scientific reports

OPEN

Clickbait detection in news headlines using RoBERTa-Large language model and deep embeddings

Fawaz Khaled Alarfaj[1]✉, Amara Muqadas[2], Hikmat Ullah Khan[3]✉ & Anam Naz[3]

The integration of Large Language Models with Artificial Intelligence is transforming digital news analysis, particularly through progressions in natural language processing. Among the emerging applications, clickbait headline detection has become a significant but challenging research area. The existing research studies using Machine Learning (ML) and Deep Learning (DL) algorithms systems for news headlines analysis are limited to traditional ML and DL models. The proposed study introduces RoBERTa-Large, a transformers-based architecture, for the automated detection of clickbait news headlines. The proposed RoBERTa-Large- effectively captures complex contextual dependencies and semantic relationships within text based on its integration with self-attention mechanism. The model is evaluated against state-of-the-art ML and DL approaches to assess its classification capabilities. A diverse set of textual features including Term Frequency-Inverse Document Frequency (TF-IDF), Part of speech tagging (PoS), n-gram representations, and advanced word embeddings such as word2Vec, and FastText and Sentence Embeddings are employed to encode linguistic information from dataset. A comprehensive empirical analysis indicates that RoBERTa-Large achieves the highest classification accuracy of 97% outperforming relevant existing studies. Moreover, with Explainable AI (XAI) methods, like LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations), for better understanding of the results and explainability.

Keywords Machine learning, Deep learning, Clickbait detection, Natural language processing, Word embeddings, News detection, Large language models

Clickbait has become a widespread issue in today's digital world, where online platforms compete to grab users' attention. It refers to sensational, or misleading headlines designed to attract clicks, often at the cost of delivering genuine or relevant content. This manipulative activity not only misguides readers but also undermines the credibility of online media. Detecting clickbait has therefore emerged as an important research area, aiming to ensure information quality and promote trustworthy digital communication[1] [2]. Based on the analysis the ratio of 1.67 million Facebook posts from 153 media organizations revealed that mainstream media contained 33.54% clickbait headlines, but unreliable media exceeded this rate at 39.26%. The study showed 64.7% of people experienced considerable exposure to clickbait headlines while 97.5% of them confirmed they clicked on these attention-grabbing headlines[3]. Digital news and social media have become commonplace in using clickbait headlines that overestimate or mislead the content to increase clicks. Though such tricks are quite effective in drawing attention, they also diminish the authority of internet sources, lower the level of information, and provoke ethical issues of the producers and readers. Some examples of news headlines either a clickbait or not clickbait in English language show in Table 1.

In digital era, thanks to Artificial Intelligence (AI), many studies have employed various methods, including Natural Language Processing (NLP), and ML techniques, to automatically detect and classify clickbait content[4]. These approaches analyze linguistic patterns, semantic structures, and contextual cues within headlines or articles. However, despite significant progress, existing models still face challenges such as handling ambiguous language, detecting subtle manipulations, and maintaining accuracy across diverse platforms and writing styles[5].

In this study, main aim to design a DL model of sentence embedding used to employ contextual semantic characteristics to the task of finding better clickbait in English news headlines. Besides, we include the XAI

[1]Department of Management Information Systems, School of Business, King Faisal University, Al Ahsa, Saudi Arabia. [2]Department of Computer Science, University of Sargodha, Punjab, Pakistan. [3]Department of Information Technology, University of Sargodha, Punjab, Pakistan. ✉email: falarfaj@kfu.edu.sa; dr.hikmat.niazi@gmail.com

A couple did a stunning photo shoot with their baby after learning she had an inoperable brain tumor	Clickbait News
Has a new chapter on Maryam Nawaz been added to 10th-grade textbook?	
Video of dogs attacking woman in Karachi is from India, not Pakistan	
NADRA has not removed Azad Jammu and Kashmir from ID cards	Non-Clickbait News
On the last business day of July, the dollar became more expensive	
Iran says ready for nuclear talks if West is ‘serious	

Table 1. Example of news headlines clickbait and non-clickbait.

method to explain and justify the model predictions to promote transparency and facilitate decision-making[6]. Linguistic features and semantic representation make it possible for the proposed model to accurately retrieve both shallow textual information and deep associations between contexts[7,8]. Various embeddings and features complement other normal feature engineering methods that have been employed to ensure that the model has a good representation of the text data, also include Transformer base model. In this respect, this study aims to solve the major problems and enhances the performance of the proposed hybrid approach in tackling the issue of clickbait news detection and providing the recognition of real news online[9,10]. The contribution of this research is as follows:

- Introduced a state-of-the-art transformer-based model RoBERTa specifically fine-tuned for English news headlines to accurately classify them as clickbait or non-clickbait, achieving a remarkable accuracy of 97%.
- Conducted a comprehensive analysis between traditional ML, DL, and Transformer-based models across multiple feature extraction techniques.
- Explored diverse feature engineering methods, including TF-IDF, N-Grams, POS tagging, and semantic embedding techniques such as Word2Vec, FastText, and Sentence Embeddings, to enhance model representation and contextual understanding.
- Implemented and compared T5 and DistilBERT models to evaluate their effectiveness and generalization capability in clickbait detection tasks on English news datasets.
- Applied XAI techniques, including LIME and SHAP, to interpret model predictions, providing deeper insights into decision-making processes and ensuring model transparency and trustworthiness.

The paper is organized as follows; “Related work” presents analysis of existing studies with ML and DL. “Proposed methodology” provides proposed research methodology and “Experimental setup” discusses experimental setup sharing the details about dataset and evaluation measures. In “Results and discussion”, present the results and discussion of proposed research methodology for the identification of news clickbait. In “Conclusion” shares the conclusion of this study and future directions.

Related work

The latest NLP and AI developments have impacted automated content classification tools with a particular emphasis on detecting toxic headlines. Researchers examine different detection techniques for deceptive news headlines that combine both basic ML approaches like Support Vector Machine (SVM) and Random Forest (RF) with DL systems including Long Short-term memory (LSTM) and Bidirectional long short-term memory (Bi-LSTM). Exploratory analysis of deceptive headlines depends on the combination of three textual features which include TF-IDF as well as n-grams and part-of-speech tags. Word embedding techniques such as Word2Vec and Glove also help extract contextual meanings from texts. Although research has produced valuable results the current methods face challenges in achieving dataset generalization and processing semantic meaning properly. Therefore, robust methods need improvement. A review of existing clickbait detection literature and methodologies evaluates their key contributions and technical advantages and drawbacks.

Existing study with ML models

The current literature used ML methods to classify clickbait content in the news headlines significantly and relied on many textual attributes including TF-IDF, POS tagging, and n-gram patterns to train classification models. Initial literature on clickbait detection used classical ML techniques. As an example, Cao et al.[11] constructed a model on the 2017 clickbait challenge data with the top ranked features, with results of 82% accuracy with relatively low F1 of the clickbait category (0.61) indicating that they have a challenge with class imbalance. Equally, Sisodia et al.[12] put through a test of 100,000 headlines with several ML learners and found that Random Forests was the best (91.16%). Although these examples underscore the promise of feature-engineered ML models, these models are very reliant on hand-coded feature selection and potentially lack semantic sensitivity.

Recent literature has investigated more extensive sets of ML and DL processes. Compared to 6 classifiers (SVM, Logistic Regression, and LSTM), Ahmad et al.[13] found that the neural models (especially LSTM) and transformers (BERT) yielded higher accuracy on 32,000 headlines (3–4% higher than traditional baselines). The study by Mowar et al.[14] offers an extension of clickbait detection to YouTube content with a stacking ensemble that obtained a high accuracy of 95.38%, which highlights the advantage of learning as an ensemble. Bajaj et al.[15,16] also compared the classifiers and reported that Random Forest was still effective (89%) with a small amount of features, but could be not as semantic as neural models. Al-Sarem et al.[16] have shown that features of social media in Arabic are very strong when using the ML-based techniques in Arabic social media; SVM using features selected by ANOVA has 92.16% accuracy. This demonstrates that performance can be enhanced using language-specific means, but these might not be applicable to other languages or fields. Vincent et al.[17] created a model based on the NLP to identify the clickbait headline and differentiate between the real and those full of clickbait news. High accuracy of the model was at 71% and it utilized Bi-LSTM, Decision Tree, and K-Nearest Neighbor. Genic et al.[18] expanded the Turkish clickbait data to 48,060 samples with 8859 tweets. With the help of this larger dataset, the models of Artificial Neural Network (ANN), Logistic Regression (LR), Random Forest (RF), LSTM, Bi-LSTM, and Ensemble Classifier were evaluated. The findings indicated that basic ML models such as LR and RF had moderate accuracy (85–86%) and DL models such as LSTM, ANN, and Bi-LSTM performed much better compared to them with the latter performing on 97%. This implies that contextual dependencies in headlines are better represented using sequence-based models than using the traditional classifiers. NB and LSTM were used in Adrian et al.[19] to detect social media headlines. The suggested LSTM showed better performance with an accuracy of 96%, and NB showed low performance, which indicates that statistical models are not productive in the processing of intricate linguistic features.

All in all, the direction in these studies indicates that more conventional ML algorithms (LR, NB, RF) tend to perform sufficiently, when the datasets are small, or when the features available are limited, however, deep learning-based models (LSTM, Bi-LSTM, ANN) tend to have a high level of performance because of the ability to take into account sequential effects, semantic peculiarities, and word embeddings. Combined, the literature demonstrates that there are obvious trends, namely: (i) ML models such as SVM and RF can be used as strong baselines, but depend on features that are hand-crafted, (ii) DL models like LSTM can be used to better represent contextual representations of sequential data, and (iii) Transformer-based models (BERT, RoBERTa) are capable of achieving the highest level of accuracy because they base on features that represent global semantic relationships. Nonetheless, most of the earlier literature focuses on performance indicators, but not on the interpretability of the models. The gain in performance is, however, at the expense of a more complex computation and a greater requirement of large, balanced datasets. Summary of existing research studies employing ML models for clickbait detection are displayed in Table 2.

Existing study with DL models

Recently, DL models became a common feature in clickbait detection owing to the ability of architectures like LSTM, Bi-LSTM, GRU, and CNN to capture the context-specific and sequential patterns in news headlines. Naeem et al.[20] introduced an LSTM-based approach that utilized linguistic analysis (POS tagging) and yielded 97% accuracy, which supports the argument that sequence models are effective in syntactical signals. On the same note, Kaur et al.[21] proposed a hybrid CNN-LSTM Bi-directional model which achieved a 95% accuracy and demonstrated that the addition of convolutional layers and recurrent units improves local and sequential features representation. This was verified by Thakur et al.[22] who showed that hybrid models (RCNN + LSTM with embeddings) achieve better accuracy as compared to single standalone models with 94% accuracy. Transformer-based models such as BERT, XLNet, and RoBERTa that Rajapaksha et al.[23] have experimented with Twitter data only obtained 85% results, lower than LSTM-based models. It means that although transformers are potent, their work performance may be dependent on datasets, especially in the case of short or noisy texts like tweets. However, Marreddy et al.[24] demonstrated that transformer variants (ALBERT, RoBERTa-Large, ELECTRA) pre-trained on larger datasets performed powerfully, reaching F1-scores in the range of between 0.94 and indicating that transformer models need enough data volume and task-specific fine-tuning to be able to outperform classical DL models. Other alternative approaches have been investigated, including ensemble learning. Jain et al.[25]. developed an ML based stacking classifier integrating ML learners (KNN, SVM, LR, XGB, NB) and RF meta-classifier with 88.5% on Instagram posts. This was still lower than the results of DL, but it is important to note that even weak learners have a variety that makes them useful in a real-time setting.

