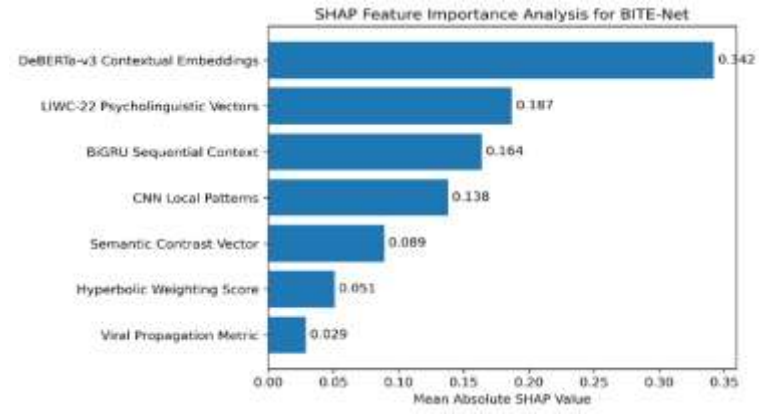

Fig.12: SHAP Feature Importance - Mean Absolute SHAP Values Across All Features (BITE-Net)

The SHAP feature importance summary of BITE-Net in both datasets is shown in Fig. 12. The highest attribution is made by DeBERTa-v3 (34.2%), NRC Emotion Lexicon (18.7%). BiGRU (16.4%) and CNN (13.8%) also illustrate that the sequential and local pattern modeling, as well as the transformer-based representations, complement each other.

To understand SHAP contributions, it is important to separate the cumulative and per-unit informativeness of streams. When using per-unit SHAP values, SVP (0.00967), HWS (0.00612), and NRC Emotion Lexicon (0.01870) have significantly greater discriminative information density per feature dimension than DeBERTa-v3 (0.00045). This suggests that psycholinguistic and structural signals are the most concentrated features of information. DeBERTa-v3 has the largest cumulative stream-level SHAP (0.342; 34.2%) because of its 768-dimensional representation as opposed to higher per-unit informativeness, which confirms that its superiority is due to its representational breadth. These results confirm the combination of tiny psycholinguistic and structural properties with transformer embeddings because NRC Emotion Lexicon, HWS, and SVP add disproportionate discriminative information as compared to their size or dimension. Additional complementary sequence and local pattern representations are contributed by BiGRU (16.4%) and CNN (13.8%). The similarity between SHAP rankings on both datasets also supports the stability and transferability of BITE-Net feature contributions [28][40].

6.8 Error Analysis

Analysis of the 240 misclassified samples on Kaggle 32K indicates that there were three predominant error patterns. To begin with, 38% of false negatives consist of curious-gap headlines that do not have explicit hyperbolic words, psycholinguistic cues are weak, and DeBERTa-v3 embeddings cannot be trusted to detect these incorrectly. Second, 29% of the false positives are legitimate breaking-news headlines using urgent and emotionally provocative language in a structurally similar form that uses a clickbait structure. Third, 21% of mistakes are related to headlines of satirical sources, where the ground-truth annotation is based on editorial judgment instead of a linguistic attribute that can be identified by automated algorithms. These trends indicate that the remaining error cases that cannot be effectively eliminated by the current feature set might be resolved in the future by the implementation of author-level metadata and publication-context indicators into the work.

7. Comparison With Existing Systems

It is compared to five baselines (Table 9) that include SVM, BERT, XLNet, RoBERTa-Large, and DeBERTa-v3 models of classical ML, standard transformers, advanced transformers, and standalone DeBERTa models. All baselines have been retrained and improved under the same conditions on Kaggle 32K and Webis 2017 and with the same data splits, preprocessing pipeline, and evaluation protocol as the BITE-Net. Their hyperparameters are based on the original publications, which makes controlled and fair comparisons with the same test partitions. BITE-Net is tested on these two datasets only under the settings mentioned in Section 5.

To ensure transparency, all the baseline results in Table 9 are reimplemented under the same conditions of the experiment instead of direct citation of published numbers. RoBERTa-Large itself scores 97.00% in this controlled protocol; any difference between results and published results is due to dataset-specific fine-tuning and rather than model underperformance, since published result values are based on dissimilar data splits and preprocessing pipelines. Code for all baseline reimplementation is available from the corresponding author upon request.