There are also interesting trends that can be discovered in language-specific studies. Razaque et al.[26] to solve the problem of news clickbait detection, proposed Clickbait Security browser extension using legitimate and illegitimate list search. Apply deep recurrent neural network algorithms and achieve highest accuracy for the detection of malicious and safe links. Fakhruzzaman et al.[27] proposed a neural network with multilingual bidirectional encoder representations from transformer model BERT was proposed for the detection of news clickbait. Applied BERT on Indonesian news headlines collected from various news sites with sentence embedding's features and achieved 91% accuracy for the identification of news headlines. Kumari et al.[28] to

Ref	Models	Dataset	Features	Results	Limitation
[11]	RF	Clickbait Challenge 2017	Text features	82	Relies only on top sixty features; lacks contextual semantic representation.
[12]	AdaBoost	Clickbait news	Text features	91	Focused on feature engineering; limited generalization beyond dataset.
[13]	SVM	online news outlets	Text features	97	Compared multiple algorithms but lacked interpretability and semantic analysis.
[14]	LR	BollyBAIT	Word embeddings	95	Achieved accuracy with minimal features; limited exploration of deep models.
[15]	NB	Clickbait headlines	non-word features	89	Focused only on Arabic dataset; domain-specific and lacks cross-lingual evaluation.
[16]	RF	Arabic news	FV-ANOVA	92	Lower accuracy with traditional ML models; limited contextual representation.
[17]	DT	Indonesia news	Text features	71	Expanded dataset but relied on conventional ANN/LSTM; limited interpretability.
[18]	RF	Turkish clickbait	Word2vec	97	Compared to NB and LSTM, it lacks robust dataset diversity.
[19]	NB	Clickbait news	Text feature	96	Limited exploration of deep models.

Table 2. Summary of existing research studies employing ML models for clickbait detection.

address the issue of identification of news clickbait, they proposed method gathers data from a dissimilarity matrix and analyzes those using ML classifiers after using SBERT to generate features from headlines and paragraphs. With an accuracy of 0.84 the SVM classifier with six dissimilar sentences earned the best whenever evaluated on two real-world datasets. Broscoteanu et al.[29] used various DL and ML models include RF, SVM, BiLSTM and BERT applied on Romanian Clickbait Corpus dataset that consists of 8313 news headlines samples labeled as clickbait or non-clickbait for the identification of news clickbait. Ensemble base model achieved highest accuracy eighty-nine for the detection of news clickbait. This was confirmed by Yadav et al.[30] method for detecting news clickbait using BiLSTM was proposed and uses a dataset from multiple sources to produce a Bi-LSTM clickbait detection algorithm and collected from various news sites including Vilanova, The Odyssey, Buzzfeed and that scoop Bi-LSTM perform better than compared to different DL methods include SVM, GRU, with 93% accuracy.

Hashmi et al.[31] presented a robust framework for the detection of fake content using three datasets include WELFake, Fake News Prediction and Fake Newsnet. Used hybrid approach of both ML and DL models such as CNN + LSTM and RNN with fast Text embedding, achieved highest accuracy 97% for the identification of news clickbait. Furthermore, transformer-based models such as XLNet, BERT and RoBERTa-Large on three datasets and achieved the best accuracy. Abualigah et al.[32] proposed a news methods had been developed for the identification of news clickbait and fake news content using unsupervised learning model with Glove feature embeddings. Summary of existing research studies employing DL models for clickbait detection are displayed in Table 3.

The purposed methods include DL algorithms CNN, DNN and LSTM applied on Corpus fake news dataset and achieved best accuracy for the detection of news clickbait and Mallik et al.[33] A hybrid framework was developed for fake news detection using DL models include LSTM and BERT with word2vec embeddings. This approach was to use stacked LSTM layers to acquire topic-relevant salient features from news articles and generate context-free, data-agnostic feature vectors. Performance is enhanced by hyper parameter adjustments. On four datasets, it performs better than both conventional and novel models. Alhanaya et al.[34] a DL approach utilized CNN to detect Arabic news content within Arabic dataset. The performance of three optimizers Adam, RMSprop and Adagrad was compared. After Pre-processing and Word2Vec addresses yielded the best results for the CNN model, with Adam and RMSprop displaying the highest effectiveness for the detection of news clickbait content Even though many studies have investigated the use of both ML and DL methods of detecting clickbait, there are still several limitations. Other classical ML-based approaches like RF, SVM and LR usually use handcrafted features, and they do not reflect the underlying semantic and contextual meaning of headlines. DL models such as LSTM, Bi-LSTM, CNN, and hybrid structures demonstrate better results, but are often trained using language specific or domain limited datasets and as such their generalizability across various news sources and languages is not as strong. BERT, RoBERTa, and XLNet are transformers that have been shown to be highly accurate but are computationally intensive and frequently lack enough interpretability of their results. Furthermore, much of the literature focuses on reporting performance measures without offering critical analysis of why some models perform better than others, and there is a lack of literature that integrates explainability models that would close the gap to understandable and interpretable clickbait detection.

In recent research Ahmad et al.[35] pretrained language models which are based on transformers (including BERT, GPT-2, and XLNet) are used to solve the problem of misinformation on social networks on the Internet. The authors provided proof of the potential of attention mechanisms to include contextual subtlety in short texts using a large corpus of tweets about the 2020 U.S. election and resulted in a considerable decrease in the

Ref	Models	Dataset	Features	Results	Limitation
20	LSTM	News Headlines	POST	97	Limited to linguistic and POS features.
21	CNN	News headlines	Glove	95	Assessed only on a specific dataset, reducing generalizability.
22	LSTM	English news	Word embeddings	94	Combined DL models but lacked interpretability and scalability evaluation.
23	BERT	Webis-Clickbait-17	Word2vec	85	Computationally expensive and resource heavy.
24	ALBERT	Telugu clickbait headlines	Word2vec, Glove, fast text	93	Limited to Telugu dataset.
25	XGB	Twitter and Instagram post	FastText	88	Stacking classifiers effective but complex and less interpretable.
26	C-LSTM	News Headlines of US publishers	Word Embeddings	98	Proposed extension limited to malicious.
27	BERT	Indonesian news headlines	Sentence embedding	91	Lacks cross-domain testing.
28	SBERT	Webs Clickbait Corpus 2017	Sentence embedding	84	dataset specific.
29	BiLSTM	Romanian Clickbait Corpus	Word2Vec	89	Ensemble methods effective but dataset (Romanian) small and limited.
30	BiLSTM	News headlines clickbait	Word2Vec	93	Bi-LSTM effective but dataset size limited and domain-specific.
31	BERT	WELFake, Fake Newsnet	Fast Text	97	Lacks explanation of decision process.
32	DNN	Corpus fake news	Glove	98	Limited evaluation scope.
33	BERT	Fake news detection	Word2Vec embedding	84	Computationally heavy and dataset specific.
34	CNN	Arabic dataset.	Word2Vec embedding	77	Results are tied to specific optimizers and dataset.
35	XLNet	Tweets on 2020 U.S. election	Transformer-based embeddings	87	Domain-specific limited to text-only, scalability and real-world deployment not fully addressed

Table 3. Summary of existing research studies employing DL models for clickbait detection.

levels of misinformation. Although the research points to the potential of transformer architectures to address the problem of misinformation, it is limited by domain dependency (only to the U.S. election) and inability to analyze in multilingual and multimodal as well as difficulty in scaling to real-time, large-scale social media settings.

Critical analysis of prior approaches

Although pipeline-only models like BERT, RoBERTa, XLNet, and GPT-based models can be trained to exhibit state-of-the-art results in tasks of detecting clickbait and misinformation, they can be computationally intensive and can be unable to adapt to new domains in case of limited training data. Conversely, hybrid ML/DL methods will also be useful since they will be able to combine feature-based models (e.g., TF-IDF, POS-tagging) with contextual embeddings thus taking advantage of both explicit linguistic features and deep semantic representations[36]. These hybrid models can frequently be competitive with training costs cut and inference speed increased, and such models can be conveniently deployed on resource constrained systems. Besides, the explainability of conventional ML features is complementary to the black-box characteristics of transformers that enhances model explainability.

Proposed methodology

Research introduces a robust detection method that combines traditional ML with advanced DL approaches to improve clickbait identification systems. The framework begins with comprehensive data preprocessing, feature engineering and train model also explain. For the analysis of results, present experimental setup. The proposed methodology framework shows in Fig. 1.

Data preprocessing

The preprocessing of data is a crucial component in the pipeline of any model that enhances the accuracy of the model, improves the quality of data, increases interpretability, and lowers the complexity of the data. The raw textual data used in this paper was a collection of English news headlines that were categorized as a clickbait or non- clickbait. All the headlines were removed and rearranged in a systematic manner into a structured form that would be analyzed. The initial step in the preprocessing involved cleaning the data, which involved all the text being turned into lower case letters and deleting all the unnecessary information including punctuation marks, special characters, and additional spaces to reduce the noise and ensure uniformity among samples. The process of tokenization was subsequently done to break down each headline into separate words, which allows further linguistic analysis. The common but meaningless words, known as stop words, such as is, the, were filtered to exclude all non-meaningful words. This was followed by lemmatization which transformed every word into a base or authorized form (e.g., running run) and therefore enhanced semantic consistency. Words were also reduced to their root forms and in some cases, this was done by using stemming[37].

Feature engineering

Feature engineering is a crucial step used for selecting relevant features from datasets to improve model performance. In this research, Textual features include TF-IDF, POS and N-gram for ML models. DL models

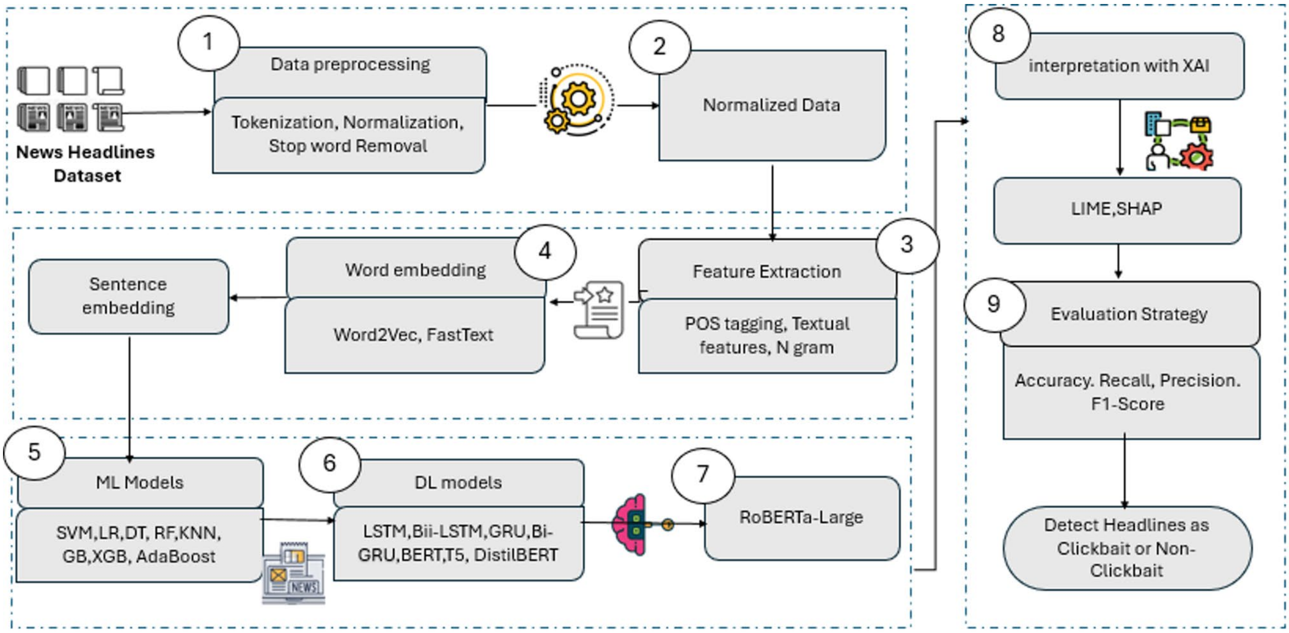

1
Data preprocessing
Tokenization, Normalization, Stop word Removal
News Headlines Dataset
2
Normalized Data
3
Feature Extraction
POS tagging, Textual features, N gram
4
Word embedding
Word2Vec, FastText
Sentence embedding
5
ML Models
SVM,LR,DT, RF,KNN, GB,XGB, AdaBoost
NEWS
6
DL models
LSTM,Bii-LSTM,GRU,Bi-GRU,BERT,T5, DistilBERT
7
RoBERTa-Large
8
interpretation with XAI
LIME,SHAP
9
Evaluation Strategy
Accuracy. Recall, Precision. F1-Score
Detect Headlines as Clickbait or Non-Clickbait

Fig. 1. Proposed methodology framework illustrating the complete pipeline of clickbait detection.

with word embeddings are applied on dataset. Various word embeddings include FastText, word2vec and advance sentence embeddings are applied on English news headlines dataset.