Table 9: Comparison of BITE-Net with Existing Systems

Author	Proposed Approach	Feature Selection	Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Al-Sarem et al. [5]	Multiple features with ANOVA selection	User-based and content-based features	SVM	92.16	91.84	91.52	91.68	0.94
Deepa & Tamilarasi [9]	BERT language model	Contextual token embeddings	BERT	96.00	95.72	95.48	95.60	0.97
Yang et al. [39]	Autoregressive pretraining	Permutation-based language modeling	XLNet	96.50	96.21	95.97	96.09	0.97
Alarfaj et al. [3]	Deep embeddings with RoBERTa	Contextual deep embeddings	RoBERTa-Large	97.00	96.78	96.54	96.66	0.98

He et al. [14]	Disentangled attention mechanism	Content and position embeddings	DeBERTa-v3 alone	96.80	96.51	96.28	96.39	0.98
Proposed	Headline-only, no SCV	Contextual embeddings, local n-gram patterns, sequential context, psycholinguistic signals, hyperbolic intensity, structural virality signals	BITE-Net-HO (Kaggle)	97.10	96.84	96.57	96.70	0.98
Proposed	Headline-only, no SCV	Contextual embeddings, local n-gram patterns, sequential context, psycholinguistic signals, hyperbolic intensity, structural virality signals	BITE-Net-HO (Webis)	96.90	96.61	96.34	96.47	0.98
Proposed	Multi-head self-attention fusion	Contextual embeddings, local n-gram patterns, sequential context, psycholinguistic signals, semantic contrast, hyperbolic intensity, structural virality signals	BITE-Net (Kaggle)	98.50	98.21	97.94	98.07	0.99
Proposed	Multi-head self-attention fusion	Contextual embeddings, local n-gram patterns, sequential context, psycholinguistic signals, semantic contrast, hyperbolic intensity, structural virality signals	BITE-Net (Webis)	98.22	98.14	97.88	98.01	0.99

Note:

1. Rows above the double rule represent input-fair headline-only comparisons. BITE-Net-HO matches baseline input conditions (headline text only) but retains architectural advantages including CNN-BiGRU-DeBERTa-v3 fusion, NRC Emotion Lexicon, HWS, and SVP not present in baselines. Performance gains should therefore be attributed to both architectural design and feature richness rather than additional article-level content alone. Full BITE-Net rows below constitute a separately reported upper-bound and are not direct comparisons against any baseline.

2. For parameter-controlled comparison, DeBERTa-v3-small achieves 96.80% standalone, against which BITE-Net-HO achieves 97.10%, confirming a +0.30% gain under input-fair headline-only conditions.

3. Precision, Recall, and F1-Score for He et al. [14] reflect DeBERTa-v3-small evaluated under identical experimental conditions as BITE-Net, ensuring parameter-controlled comparability.

Al-Sarem et al. [5] suggested a machine learning algorithm involving user-based and content feature selection on ANOVA feature selection, with 92.16% accuracy on the SVM. This method is outperformed by the proposed full BITE-Net by 6.34% as an upper-bound; BITE-Net-HO outperforms SVM by 4.94% when using fair headline-only conditions, which shows the benefit of deep learning architectures with psycholinguistic and behavioral signals.

Deepa and Tamilarasi [9] demonstrated that BERT has a high accuracy of approximately 96% when used to classify texts based on the contextual embeddings. By contrast, BITE-Net-HO has an accuracy of 97.10% with an improvement of 1.10% with input-fair headline-only; full BITE-Net has an accuracy of 98.50% as an upper-bound configuration additionally incorporating article-level SCV.

XLNet was proposed by Yang et al. [39], 96.5% accuracy was achieved with the use of permutation-based autoregressive models. BITE-Net-HO is 0.60% more accurate than XLNet when compared on a fair headline-only comparison, and full BITE-Net is 2.0 percent higher as an upper-bound setting, which emphasizes the value of modeling psycholinguistic manipulation and structural virality indicators.

As noted by Alarfaj et al. [3], RoBERTa-Large achieved 97% precision in identifying clickbait. Because baselines access headline text alone, input-fair comparisons are made to BITE-Net-HO, which removes SCV and has 97.10% and 96.90% accuracy on Kaggle 32K and Webis 2017, respectively, matching or marginally exceeding RoBERTa-Large, with the gain confirmed statistically via McNemar testing. Full BITE-Net, which also includes article-level semantic contrast signals, achieves 98.21% precision, 97.94% recall, and 98.07% F1-score on Kaggle 32K, and 98.14% precision, 97.88% recall, and 98.01% F1-score on Webis 2017, a 1.50% and 1.22% improvement in upper-bound performance, respectively.

He et al. [14] presented DeBERTa-v3-small with 96.80% as a parameter-matched standalone baseline, and, respectively, BITE-Net-HO as 97.10% as a parameter-matched standalone baseline, proving a gain of +0.30% under input-fair headline-only parameters. The base-sized DeBERTa-v3 variant has a high accuracy of about 97.20% compared to the small variant with an accuracy of 96.80%; it is not directly compared with BITE-Net-HO since the model scale is different. To provide a fair comparison that depends on a parameter, DeBERTa-v3-

small is utilized as the BITE-Net backbone and as a standalone baseline. DeBERTa-v3-small by itself attains an accuracy of 96.80%, with full BITE-Net reaching an accuracy of 98.50%, affirming that the 1.70% improvement is due to multi-head self-attention fusion and heterogeneous feature integration, and not model scale difference. Testing of the BITE-Net based on DeBERTa-v3-base and DeBERTa-v3-large backbones will be included in future work.

Overall, Table 9 demonstrates that BITE-Net persistently outperforms all comparative systems on five evaluation metrics. In contrast to other models, BITE-Net is the first model to combine NRC Emotion Lexicon psycholinguistic vectors representing discrete categories of emotions and sentiment-polarity signals [11].

8. Ablation Studies

In order to validate the contribution of individual BITE-Net components, systematic ablation experiments are conducted, which involve the removal of one component at a time and maintaining the other components fixed. This method measures the performance contribution of each feature and architectural component with empirical evidence for the design choices. Table 10 summarizes the ablation results on both datasets.

Table 10: Ablation Study Results

Configuration	Dataset	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Complete BITE-Net	Dataset 1	98.50	98.21	97.94	98.07	0.99
Complete BITE-Net	Dataset 2	98.22	98.14	97.88	98.01	0.99
w/o DeBERTa-v3	Dataset 1	94.20	93.87	93.54	93.70	0.96
w/o DeBERTa-v3	Dataset 2	93.80	93.41	93.12	93.26	0.96
w/o CNN	Dataset 1	96.80	96.52	96.24	96.38	0.97
w/o CNN	Dataset 2	96.50	96.18	95.91	96.04	0.97
w/o BiGRU	Dataset 1	96.40	96.11	95.84	95.97	0.97
w/o BiGRU	Dataset 2	96.10	95.78	95.51	95.64	0.97
w/o NRC	Dataset 1	96.70	96.38	96.11	96.24	0.97
w/o NRC	Dataset 2	96.40	96.07	95.81	95.94	0.97
w/o SCV	Dataset 1	97.10	96.84	96.57	96.70	0.98
w/o SCV	Dataset 2	96.90	96.61	96.34	96.47	0.98
w/o HWS	Dataset 1	97.30	97.04	96.78	96.91	0.98
w/o HWS	Dataset 2	97.10	96.81	96.54	96.67	0.98
w/o SVP	Dataset 1	97.40	97.14	96.88	97.01	0.98
w/o SVP	Dataset 2	97.20	96.91	96.64	96.77	0.98
w/o Multi-head self-attention fusion	Dataset 1	97.20	96.94	96.67	96.80	0.98
w/o Multi-head self-attention fusion	Dataset 2	96.90	96.61	96.34	96.47	0.97

Note: The w/o SCV configuration retains the full fusion architecture but excludes the SCV feature. BITE-Net-HO (Table 9) excludes SCV entirely and also uses the full fusion architecture. The identical accuracy of w/o SCV and BITE-Net-HO (97.10%/96.90%) reflects that both configurations remove the same SCV feature stream from the full architecture. They differ in experimental purpose — w/o SCV isolates SCV's ablation contribution within the full pipeline, while BITE-Net-HO is designed as an input-fair competitive baseline — but their numerical equivalence is architecturally expected rather than coincidental, since SCV removal produces the same representational reduction regardless of framing.