Textual features

Textual features are measurable attributes extracted from text data to represent its content and structure in a form suitable for machine learning, NLP, or statistical analysis. Determining whether news headlines are clickbait or not depends on the extraction of important linguistic patterns. The features of the text have been obtained with the help of the TF-IDF) scheme, according to which a weight of each term is provided by the frequency of the word within a single document in comparison with the frequency of occurrence within the whole collection of texts. The achievement of this task happens through TF-IDF analysis and POS tagging processes as well as N-gram models that allow measurement of word significance while maintaining text structure and content relationships[11].

TF-IDF TF-IDF is a number that indicates according to which a weight of each term is provided by the frequency of the word within a single document in comparison with the frequency of occurrence within the whole collection of texts. Clickbait headlines employ exaggerated terms as their writing style while non-clickbait headlines stick to a neutral tone in their content.

POS tagging POS tagging was used to extract syntactic structures of clickbait headlines. Every token was marked with its grammatical role (e.g., noun, verb, adjective), which made it possible to extract linguistic features that represent sensational or misleading styles. Clickbait headlines construct suspense through overused adjectives and verbs and adverbs but non-clickbait headlines stick to standard proper uses of nouns and descriptive names[38]. Insights from POS Tagging Verbs (VBG), Nouns (NN), adjectives (JJ), and verb etc. POS tagging example shows in Fig. 2.

N-gram Ngram is a contiguous sequence of *n* number of items words, characters out of the textual or spoken material provided. N-grams assist in analyzing common phrase patterns that appear in clickbait headlines through their observation of bigrams and trigrams[39].

Word embedding

Most of the NLP techniques employed for word embedding are the requirement of representing the word as vectors and capturing semantic relations. These techniques are used to capture word meanings and deal with homophones. Word embedding is used in text classification, sentiment analysis text summarization, and language modeling. Word embedding that are available include word2vec, glove and advance sentence embedding. In NLP tasks, such embedding's are essential for ML and DL models since they acquire syntax and semantic details[33].

Word2Vec It represents and captures semantic relationship that predicts surrounding word in context. Due to the deep neural network, it has an especially low computational cost. From it we have skip gram gives more importance to frequent context word prediction and second is CBOW in this model target word is predicted by the model using its context. For the identification of sporadically occurring terms, it proves useful while it maintains the bag of words architecture. The primary purpose of this techniques is to use for maximizing the probability of predicting words efficient training[33]. The Word2Vec Skip-gram model functions to optimize the prediction probability of context terms relative to the target word. The implementation of the probability uses the SoftMax function show in Eq. (1).

$$P\left(\omega_{c} \mid \omega_{t}\right) = \frac{exp\left(V'_{\omega_{c}} \cdot V'_{\omega_{t}}\right)}{\sum_{\omega \epsilon V} exp\left(V'_{\omega_{c}} \cdot V'_{\omega_{t}}\right)} \tag{1}$$

FastText This feature is the extension of word2vec used for fast and efficient training and support for sentence classification. It is made for high throughput and it is used for tasks such as text classification, language modeling and word embedding[40]. The context probability of a word can be calculated through Eq. (2).

$$P\left(\omega_{c} \mid \omega_{t}\right) = \frac{exp\left(\sum_{g\in G} V'_{g} \cdot V_{\omega_{c}}\right)}{\sum_{\omega \epsilon V} exp\left(\sum_{g\in G_{(wt)}} V'_{g} \cdot V_{\omega}\right)} \tag{2}$$

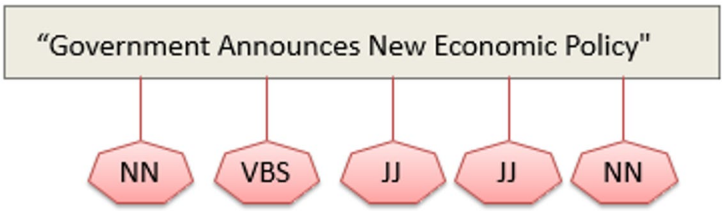


Fig. 2. Standard POS tagging process demonstrated on a sample news headline.

Sentence embedding's This technique is used for understanding the semantic meaning of a sentence and for the classification of similar sentences and information retrieval. More specifically, they can convey the meaning of not only specific words but the whole sentence[41]. NLP data is used for encoding semantic meaning putting sentence. Suitable for problem solving at the sentential level, for example text categorization, translation identification, and measuring meaning resemblance. The use of sentence embedding is common in summarization, question answering comparison among others[42]. The procedure involves calculating the weighted average of word embedding's followed by Principal Component Analysis or Singular Value Decomposition to eliminate shared factors. To compute sentence embedding can be used in Eq. (3). How sentence embedding work with English news headline show as Fig. 3.

$$Sentence\ embedding = \frac{1}{N} \sum{}_{i-1}^{N} a_i v_{wi} \tag{3}$$

Experimental setup

To use the method, we needed to use a large set of properly labeled English news headlines in experiments with both ML and DL algorithms.

Dataset

This research used a collection of data that was edited by Chakraborty et al. that is publicly available on GitHub and is permitted to be used by the author composed of 32,000 news headlines (clickbait 16,000 and non-clickbait 16,000) to maintain the balance of the classes. Every case has two variables, the text of the headline, and a binary variable that reflects whether the case is consistent or not.

To train the model a systematic preprocessing pipeline was used to prepare the data. This involved text normalization, deleting stop words, punctuation, digits, and special characters, stemming, lemmatization and tokenization. The steps guaranteed linguistic consistency and minimized noise, turning the dataset into both the traditional ML and the DL work. Nevertheless, one should keep in mind that the data is restricted to English language news headlines, and this can limit the ability to generalize the research findings to other languages or other media scenarios. Moreover, since the data is based on a particular period and sphere, there can be certain biases, especially in the aspect of the writing style, cultural orientation, and platform preferences. These remarks explain why future research work is essential in multilingual, multimodal, and domain-adaptive dataset research.

Baseline models

ML is a subset of AI that enables systems to learn from data and improve their performance on specific tasks without being explicitly programmed. Several features engineering with TF-IDF, POS tagging features and N gram with ML models include SVM that determines the best hyper plane for a dataset to have the largest margin between two classes on it, KNN is lazy learning algorithm which essentially sorts data points according to their proximity to *k* nearest neighbors[43], DT that divides data based on feature values making branches until arriving at a target, easy to understand, LR that used for handling discrete values or those values which lie within a range

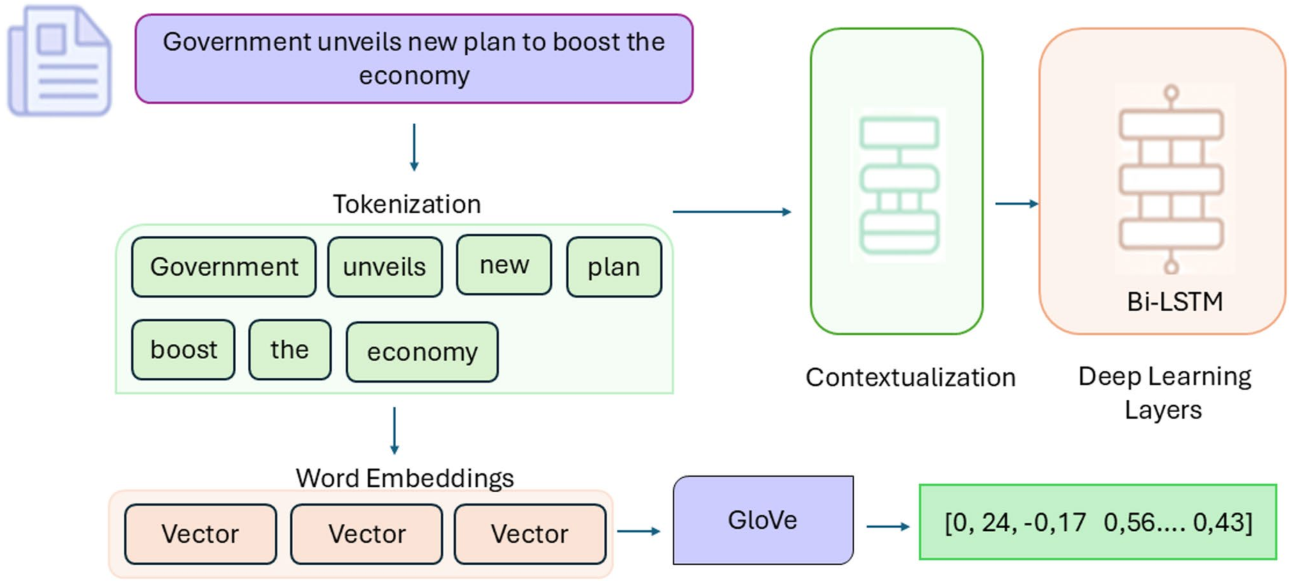


Fig. 3. Sentence embedding workflow showing how semantic representations are generated from English news headlines.

of 0 and 1 only, GB that is use for Analyzing large data, futuristic data populating, fraud detection, and ranking systems, XGB is a novel gradual accruing technique developed for efficiency of algorithm. It also includes some techniques to prevent over fitting; RF is an ensemble method that forms several decision trees and makes the final prediction by employing majority classification and AdaBoost ensembles learning mechanism that by adjusting weights for misclassified data points tries to enhance weak classifiers a lot. DL are advanced ML models that use different layers of artificial neural network to simulate human learning. DL model were also apply on English news clickbait dataset include LSTM that generated to solve the issue of passing fades and Bidirectional Bi-LSTM that process of input sequence with both forward and backward in recurrent neural network model, GRU is a type of RNN model use for sequential data[44] and Bi-GRU that uses two GRUs to process the input sequences in both a forward and backward direction with word2vec, FastText and sentence embedding's. As a DL architecture Transformer model serves NLP tasks through machine translation and text classification as well as text generation. Transformers operate entire sequences simultaneously through parallel processing which enables effective handling of large-scale text processing. Bidirectional Encoder Representations from Transformers (BERT) functions as a DL model based on the Transformer architecture which exclusively uses its encoder subsection and Robustly Optimized BERT Pre training Approach (RoBERTa-Large) are applied on English news dataset[45] .

Proposed model

The robustly optimized BERT pre training approach named RoBERTa-Large provides performance improvement through optimized training procedures. The transformer-based architecture remains the same as BERT, yet the next sentence prediction is eliminated while the model receives a larger dataset combined with dynamic masking to enhance contextual understanding. Research confirmed next sentence prediction removal makes no notable impact on model functionality thus RoBERTa-Large eliminates the next sentence prediction from its system[46]. The approach of RoBERTa-Large revolves around Masked Language Model without the next sentence prediction task since it prefers to concentrate on MLM exclusively. RoBERTa-Large uses a dynamic masking framework to improve generalization because clickbait headlines frequently employ deceptive or complex wording. RoBERTa-Large achieves better detection of these subtle elements because of its improved capabilities. The high precision requirements of NLP tasks associated with clickbait detection favor the use of RoBERTa-Large models due to their updated features from BERT[23]. The architecture diagram of RoBERTa-Large is shown in Fig. 4.

RoBERTa is a transformer-based model that is trained with dynamic masking and pretrained on a larger scale with a state-of-the-art performance on numerous text classification tasks with only a fraction of the computational budget of LLMs. Second, newer LLMs like GPT-3.5 or LLaMA-2 are more powerful, but demand a lot of computational resources, and are therefore less reproducible in experimentation in academia. Lastly, by working on RoBERTa, it is possible to make a direct comparison with the previous work of clickbait detection and methodological consistency is guaranteed.