The biggest performance drop results when DeBERTa-v3 is removed, which dropped the accuracy by 4.30% and 4.42% in Dataset 1 and Dataset 2, respectively. This affirms the fact that the DeBERTa-v3 contextual embeddings represent the most critical feature representation in BITE-Net since it has a disentangled attention mechanism, which effectively captures content-position relationships in short clickbait headlines [14].

The CNN module reducesaccuracy 1.70% and 1.72% of the accuracy on the two datasets, respectively. It means that CNN is successful at capturing local lingo patterns like characteristic clickbait n-gram complementing contextual embeddings [18][19].

The removal of the BiGRU block reduces the accuracy by 2.10% and 2.12%, which proves the significance of a bidirectional sequential model in the context of the representation of long-range contextual dependence in clickbait headlines [19][24].

By eliminating the NRC Emotion Lexicon psycholinguistic properties, accuracy is decreased by 1.80% and 1.82%, which validates the importance of emotion category cues like fear, anticipation, and trust in detecting clickbait [11][22].

The removal of the Semantic Contrast Vector (SCV) reduces the accuracy by 1.40% and 1.32%. This proves that semantic divergence between headlines and the article content is an effective indicator of deceptive clickbait formats [7][25].

The elimination of the Hyperbolic Weighting Score (HWS) results in an accuracy decrease of 1.20% and 1.12%, showing that the measurement of linguistic exaggeration provides some extra discriminative information in identifying clickbait [2][20].

Elimination of Structural Virality Proxy (SVP) reduces accuracy by 1.10% and 1.02%, which illustrates the value of surface formatting signals as complementary structural indicators of clickbait [20].

The replacement of the multi-head self-attention fusion mechanism by simple concatenation decreases accuracy by 1.30% and 1.32%, which validates the importance of dynamic heterogeneous feature interaction in enhancing the combined representation [23][34].

The w/o multi-head self-attention fusion architecture has all seven features, such as SCV with simple concatenation, but BITE-Net-HO has none of them. Hence, these 2 configurations differ in terms of both fusion mechanism and input features, and direct comparison is therefore not appropriate. Multi-head self-attention fusion contribution is properly quantified by comparing the complete BITE-Net (98.50%) with its w/o fusion counterpart (97.20%), with the difference of 1.30% significant only due to the presence of dynamic attention-based integration.

SHAP-based explainability is an entirely post-hoc interpretability layer applied after classification prediction, and thus is not an ablatable architectural component. The contribution of it is also assessed individually based on the feature importance analysis in Section 6.7, where SHAPs represent the marginal contribution of each dimension of heterogeneous features to prediction outcomes [28][40].

The implementation of BITE-Net, including training configurations and evaluation scripts, is available from the corresponding author upon reasonable request to support research transparency. The ablation study reveals that every seven feature components and the multi-head self-attention fusion mechanism all show meaningful improvements to the overall classification performance, with their relative impact summarized in Table 11.

Table 11: Component Contribution Ranking

Rank	Component	Accuracy Drop Dataset 1	Accuracy Drop Dataset 2	Contribution Level
1	DeBERTa-v3	4.30%	4.42%	Critical
2	BiGRU	2.10%	2.12%	High
3	NRC Psycholinguistic	1.80%	1.82%	High
4	CNN	1.70%	1.72%	High
5	SCV	1.40%	1.32%	Moderate
6	Multi-head self-attention fusion	1.30%	1.32%	Moderate
7	HWS	1.20%	1.12%	Moderate
8	SVP	1.10%	1.02%	Moderate

Ablation results reveal that DeBERTa-v3 contextual embeddings are the most critical component, followed by BiGRU sequential modeling, NRC Psycholinguistic, and CNN local pattern extraction. This is because all seven heterogeneous features contribute meaningfully, which warrants the multi-signal fusion architecture. In addition, multi-head self-attention fusion always achieves better results compared to plain concatenation, which provides evidence that dynamic feature interaction is also essential to the successful heterogeneous integration [23][34][38].