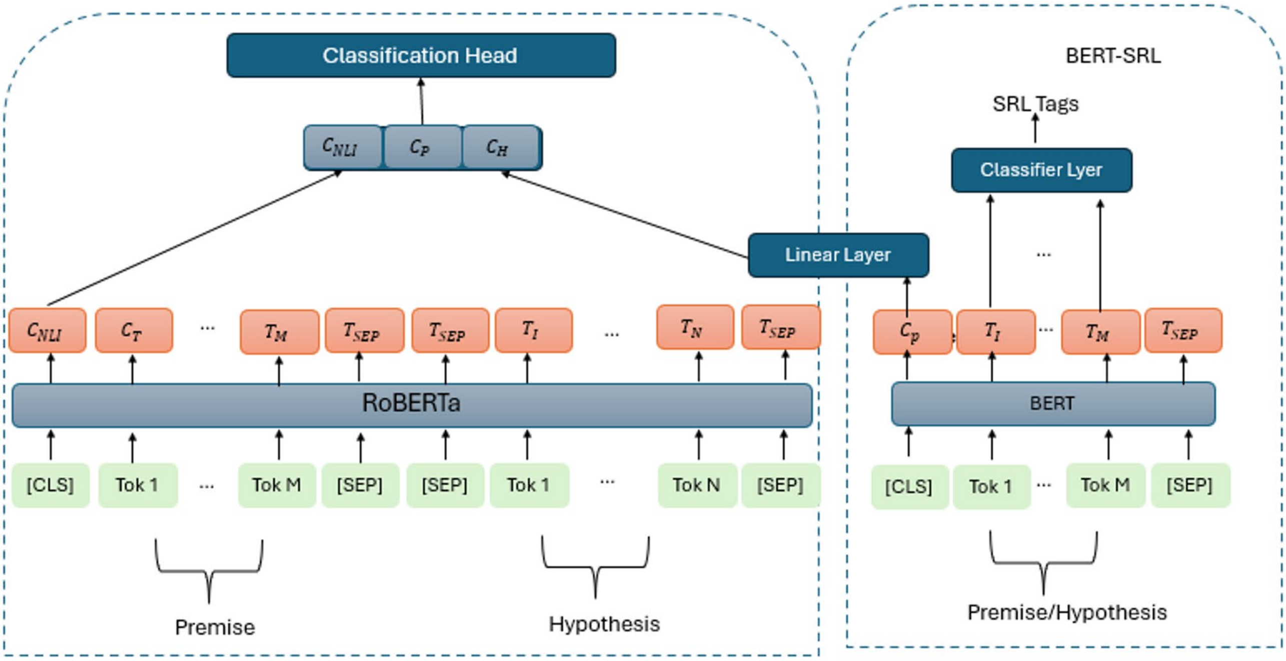


Fig. 4. Architecture of the RoBERTa-Largemodel depicting the deep transformer layers and fine-tuning process for clickbait classification.

Evaluations metrics

For evaluating the performance of DLmodel and to ensure the model is effective, give accurate results for some problems, perform various evaluation metrics include measure accuracy, precision, recall, f1 score and ROC-AUC[47].

Accuracy: The correctly prediction of news headlines out of the total instance.it give generally idea of the model.

$$Accuracy = \frac{(True\ Positives\ +\ True\ Negatives)}{Total\ Instances} \tag{4}$$

Precision: The percentage of positive instances out of the predictive news headlines.

$$Precision = \frac{(True\ Positives\ +\ False\ Positives)}{True\ Positives} \tag{5}$$

Recall: The proportion of correct positive instance out of the actual news headlines. This is evaluated to ensure most clickbait headlines are detected.

$$Recall = \frac{(True\ Positives\ +\ False\ Negatives)}{True\ Positives} \tag{6}$$

F1-score: The harmonic means of recall and precision that ensures the balance of precision and recall. This is preferred when the dataset is imbalanced.

$$F1 - Score = 2 * \frac{(Precision\ +\ Recall)}{(Precision *\ Recall)} \tag{7}$$

Receiver operating characteristic—area under curve

AUC-ROC means the way the model fits in prediction and facilitates true positive rate against the false positives and was applied when dealing with class imbalance, comparability of different kind of models.

In this paper, we use F1-score, precision and recall as our top priority since clickbait detection is a binary classification task, and the tendency of imbalanced data is inherent in it. Although accuracy gives overall analysis, it may not be accurate when there are more non-clickbait samples of data as compared to the clickbait samples. The common sense is that one has to be accurate to make certain that the marked clickbait headlines are false, minimizing the number of false alarms. Recall, conversely, prevents the possibility of missing real cases of clickbait, which is essential to reduce the dissemination of misinformation. The harmonic measure of precision and recall, F1-score, is a more balanced measure, especially with class imbalance. We also put ROC-AUC as a complementary measure to have the overall discriminative power of the models over thresholds. Nonetheless, due to the practical necessity to both properly detect clickbait and not to miss too many false positives, F1-score is the most appropriate measure of such a task.

Results and discussion

In this section, present exploratory data analysis (EDA) techniques that is a process of assessing and describing dataset with the aim of identifying trends, outlying observations and assessing various hypotheses graphically and numerically. The Fig. 5 shows the percentage of clickbait and non-clickbait headlines. The sample size is almost even in terms of distribution, which does not result in the bias of the classification models that prioritize one of the classes. A balanced dataset gives a level playing field of training as well as enables models to generalize more in testing. The headline length analysis is a significant process in preparing and exploring textual data more often in cases like clickbait detection. The summary statistics for headline lengths show in Fig. 6. This value is a comparison of the lengths of headlines in the category of clickbait and non-clickbait. The headlines generated by clickbait are often shorter and punchier whereas non-clickbait headlines are longer and more descriptive. This comparison suggests length is a beneficial discriminative attribute, given that models can use structural cues, in addition to semantics. Figure 7 presents the analysis of the frequency of words in three categories. Clickbait headlines have been known to be full of terms that arouse curiosity like shocking and revealed. Non- clickbait headlines, on the contrary, are based on the topic specific factual words. The general corpus is an indicator of amalgamation though the lexical distinction is easy enough to the classification cues of the models. Applying the method makes it easier for researchers and data scientists to notice words within articles that are often found in clickbait pieces. In comparison, sensible headlines usually skip flamboyant words and feature terms such as "report", "announces" or "study". The word clouds that were generated demonstrate the most common patterns of lexicons in clickbait and non-clickbait headlines. The use of words shocking, amazing, secret, etc. prevails in the category of clickbait which represents the reliance on the vocabulary of sensations and curiosity. The non-clickbait category, on the other hand, has been described as having neutral and fact-based words like report, government and update. These comparisons validate the lexical inclination of clickbait headlines to capitalize on exaggerated adjectives and action-oriented expressions to pay attention to user desires, as it was discovered in the literature previously.

Results of ML models

TF-IDF served as the feature representation for clickbait detection and the outcome provided valuable information about the performance of distinct MLmodels. The highest accuracy of 92% was reached by SVM

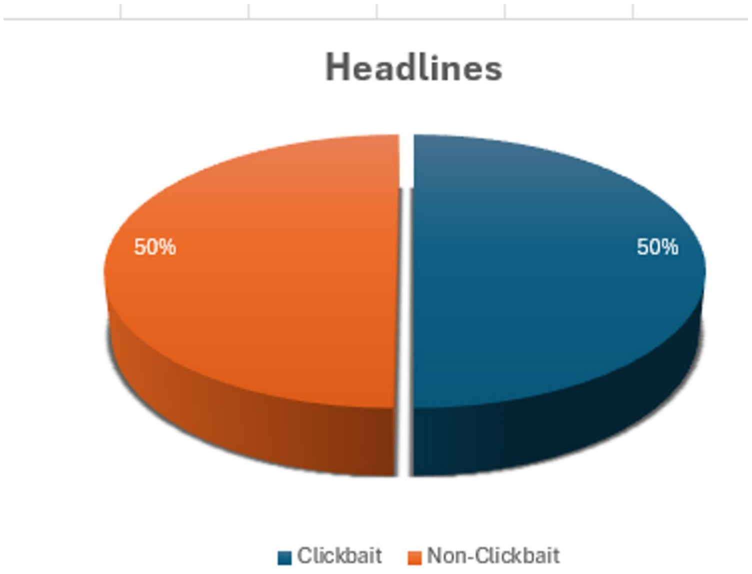


Fig. 5. Data distribution of the English news dataset across different categories, showing dataset balance.

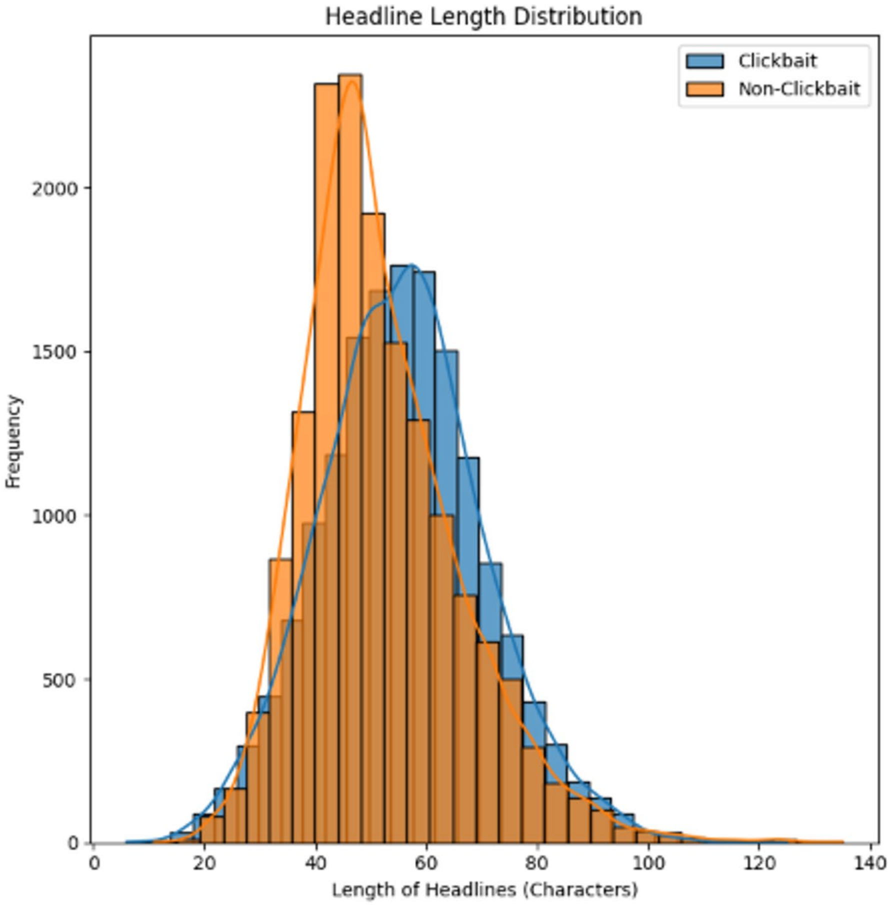
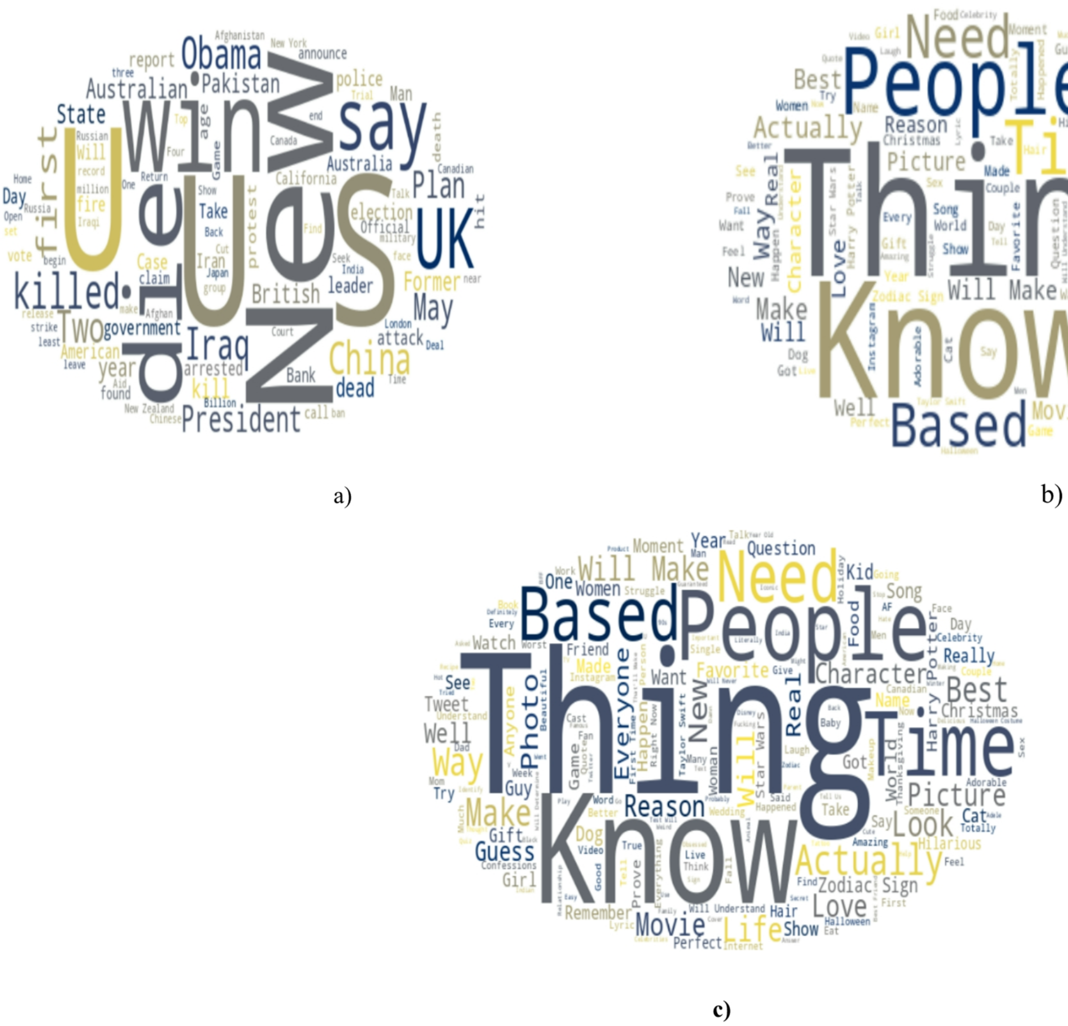


Fig. 6. Analysis of news headline length showing the variation in average word count across clickbait and non-clickbait samples.



a)

b)

c)

Fig. 7. Visualization of most frequent words in (a) Non-clickbait, (b) Clickbait, and (c) Overall news headlines, emphasizing linguistic differences between categories.

combined with LR which proves their exceptional ability in clickbait and non-clickbait headline classification, as displayed in Table 4. When processing high-dimensional textual data using TF-IDF the model offers effective capture of word importance patterns at the same time. DT achieved 88% accuracy along with KNN performed slightly better for the identification of news clickbait and non-clickbait. RF reached 90% accuracy in the prediction model. The accuracy rates for GB and XGB models were lower than other methods since they reached 79% and 87% respectively. The detection of clickbait becomes more feasible when researchers leverage POS patterns because they accurately identify the core syntactical structures used in clickbait headlines. POS tagging functions as a fundamental characteristic in clickbait detection because it performs analysis on the grammatical design of news headlines. Performance of all applied ML classifiers with feature selection results are displayed in Table 4. The best performance reached with POS tagging reached 65% GB along with RF and XGB. The ensemble learning methods excel at detecting refined syntactic patterns in POS sequences which enables them to enhance separation of clickbait from non-clickbait headlines, comparatively shown in Fig. 8.

The classifier models DT and AdaBoost achieved 64% accuracy while detecting POS tag distributions in headlines, but their ability remained moderate. SVM and LR achieved the same 63% accuracy rate in news headline identification as clickbaits or non-clickbait. KNN showed similar success along with 61% accuracy and high values of f1-score, recall and precession. N-Gram features serve as crucial elements of clickbait detection since they recognize common word arrangements found within clickbait headlines. The systems of SVM & Logistic Regression achieved the best results by reaching 92% accuracy for news clickbait classification. Text classification success stems from using N-Gram features with these models because they teach specific word sequences that define clickbait headlines. The performance of RF and DT is comparable to each other as they produce detection rates of 90% and 88%. The text classification performance of these models becomes stronger when N-Grams enable them to create better detection rules for clickbait. XGB and KNN models demonstrate better outcome for detecting news clickbait headlines while achieving 87% accuracy with Ngram feature as the input. The AdaBoost model achieved superior outcomes on English News detection by using Ngram features to deliver an 78% success rate. This Fig. 8 gives the accuracy of various ML classifiers. Ensemble models and tree-

Features	Models	Accuracy	Precision	Recall	F1-Score
TF-IDF	**SVM**	**92**	**92**	**92**	**92**
	LR	92	92	92	92
	DT	88	88	88	88
	KNN	88	87	88	88
	GB	79	84	80	79
	RF	90	90	90	90
	XGB	87	89	87	87
	AdaBoost	78	82	78	78
POS	SVM	63	65	64	63
	LR	63	65	64	64
	DT	64	64	64	64
	KNN	61	61	61	61
	GB	**65**	**66**	**66**	**65**
	RF	65	65	65	65
	XGB	65	66	66	66
	AdaBoost	64	65	65	65
Ngram	**SVM**	**92**	**92**	**92**	**92**
	LR	92	92	92	92
	DT	88	87	87	87
	KNN	87	87	87	87
	GB	79	84	80	79
	RF	90	90	90	90
	XGB	87	89	87	87
	AdaBoost	78	82	78	78

Table 4. Performance of ML classifiers with feature selection (results in %). Significant values are in bold.

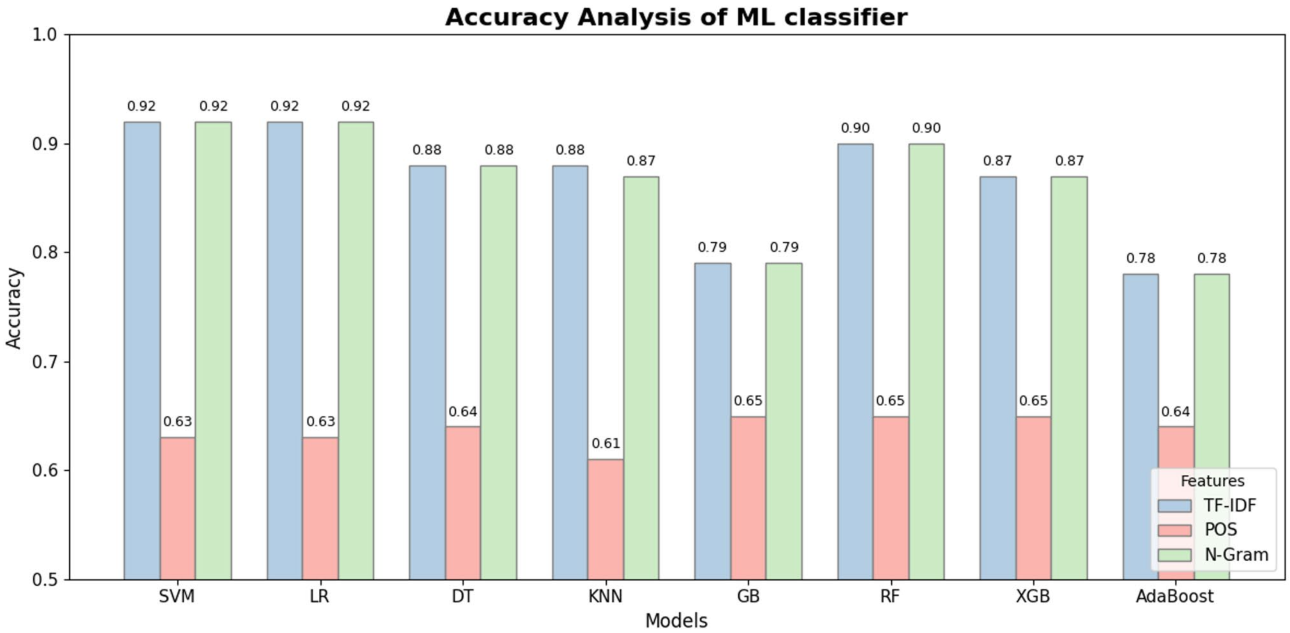


Fig. 8. Comparison of accuracy scores among ML classifiers applied for clickbait detection, demonstrating model performance variations.

based models are always more effective than simple linear models, indicating that complex non-linear decision boundaries are more appropriate in finding out the subtle linguistic patterns of clickbait detection.

Results of DL model

As the generalizability of LSTM and the analysis of Bi-LSTM has illustrated, setting proper parameters is essential for training and test of models, as well as for desired performance enhancement. Removes the first and last tokens of sequence since the inputs contain variable length and set them up to 3000 as an input vector length. The first subset is Vocabulary size 2000 consisting of the most basic 2000 words on which model should ideally be trained, and output dimension is 300. The numbers of units are set at 120 that specify the pattern learning capacity of the model and dropout rate 0.5. Activation function is sigmoid. To make it efficient and adaptive learning, an Optimizer set for Adam and learning rate set 0.001 by default. To avoid over fitting, the numbers of Epochs are set at 10 and Batch size is set to value 32. Apply different DL models include LSTM and Bi-LSTM, GRU and Bi-GRU with three embedding features such as Word2Vec, FastText and sentence embeddings.

The LSTM model has a clear difference in performance with regard to the type of embedding employed. Using Word2Vec embeddings, the model attains 85% accuracy and F1-score of 85%, which means that it learns sequential relationships. Nevertheless, Word2Vec does not understand word contextualities because it gives fixed word representations. Performance is greatly improved when Sentence Embeddings are added, reaching a high accuracy of 94% and a final F1-score of 94% indicating that the contextual understanding of a complete sentence is essential in detecting clickbait since the model is able to comprehend the complete sentence, not just a single word representation. Conversely, FastText embeddings achieve merely 77% accuracy and 78% F1-score, as although the embeddings are able to capture subword information, FastText remains unsuccessful at short and high-context headline, which is common with clickbait materials. The Bi LSTM model is more effective than the standard LSTM in all forms of embedding because it gives a text a forward and backward analysis, which makes the context easier to understand. Bi-LSTM has 87% accuracy using Word2Vec embeddings, a slight improvement over unidirectional LSTM demonstrating that having a flow in both directions assists in preserving the dependency between words before and after the main keyword. In combination with Sentence Embeddings, the Bi-LSTM gets the highest overall performance of 95% accuracy and 96% F1-score. This demonstrates that both subtle emotional signals and manipulative language, which constitutes clickbait, can be comprehended in semantic-level features of contextual embeddings, processed in both directions. Using FastText embeddings, the model has a 79% accuracy, with a slight improvement that can be attributed to the fact that the contextual advantage of FastText is limited.GRU model is a simpler but equal in power substitution to the LSTM, and its performance is preserved, but at a reduced computational cost. Sentence Embeddings GRU has an accuracy and F1-score of 94% and 96%, respectively, as with Bi-LSTM, indicating that contextual embeddings can be used with only a few fewer parameters. This performance makes GRU an appropriate choice when the speed of the models is an important factor and it is used in real time.

GRU holds a sTable 89% accuracy with Word2Vec embeddings, in support of the fact that without deep-contextual layers, in it can still learn meaningful sequential patterns. But with FastText, the performance decreases to 78%, which supports the fact that subword-level representations play a minor role in comprehending the intent or the tone of brief English headlines.Bi-GRU model builds on the strengths of GRU by working on both directions, which further improves the contextual understanding. Sentence Embeddings produce Bi-GRU with the best accuracy of 95% and an F1-score of 96%, which is equivalent to the best Bi-LSTM. This demonstrates that bidirectional recurrence that is combined with sentence-level semantics provides strong feature extraction and semantic model generalization. The second setup (probably retended with Word2Vec or FastText) attains about 84% accuracy and 83% F1-score, which implies that the results of these models are significantly lower when there is no contextualized input. The findings highlight that Bi-GRU similarly to Bi-LSTM is most advantaged by the embeddings which encode full sentence meaning. BERT model performed a clickbait classification task on English news headlines to distinguish between clickbait and non-clickbait categories. BERT applied its transformer architecture to read text from both directions which successfully recognized the distinctions between deceptive headlines and factual ones through contextual understanding. BERT proved its superiority over ML traditional models and DL approaches through many experimental tests by achieving higher accuracy rates with better precision and recall together with F1-score. The 95% accuracy rate exhibited by BERT exhibited higher performance in extracting misleading headlines compared to LSTM, GRU, and Bi-GRU models thus establishing its efficiency for detecting exaggerated headlines. Performance of all applied DL classifiers with feature selection results are displayed in Table 5.

The results obtained show BERT functions as an advanced clickbait detection system for news headlines. BERT model performance assessment for clickbait detection included multiple iterations of training along with validation loss evaluation and accuracy assessment. The training loss visualizes how the model learns during its operations. The model starts with a high training loss before it gradually learns from dataset inputs which reduces the loss values. The model prevents overfitting through the validation loss pattern which follows a similar trend. Performance data for the model appears through the accuracy graph as training advances with each epoch. The model begins by achieving poor results but develops advanced identification capabilities between clickbait and non-clickbait headlines throughout training. BERT demonstrates strong ability to separate clickbait from non-clickbait news headlines through its confusion matrix thus becoming a robust transformer-based model for automatic fake news detection systems and content moderation solutions.