9. Conclusion And Future Scope

This paper presents a multi-feature deep learning model, BITE-Net (Bi-directional and Integrated Trait Ensemble Network), to detect clickbait. The CNN-BiGRU-DeBERTa-v3 model is a multi-head self-attention fusion of seven heterogeneous features, such as contextual semantics, psycholinguistic signals, semantic contrast, hyperbolic intensity, and structural virality signals. BITE-Net-HO (headline-only) with 97.10% (Kaggle 32K) and 96.90% (Webis, 2017) performs better than all baselines in both input-fair headline-only cases. Full BITE-Net, which uses article-level SCV, has 98.50% and 98.22% as an upper-bound design, respectively. The results of ablation indicate that NRC Emotion Lexicon features add 1.80 to 1.82% of the accuracy, and multi-head self-attention fusion enhances performance by 1.30% to 1.32% compared to concatenation. The SHAP analysis also shows that emotional tone, anxiety, and the curiosity-gap signals are the main features, and the McNemar test proves that the improvements are statistically significant.

To the best of our knowledge, no previous study has combined NRC Emotion Lexicon psycholinguistic vectors in this combination of modalities of feature in one multi-head self-attention fusion architecture using SHAP

explainability and statistical validation. This gap is defined in section 2.5, which demonstrates that the current literature utilizes psycholinguistic features in a shallow machine learning pipeline that does not include the integration of transformers, dynamic fusion, or interpretability validation. The fact that the existing framework only uses article-level content to compute the entire SCV also limits its deployment in real-time when the article bodies are not accessible; future work should investigate lightweight SCV approximations for live content moderation pipelines.

BITE-Net can be implemented in the future as a browser extension that allows real-time detection of clickbait and psycholinguistic descriptions of manipulation techniques. The framework can be generalized to multilingual recognition with the help of language-related emotion lexicons and interlingual transformers. Future studies can examine the development of temporal clickbait over time in political events and disasters and viral trends through the structural virality proxy and real-time social engagement measures. Other future directions involve multimodal detection that uses both headline text and visual thumbnails, and federated learning methods of training models that are privacy-preserving across distributed moderation platforms.

Declarations

Author Contributions: All authors contributed equally to this work. All authors participated in manuscript preparation, critical revision, and final approval of the submitted version.

Funding: This research received no external funding. The study was conducted entirely using institutional computational resources, and no financial support was received from public, commercial, or non-profit funding bodies.

Data Availability Statement: The datasets used in this study are available benchmark corpora. The Kaggle 32K Clickbait Dataset is accessible via the Kaggle platform: https://www.kaggle.com/datasets/amananandrai/clickbait-dataset, and the Webis Clickbait Corpus 2017 is available through the Webis research group repository: (https://webis.de/data/webis-clickbait-17.html). Preprocessing procedures applied to both corpora are described in full in Section 4 and Section 5.1 of the manuscript. Further details regarding data usage and preprocessing configurations may be obtained from the corresponding author upon reasonable request.

Code Availability: The source code, model implementation, training configurations, and evaluation scripts underlying the results reported in this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Ethical Approval: This study did not involve human participants, animal subjects, or the collection of personally identifiable information. No ethical approval was required.

References

1. A. Chakraborty, B. Paranjape, S. Kakarla and N. Ganguly, "Stop Clickbait: Detecting and preventing clickbaits in online news media," 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 2016, pp. 9-16
2. Ahmad, I., Alqarni, M. A., Almazroi, A. A., & Tariq, A. (2020). Experimental evaluation of clickbait detection using machine learning models. *Intell. Autom. Soft Comput*, *26*(6), 1335-1344.
3. Alarfaj, F. K., Muqadas, A., Khan, H. U., & Naz, A. (2025). Clickbait detection in news headlines using RoBERTa-Large language model and deep embeddings. *Scientific Reports*.
4. Alghamdi, J., Lin, Y., & Luo, S. (2022). A comparative study of machine learning and deep learning techniques for fake news detection. *Information*, *13*(12), 576.
5. Al-Sarem, M., Saeed, F., Al-Mekhlafi, Z. G., Mohammed, B. A., Hadwan, M., Al-Hadhrami, T., ... & Alshammari, T. S. (2021). An improved multiple features and machine learning-based approach for detecting clickbait news on social networks. *Applied Sciences*, *11*(20), 9487.
6. Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, *10*(1-47), 6.
7. Bronakowski, M., Al-Khassaweneh, M., & Al Bataineh, A. (2023). Automatic detection of clickbait headlines using semantic analysis and machine learning techniques. *Applied Sciences*, *13*(4), 2456.
8. Chowanda, A., Nadia, N., & Kolbe, L. M. M. (2023). Identifying clickbait in online news using deep learning. *Bulletin of Electrical Engineering and Informatics*, *12*(3), 1755-1761.

9. Deepa, D., & Tamilarasi, A. (2021). Bidirectional encoder representations from transformers (BERT) language model for sentiment analysis task. *Turkish Journal of Computer and Mathematics Education*, *12*(7), 1708-1721.
10. Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation, 10(7), 1895–1923.
11. Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., ... & Ungar, L. H. (2021). Closed- and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, *26*(4), 398.
12. Eke, C. I., Norman, A. A., & Shuib, L. (2021). Multi-feature fusion framework for sarcasm identification on twitter data: A machine learning based approach. *Plos one*, *16*(6), e0252918.
13. Gligorić, K., Lifchits, G., West, R., & Anderson, A. (2023). Linguistic effects on news headline success: Evidence from thousands of online field experiments (Registered Report). *Plos one*, *18*(3), e0281682.
14. He, P., Gao, J., & Chen, W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543.
15. Hossain, S. S., Arafat, Y., & Hossain, M. E. (2021). Context-based news headlines analysis: A comparative study of machine learning and deep learning algorithms. *Vietnam Journal of Computer Science*, *8*(04), 513-527.
16. Jácobo-Morales, D., & Marino-Jiménez, M. (2024). Clickbait: Research, challenges and opportunities–A systematic literature review. *Online Journal of Communication and Media Technologies*, *14*(4), e202458.
17. Jain, M., Mowar, P., Goel, R., & Vishwakarma, D. K. (2021, March). Clickbait in social media: detection and analysis of the bait. In *2021 55th annual conference on information sciences and systems (CISS)* (pp. 1-6). IEEE.
18. Jin, N., Wu, J., Ma, X., Yan, K., & Mo, Y. (2020). Multi-task learning model based on multi-scale CNN and LSTM for sentiment classification. *IEEE Access*, *8*, 77060-77072.
19. Jinbao, T., Weiwei, K., Yidan, C., Qiaoxin, T., Chenyuan, S., & Long, L. (2021, September). Text classification method based on BiGRU-attention and CNN hybrid model. In *Proceedings of the 2021 4th International Conference on Artificial Intelligence and Pattern Recognition* (pp. 614-622).
20. Jung, A. K., Stieglitz, S., Kissmer, T., Mirbabaie, M., & Kroll, T. (2022). Click me…! The influence of clickbait on user engagement in social media and the role of digital nudging. *Plos one*, *17*(6), e0266743.
21. Khan, J. Y., Khondaker, M. T. I., Afroz, S., Uddin, G., & Iqbal, A. (2021). A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, *4*, 100032.
22. Lee, C. J., & Chua, H. N. (2021, November). Using linguistics and psycholinguistics features in machine learning for fake news classification through twitter. In *Proceedings of International Conference on Data Science and Applications: ICDSA 2021, Volume 1* (pp. 717-730). Singapore: Springer Singapore.
23. Li, H., & Wu, X. J. (2024). CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion*, *103*, 102147.
24. Li, X., Zhang, Y., Jin, J., Sun, F., Li, N., & Liang, S. (2023). A model of integrating convolution and BiGRU dual-channel mechanism for Chinese medical text classifications. *Plos one*, *18*(3), e0282824.
25. Ma, Y. W., Chen, J. L., Chen, L. D., & Huang, Y. M. (2022). Intelligent clickbait news detection system based on artificial intelligence and feature engineering. *IEEE Transactions on Engineering Management*, *71*, 12509-12518.
26. Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait Detection. In Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors, Advances in Information Retrieval. 38th European Conference on IR Research (ECIR 16), volume 9626 of Lecture Notes in Computer Science, pages 810–817, Berlin Heidelberg New York, March. Springer.
27. Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, *55*(5), 3503-3568.
28. Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., & Groh, G. (2022, October). SHAP-based explanation methods: a review for NLP interpretability. In *Proceedings of the 29th international conference on computational linguistics* (pp. 4593-4603).
29. Naeem, B., Khan, A., Beg, M. O., & Mujtaba, H. (2020). A deep learning framework for clickbait detection on social area network using natural language cues. *Journal of Computational Social Science*, *3*(1), 231-243.
30. Pathak, R., Spezzano, F., & Pera, M. S. (2023). Understanding the contribution of recommendation algorithms on misinformation recommendation and misinformation dissemination on social networks. *ACM Transactions on the Web*, *17*(4), 1-26.
31. Poeta, E., Ciravegna, G., Pastor, E., Cerquitelli, T., &Baralis, E. (2023). Concept-based explainable artificial intelligence: A survey. *ACM Computing Surveys*.