Proposed model results

At the time of implementing RoBERTa-Large for clickbait detection researchers need to focus on adjusting hyperparameters because this step affects the model's overall performance. The learning rate of 2e-5 is specifically designed to make weight updates small enough for avoiding overshooting during effective convergence. The model uses a batch size of 32 as it meets both computational constraints and stability needs to perform effective gradient updates without increasing memory usage. All RoBERTa-Large hyperparameter are displayed in Table 6.

Model	Embeddings	Accuracy (%)	Precision	Recall	F1-Score
LSTM	Word2Vec	85	87	85	85
	FastText	77	79	78	78
	Sentence Embeddings	94	95	96	94
Bi-LSTM	Word2Vec	87	88	87	87
	FastText	79	80	80	80
	Sentence Embeddings	95	96	96	96
	Word2Vec	89	89	89	89
GRU	FastText	78	79	79	78
	Sentence Embeddings	94	94	95	96
	Word2Vec	89	89	89	89
Bi-GRU	FastText	84	84	84	83
	Sentence Embeddings	95	97	97	96
BERT	Pre-Trained Embeddings	95	95	96	95
T5		87	88	89	88
DistilBERT		91	92	91	92
RoBERTa-Large		**97**	**98**	**97**	**98**

Table 5. Performance of DL classifiers with feature selection (results in% acc). Significant values are in bold.

Parameters	Values
Learning rate	2e-5
Batch size	32
Max sequence length	128
Weight decay	0.01
Warmup steps	500
Gradient accumulation	1
Optimizer	Adam
Number of epochs	10
Loss function	CrossEntropy

Table 6. RoBERTa-Large hyperparameter.

The chosen sequence length of 128 allows most news headlines to fit smoothly within computational boundaries. A weight decay rate of 0.01 serves as an overfit prevention method by imposing penalties on weight modification. Training starts with 500 warmup steps that regulate the learning rate increase to promote stable optimization during the first part of the training process. The model performs weight updates for every batch through its gradient accumulation step set to 1. This optimizes the training process. The model undergoes 10 epochs of training to enable the suitable discovery of complex patterns in the dataset with minimal overfitting risk. The researchers utilized the RoBERTa-Large model which represented an optimized version of BERT for classifying English news headlines according to their clickbait nature. The model exploited strong pretraining approaches along with adaptive masking methods to understand complex abnormal patterns that occur in deceptive headlines. RoBERTa-Large delivered 97% accurate results in its classification tasks which exceeded performances of standard ML techniques and alternative DL approaches. The extensive training process across a big corpus enables RoBERTa-Large to understand semantic meanings more deeply and thus deliver superior performance. This confirms RoBERTa-Large as an excellent transformer model for identifying factual or misleading news headlines. The model begins with elevated training loss before decreasing it while absorbing important features from the provided data. Both training loss and validation loss demonstrate a related decreasing pattern. RoBERTa-Large demonstrates superior capability for separating clickbait from non-clickbait headlines during its extended training process according to the accuracy graph. Starting from a low point the training along with validation accuracy increases consistently through each additional training cycle. The clickbait detection capabilities of RoBERTa-Large surpass others because of its contextualized pretrained components and superior language modeling strength resulting in highest accuracy levels.

Two transformer-based models (T5 and DistilBERT) were in this research fine-tuned on the English clickbait dataset to determine how well they can detect false headlines. The T5 model demonstrated a total accuracy of 87%and a precision of 88%, recall 89% as well as F1-score of 88%, which is very good generalization and equal performance in the clickbait and non-clickbait groups. Nevertheless, DistilBERT was also found to be better than T5, with an accuracy of 91%, precision of 92%, recall 91% and F1-score of 92%. This could be explained by the fact that DistilBERT has an optimized encoder-only format, which enables it to capture the contextual subtleties more effectively and make effective sentence-level classifications without the extra sequence-to-sequence generation cost of T5. The increased accuracy and F1-score of DistilBERT indicate that it had fewer false positives and was more consistent in the precision and recall thus more applicable in real-time or larger scale clickbait detection. T5, conversely, have minor performance concessions because its generative architecture does not focus on actual classification but instead on the presence of text-to-text generation. On the whole, the findings underscore the idea that DistilBERT is a more useful and accurate as compared to T5.

Results interpretation with XAI

XAI increases transparency as it makes us aware of why a model makes certain predictions. Regarding the RoBERTa model that can be utilized in clickbait identification in English news headlines, XAI approaches such as LIME, SHAP provide the appreciation of the role of the input features (words or phrases) in the inference regarding the classification. LIME and SHAP are useful XAI methods, used in order to interpret the outputs of a model visually.

To more deeply analyze the drawbacks of the suggested clickbait model of detection, an error analysis process was performed on the misclassified samples. Figure 9. shows error by headline length that has been performed to assess the level of the model performance based on the relationship between the headline length and the distribution of False Positives (FP) and False Negatives (FN). This analysis gave a hint on the impact of textual features like the size of words used in a headline as it impacts on the accuracy of the model prediction. The length (measured in terms of word count) of each headline was estimated and matched against the outcomes of the prediction of the model. False Positives (FP) were non-clickbait headlines that were wrongly labeled as clickbait and False Negatives (FN) were those clickbait headlines that the model misinterpreted. the plots of FP versus FN versus headline length gave some interesting results, as the shorter the headline, the more it was likely to give a misclassification, probably because less context was available, whereas too long headlines gave rise to a superfluous or misleading word which shifted the boundary of the decision surface of the classifier.

A series of statistical significance tests were performed to determine whether the observed performance improvements of the proposed model can be attributed to random variation; these tests included t-test, ANOVA, Chi-square and Z-test. These tests were used to compare differences between model predictions, label distributions, and categories of headline length. The ANOVA findings showed that there was a statistically significant difference in the mean headline length between clickbait and non-clickbait classes ($p < 0.05$), which means that the headline structure is a measurably important factor in classification. The t-test showing that the improvements in models in terms of F1-score and accuracy are not just coincidental. The Chi-square test showed the high dependency between the categorical variable of the headline length and the labels of classes, stressing the importance of the structural characteristics in the detection process. This statistical validation is vital in establishing the reliability and robustness of the proposed RoBERTa based model over baseline ML and hybrid

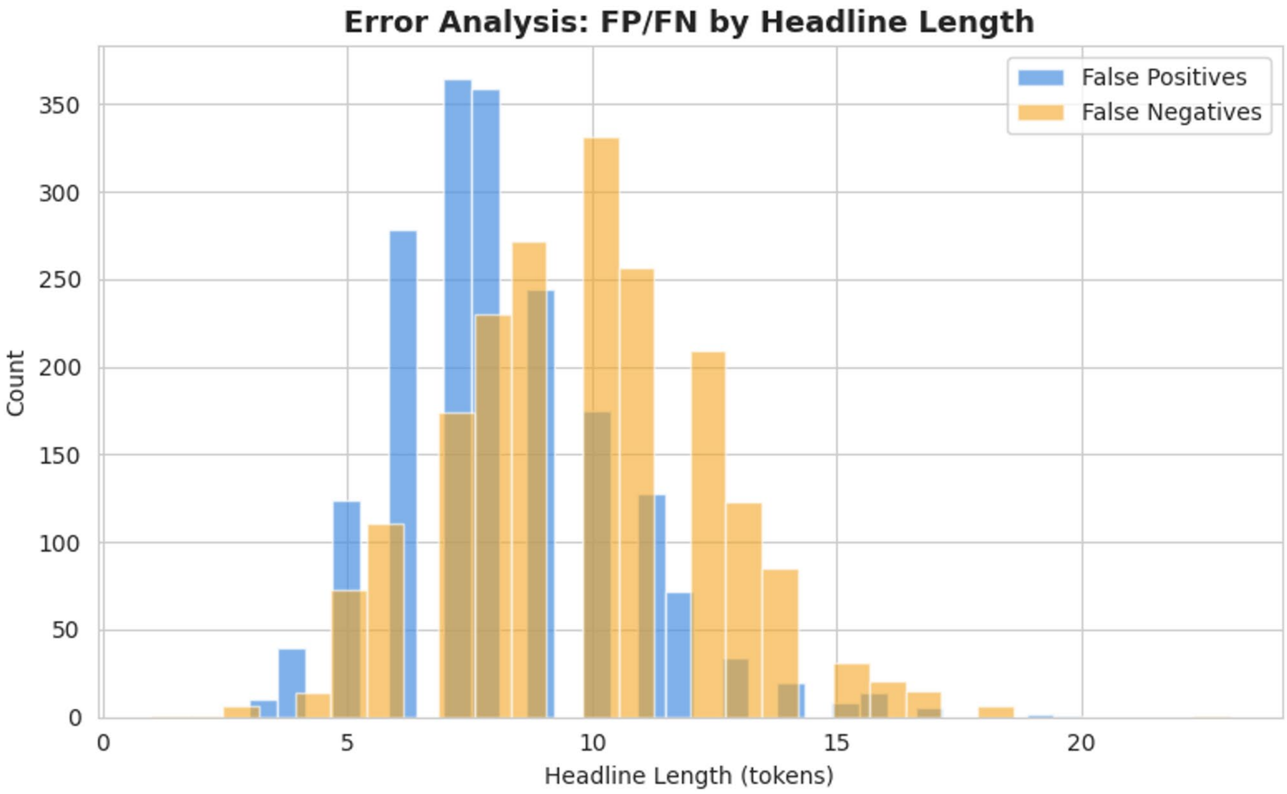


Fig. 9. Error analysis illustrating the distribution of FP and FN relative to headline length.

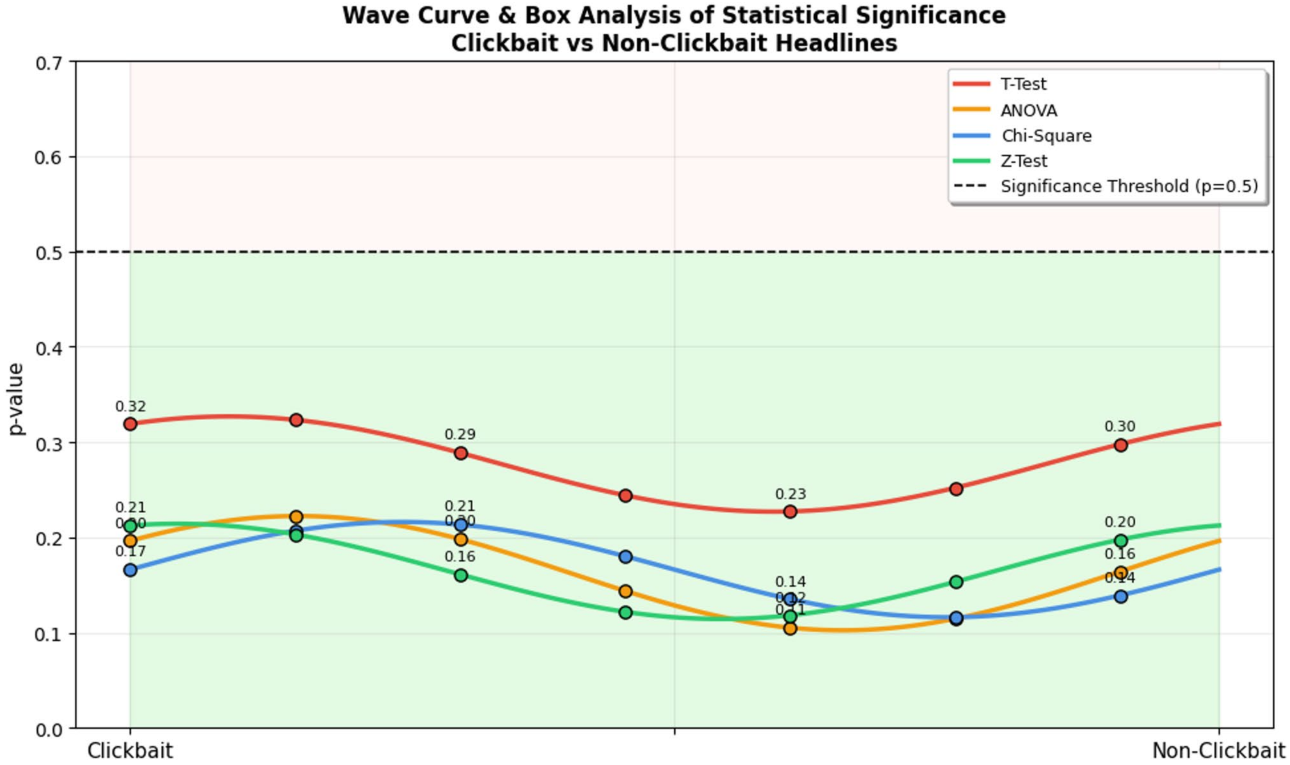
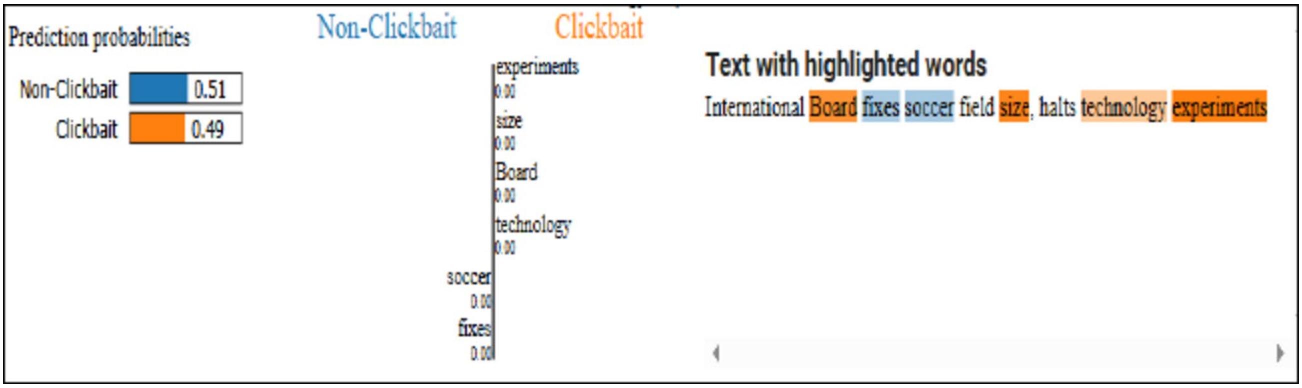


Fig. 10. Statistical significance comparison based on log-transformed p-values, indicating the relative strength of feature contributions.



International Board fixes soccer field size, halts technology experiments

Fig. 11. LIME visualization showing word-level feature importance and their positive or negative influence on the model's clickbait predictions.

DL schemes. These tests help to prove that the performance improvement is also statistically significant, which further supports the thesis that the proposed gains are not only motivated by empirical factors but rather have an analytical underpinning, which makes the model more legitimate when it concerns a real-life task, namely the recognition of clickbait. Statistical Significance Comparison (Log p-values) is shown in Fig. 10.

The LIME in the given Fig. 11 demonstrates the most perceptible words in a single news headline to cause the model to determine it as clickbait or not-clickbait, the longer the bar, the greater the conviction. To give transparency of the models, explainable AI methods (LIME and SHAP) were used on the predictions of the proposed model. The two techniques identified common language features as the ones that were conclusive in classification. As an example, LIME visualizations demonstrated that words with high affective or curiosity-related connotations were positively predictive of clickbait whereas factual nouns and named entities were better non- clickbait predictors. These results were further supported by distribution of SHAP value where sensational modifiers and verbs were ranked among high-impact features in making decisions.Such interpretations not only justify the discriminative capacity of the proposed model but also make the proposed model more trustworthy since the proposed model is based on linguistically meaningful cues and not the parasitic affinities.

Likewise, Fig. 12 shown the contribution plot of SHAP highlighted the significance of features across the whole dataset all over the globe. A positive SHAP value is associated with words of curiosity or sensations which are more likely to increase the likelihood of a clickbait label, whereas negative SHAP value is associated with accurate words that enhance non- clickbait prediction. The plot gives insight into the effect that various linguistic aspects have on the overall behavior of the model. Results suggest that on the same dataset, SVM performed the best with 92% accuracy and handled high-dimensional text well. Once applied to POS-tag-based features, XGBoost saw a drop in performance (65%) which suggests that using only syntax may not be enough to tell the differences between the labels. Sentence Embeddings achieves strong representation of the nuances in headlines, as it led to an accuracy of 95% for BiGRU across all architectures. It is evident that RoBERTa-Large at 97% and BERT at 95% managed to surpass the outcomes of both traditional and DL models. This proves that large pre-trained language models help detect small changes in language and settings, offering the most accurate results for categorizing news headlines.

Comparison with baseline models

To place the performance of our proposed model into perspective, we extend the performance of the baselines beyond traditional ML and previous DL architecture. Besides SVM, RF, LR, LSTM, BiLSTM, and RoBERTa-Large, we also use more recent foundation and distilled models: This covers our comparative analysis by filling the lightweight transformer baseline (DistilBERT), T5 and large-scale pre-trained models (RoBERTa-Large). It further emphasizes the performance/computational efficiency trade-off, improving on our analysis. Computational resource utilization with details is displayed in Table 7.

Performance interpretation and model comparison

We also give computational features of chosen models to supplement the results of accuracy. This aids in pointing out the trade-offs between lightweight ML classifiers and DL and transformer architectures that are resource intensive. Comparatively, classical ML classifiers, e.g., Logistic regression or random forest, can be trained and inferred on the same dataset in a matter of few seconds, and the training and inference do not incur any significant memory costs in terms of CPU/GPU. LSTM and Bi-LSTM models are intermediate: they perform very well but require longer training (usually minutes per epoch) and do not require a large amount of GPU memory. The above highlights the fact that transformers are the most accurate but computationally expensive,

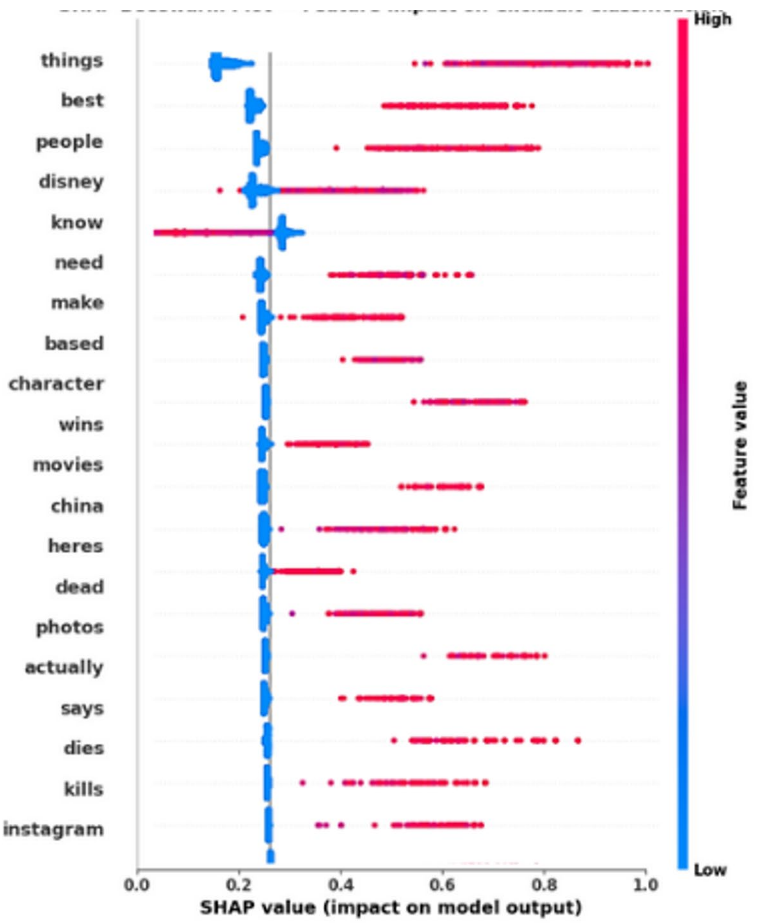


Fig. 12. SHAP summary plot presenting the overall impact and contribution of each feature on the model output across all samples.

Aspect	Details (RoBERTa-Large)
Model architecture	Transformer (RoBERTa-Large)
Framework used	PyTorch (hugging face transformers)
Training duration	3725 s (~ 1 h 2 m 8 s)
Floating point operations (FLOPs)	~ 356 M per forward pass
Average time per epoch	36.25 s
Total inference time	28 s
Inference time per sample	1.4 ms (20,000 samples)
GPU memory usage (peak)	~ 12.3 GB
Disk storage (model checkpoints)	~ 1.5 GB
CPU utilization (average)	~ 74%
RAM usage (during training)	~ 8.8 GB

Table 7. Computational resource utilization with details.

Model	Type	Parameters	Accuracy (%)	F1-score (%)	Training time (per epoch)	Inference speed	Resource demand
SVM	ML (Linear)	–	92	92	~ 15s	Fast	Low (CPU)
LSTM	RNN-based DL	10 M	95	96	~ 60s	Medium	Moderate (GPU/CPU)
BiLSTM	RNN-based DL	18 M	95	96	~ 75s	Medium	Moderate
T5	Seq2Seq Transformer	60 M	87	88	~ 85s	Slower	High (GPU/RAM)
DistilBERT	Transformer (distilled)	66 M	91	92	~ 40s	Fast	Moderate (GPU)
RoBERTa-Large	Transformer (large)	355 M	**97**	98	~ 120s	Slower	Very High (GPU ~ 12GB)

Table 8. Comparison with baseline models.

the ML models are lightweight but less precise, and the LSTM/Bi-LSTM are moderate between the two. In the comparative analysis, transformer-based models, especially RoBERTa-Large, exhibit a steady superiority to conventional ML and previous DL models like LSTM and Bi-LSTM. This higher performance is due to the deep bidirectional architecture and dynamic masking strategy of RoBERTa that allows it to detect richer contextual dependencies in news headlines. Handcrafted features in traditional ML models (e.g., TF-IDF or POS tagging) cannot capture the subtle semantics of clickbait language. Table 8 displays details of comparison with baseline models.

Likewise, sequential models such as T5 can only model long-range dependencies, whereas transformers are self-attentive and process all tokens at once, making them more likely to detect subtle clickbait signals, such as exaggeration or tone of voice. Additionally, pretraining RoBERTa on large-scale corpora with hyperparameters optimized to maximize performance (e.g., bigger batch size, longer training time, and dynamic masking of tokens) is also a factor in its strong performance on generalization across different types of news text. This enables it to perform better than even hybrid ML/DL combinations that rely on manual feature fusion. The improvement of performance is not only empirical, but statistically confirmed by ANOVA and t-test scores, which prove that the improvement is not due to mere fluctuations. Accordingly, the findings indicate that transformer based architectures such as RoBERTa are especially useful in semantic intensive and context sensitive classification tasks such as clickbait detection.

Comparative analysis with existing study

Our research compared the performance of RoBERTa-Large against clickbait detection studies which were carried out between 2021 and 2025. The previous research within this field utilized different ML and DL approaches which covered traditional NLP models together with transformer-based models and hybrid architectures. The research achievements in identifying clickbait proved effective but variation occurred from differences in data sets and variable feature engineering and model optimization procedures. Our proposed RoBERTa-Large model outperforms previous studies because it reaches the highest level of accuracy in detecting clickbait headline. Comparative analysis with Existing study models, dataset and results are diplayed in Table 9.