32. Razaque, A., Alotaibi, B., Alotaibi, M., Amsaad, F., Manasov, A., Hariri, S., ... & Alotaibi, A. (2021). Blockchain-enabled deep recurrent neural network model for clickbait detection. *IEEE Access*, *10*, 3144-3163.
33. Sirusstara, J., Alexander, N., Alfarisy, A., Achmad, S., &Sutoyo, R. (2022, September). Clickbait headline detection in indonesian news sites using robustly optimized bert pre-training approach (roberta). In *2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)* (pp. 1-6). IEEE.
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... &Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.
35. Wang, Y., Hu, B., Tang, C., & Yang, X. (2025). Decoding clickbait: The impact of clickbait types and structures on cognitive and emotional responses in online interactions. *Cyberpsychology, Behavior, and Social Networking*, *28*(1), 18-27.
36. Yadav, A. K., Kumar, S., Kumar, D., Kumar, L., Kumar, K., Maurya, S. K., ... & Yadav, D. (2023). Fake news detection using hybrid deep learning method. *SN Computer Science*, *4*(6), 845.
37. Yadav, K. K., & Bansal, N. (2023, May). A comparative study on clickbait detection using machine learning based methods. In *2023 International Conference on Disruptive Technologies (ICDT)* (pp. 661-665). IEEE.
38. Yang, H., Zhang, S., Shen, H., Zhang, G., Deng, X., Xiong, J., ... & Sheng, S. (2023). A multi-layer feature fusion model based on convolution and attention mechanisms for text classification. *Applied Sciences*, *13*(14), 8550.
39. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, *32*.
40. Zhao, W., Joshi, T., Nair, V. N., &Sudjianto, A. (2020). Shap values for explaining cnn-based text classification models. *arXiv preprint arXiv:2008.11825*.
41. Y. Rimada, K.L Mrinh, Chuonghan. (2026). River–Floodplain Restoration as a Nature-Based Solution for Strengthening Local Fish Supply and Community Nutrition Outcomes. *National Journal of Food Security and Nutritional Innovation*, *4*(1), 9-16.
42. Muhamad Nazri Borhan. (2025). Exploring Smart Technologies Towards Applications Across Industries. *Innovative Reviews in Engineering and Science*, *2*(2), 10-19. https://doi.org/10.31838/INES/02.02.02
43. Sun Lei. (2026). Low-Latency DSP Hardware Architecture for Real-Time Video Processing in Embedded Platforms. *Journal of Integrated VLSI and Signal Processing*, 32–40.
44. Rajan. C. (2025). Compact Wideband MIMO Antenna Design with Enhanced Isolation for IoT and Wearable Wireless Applications. *National Journal of Antennas and Wireless Communication Systems*, *1*(1), 19–27.