Conclusion

The identification of clickbait headlines should be accurate to prevent the diffusion of fake news and guarantee the reliability of online sources of information. This study shows hybrid approach that can be effectively used to classify clickbait headlines with the help of more sophisticated ML models SVM, LR, GB, XGB, RF and AdaBoost with TF-IDF, POS tagging features and Ngram and DL model include LSTM, Bi-LSTM, Gru and Bi-Gru with powerful features like Word2Vec, FastText and sentence embeddings. Furthermore, Transformer base model such as BERT and RoBERTa-Large also applied on English news dataset. Of these, the highest accuracy 97% was obtained by Transformer classifier such as RoBERTa-Large for the detection of news clickbait headlines. These

Ref	Year	Model	Dataset	Results(in %acc)
[24]	2021	ALBERT	Telugu clickbait	93
[48]	2022	RoBERTa	CLICK-ID	92
[49]	2023	BERT	News headlines	90
[34]	2024	CNN	Arabic dataset	77
[35]	2025	BERT, GPT-2, XLNet	Tweets on 2020 U.S. election	87
Proposed model		**RoBERTa-Large**	**English news headlines**	**97**

Table 9. Comparative analysis with existing study. Significant values are in bold.

models demonstrate its effectiveness as a transformer-based model by establishing itself as a reliable system for identifying clickbait headlines among factual headlines. This study has several limitations that should be acknowledged. First, the dataset is domain-specific, focusing only on English news headlines, which restricts generalization to other domains or informal text sources. Second, the approach does not address multilingual settings, where cultural and linguistic variations may influence clickbait patterns. Third, the study is limited to textual data and does not incorporate multimodal cues such as images or videos that are often part of online clickbait. Additionally, while hybrid ML/DL models achieve strong performance, interpretability remains limited due to the black-box nature of deep learning. Finally, transformer-based baselines were only partially explored, leaving room for deeper comparisons with more recent LLM approaches. Although this research achieves good results regarding identifying newsworthy clickbait headlines, there are a number of definite research paths to improve the research: Multilingual clickbait detectors: By using multilingual transformers, including mBERT or XLM-R, it is possible to increase coverage in a wide range of linguistic contexts. Transfer learning might also be used to assist in detection in low-resource languages in which annotated datasets are scarce. The proposed model can be deployed in real-time on lightweight and scalable architectures to implement real-time monitoring of the news feeds and social media streams. This would enable the early identification of the false content at the platform level. The incorporation of multimodal features, such as image or video thumbnails, can be utilized in numerous clickbait posts besides being based solely on textual cues. Development of multimodal fusion method Future work can focus on multimodal fusion methods, which involve NLP and computer vision models to provide strong detection. Because the strategies used in clickbait are domain-specific (e.g. politics, entertainment, health), future models would need to study domain adaptation methods in order to be accurate when used in new or changing genres of online news. Another future direction is to use XAI tools to make the decisions of deep and transformer-based models more transparent. Lastly, using real-time detection tools within social media and news sites could make the media more reliable and inform users.

Data availability

The dataset is freely available at: https://www.kaggle.com/datasets/amananandrai/clickbait-datasethttps://github.com/amaramuqadas87/Clickbait-Detection-in-News-Headlines-.

Received: 23 July 2025; Accepted: 24 November 2025
Published online: 02 December 2025

References

1. Raju, N. V. G., Nyalakanti, N., Kambampati, P., Kanthali, Y., Pandey, S. & Maithili, K. Clickbait post detection using NLP for sustainable content. *E3S Web Conf.* **430**, 01081 (2023) https://doi.org/10.1051/e3sconf/202343001081.
2. Yin, M., Wan, M., Lin, Z. & Jiang, J. Moralization-aware identity fusion for detecting violent radicalization in social media. *Inf. Process. Manag.* **63**(2), 104413 (2026). https://doi.org/10.1016/j.ipm.2025.104413.
3. Jung, A.-K., Stieglitz, S., Kissmer, T., Mirbabaie, M. & Kroll, T. Click me…! The influence of clickbait on user engagement in social media and the role of digital nudging. *PLoS ONE* **17**(6), e0266743. https://doi.org/10.1371/journal.pone.0266743 (2022).
4. Scott, K. 'Deceptive' clickbait headlines: Relevance, intentions, and lies. *J. Pragmat.* **218**, 71–82. https://doi.org/10.1016/j.pragma.2023.10.004 (2023).
5. Bronakowski, M., Al-khassaweneh, M. & Al Bataineh, A. Automatic detection of clickbait headlines using semantic analysis and machine learning techniques. *Appl. Sci.* **13**(4), 4 (2023). https://doi.org/10.3390/app13042456.
6. Khater, S. R., Al-sahlee, O. H., Daoud, D. M. & El-Seoud, M. S. A. Clickbait detection. In *Proceedings of the 7th International Conference on Software and Information Engineering*, ICSIE '18. 111–115 (Association for Computing Machinery, 2018). https://doi.org/10.1145/3220267.3220287.
7. Probierz, B., Stefański, P. & Kozak, J. Rapid detection of fake news based on machine learning methods. *Procedia Comput. Sci.* **192**, 2893–2902. https://doi.org/10.1016/j.procs.2021.09.060 (2021).
8. Muqadas, A., Khan, H. U., Ramzan, M., Naz, A., Alsahfi, T. & Daud, A. Deep learning and sentence embeddings for detection of clickbait news from online content. *Sci. Rep.* **15**(1), 1 (2025). https://doi.org/10.1038/s41598-025-97576-1.
9. Rahman, S. S., Das, A., Sharif, O. & Hoque, M. M. Identification of deceptive clickbait Youtube videos using multimodal features. In *Intelligent Computing and Optimization*. 199–208 (Springer, 2023) . https://doi.org/10.1007/978-3-031-50327-6_21.
10. Naz, A. et al. AI knows you: Deep learning model for prediction of extroversion personality trait. *IEEE Access* **12**, 159152–159175 (2024). https://doi.org/10.1109/ACCESS.2024.3486578.
11. Cao, X. et al. *Machine Learning Based Detection of Clickbait Posts in Social Media* (2017). arXiv: arXiv:1710.01977. https://doi.org/10.48550/arXiv.1710.01977.
12. Sisodia, D. Ensemble Learning Approach for Clickbait Detection Using Article Headline Features. *Informing Sci. Int. J. Emerg. Transdiscipl.* **22**, 031–044. https://doi.org/10.28945/4279 (2019).
13. Ahmad, I., Alqarni, M., Almazroi, A. & Tariq, A. Experimental Evaluation of Clickbait Detection Using Machine Learning Models. *Intell. Autom. Soft Comput.* **26**, 1335–1344. https://doi.org/10.32604/iasc.2020.013861 (2020).

14. Mowar, P., Jain, M., Goel, R. & Vishwakarma, D. K. Clickbait in YouTube Prevention, Detection and Analysis of the Bait using Ensemble Learning (2021). *arXiv*: arXiv:2112.08611. https://doi.org/10.48550/arXiv.2112.08611.
15. Bajaj, A., Nimesh, H., Sareen, R. & Vishwakarma, D. K. A Comparative analysis of classifiers used for detection of clickbait In news headlines. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*. 1410–1415 (2021). https://doi.org/10.1109/ICICCS51141.2021.9432123.
16. M. Al-Sarem et al. An improved multiple features and machine learning-based approach for detecting clickbait news on social networks. *Appl. Sci.* **11**(20), 20 (2021). https://doi.org/10.3390/app11209487.
17. Vincent, S. et al. Clickbait headline detection using supervised learning method. In *2022 IEEE International Conference on IoT & Intelligence Systems IoTaIS*. 408–412 (2022). https://doi.org/10.1109/IoTaIS56727.2022.9975866.
18. Genç, Ş & Surer, E. ClickbaitTR: Dataset for clickbait detection from Turkish news sites and social media with a comparative analysis via machine learning algorithms. *J. Inf. Sci.* **49**(2), 480–499. https://doi.org/10.1177/01655515211007746 (2023).
19. Adrian, F. et al. *Clickbait Detection on Online News Headlines Using Naive Bayes and LSTM*. 1–6 (2024). https://doi.org/10.1109/AIMS61812.2024.10512986.
20. Naeem, B., Khan, A., Beg, M. & Mujtaba, H. A deep learning framework for clickbait detection on social area network using natural language cues. *J. Comput. Soc. Sci.* **3** (202). https://doi.org/10.1007/s42001-020-00063-y.
21. Kaur, S., Kumar, P. & Kumaraguru, P. Detecting clickbaits using two-phase hybrid CNN-LSTM biterm model. *Expert Syst. Appl.* **151**, 113350. https://doi.org/10.1016/j.eswa.2020.113350 (2020).
22. Thakur, D. S. & Kurhade, S. Context-based Clickbait identification using deep learning. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*. 1–5. https://doi.org/10.1109/ICCICT50803.2021.9510141.
23. Rajapaksha, P., Farahbakhsh, R. & Crespi, N. BERT, XLNet or RoBERTa: The Best Transfer Learning Model to Detect Clickbaits. *IEEE Access* **9**, 154704–154716. https://doi.org/10.1109/ACCESS.2021.3128742 (2021).
24. Marreddy, M., Oota, S., Vakada, S., Chinni, V. C. & Mamidi, R. *Clickbait Detection in Telugu: Overcoming NLP Challenges in Resource-Poor Languages using Benchmarked Techniques*. 1–8 (2021). https://doi.org/10.1109/IJCNN52387.2021.9534382.
25. Jain, M., Mowar, P., Goel, R. & Vishwakarma, D. K. Clickbait in social media: Detection and analysis of the bait. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*. 1–6 (2021). https://doi.org/10.1109/CISS50987.2021.9400293.
26. Razaque, A. et al. Clickbait detection using deep recurrent neural network. *Appl. Sci.* **12**(1), 1 (2022). https://doi.org/10.3390/app12010504.
27. Fakhruzzaman, M., Jannah, S., Ardiati Ningrum, R. & Fahmiyah, I. Flagging clickbait in Indonesian online news websites using fine-tuned transformers. *Int. J. Electr. Comput. Eng. IJECE* **13**, 2921 (2023). https://doi.org/10.11591/ijece.v13i3.pp2921-2930.
28. Kumari, S., Singh, J. & Kumar, G. Identification of clickbait news articles using SBERT and correlation matrix. https://doi.org/10.21203/rs.3.rs-3294778/v1 (2023).
29. D. Broscoteanu and R. Ionescu, "A Novel Contrastive Learning Method for Clickbait Detection on RoCliCo: A Romanian Clickbait Corpus of News Articles," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 9547–9555. https://doi.org/10.18653/v1/2023.findings-emnlp.640.
30. Yadav, K. & Bansal, N. *Clickbait Detection Using Bi-LSTM Model*. 116–120. https://doi.org/10.1109/CIISCA59740.2023.00032 (2023).
31. Hashmi, E., Yildirim Yayilgan, S., Yamin, M., Ali, S. & Abomhara, M. Advancing fake news detection: Hybrid deep learning with FastText and explainable AI. *IEEE Access* **12**, 44462–44480 (2024). https://doi.org/10.1109/ACCESS.2024.3381038.
32. Abualigah, L., Al-Ajlouni, Y. Y., Daoud, M. Sh., Altalhi, M. & Migdady, H. Fake news detection using recurrent neural network based on bidirectional LSTM and GloVe. *Soc. Netw. Anal. Min.* **14**(1), 40 (2024). https://doi.org/10.1007/s13278-024-01198-w.
33. Mallik, A. & Kumar, S. Word2Vec and LSTM based deep learning technique for context-free fake news detection. *Multimed. Tools Appl.* **83**(1), 919–940. https://doi.org/10.1007/s11042-023-15364-3 (2024).
34. Alhanaya, R., Alqarawi, D., Alharbi, B. & Ibrahim, D. M. Mushakkal: Detecting Arabic Clickbait Using CNN with Various Optimizers. *J. Inf. Technol. Manag.* **16**(4), 64–78. https://doi.org/10.22059/jitm.2024.99051 (2024).
35. Ahmad, P. N., Shah, A. M., Lee, K. & Muhammad, W. Misinformation detection on online social networks using pretrained language models. *Inf. Process. Manag.* **63**(1), 104342 (2026). https://doi.org/10.1016/j.ipm.2025.104342.
36. Deng, Q., Chen, X., Lu, P., Du, Y. & Li, X. Intervening in negative emotion contagion on social networks using reinforcement learning. *IEEE Trans. Comput. Soc. Syst.*. 1–12 (2025). https://doi.org/10.1109/TCSS.2025.3555607.
37. Nie, W., Chen, R., Wang, W., Lepri, B. & Sebe, N. T2TD: Text-3D Generation Model Based on Prior Knowledge Guidance. *IEEE Trans. Pattern Anal. Mach. Intell.* **47**(1), 172–189. https://doi.org/10.1109/TPAMI.2024.3463753 (2025).
38. Pujahari, A. & Sisodia, D. S. Clickbait detection using multiple categorisation techniques. *J. Inf. Sci.* **47**(1), 118–128. https://doi.org/10.1177/0165551519871822 (2021).
39. Kumawat, D. & Jain, V. POS Tagging Approaches: A Comparison. *Int. J. Comput. Appl.* **118**, 32–38. https://doi.org/10.5120/20752-3148 (2015).
40. Shanbhag, A., Jadhav, S., Thakurdesai, A., Sinare, R. & Joshi, R. "BERT or FastText? A Comparative Analysis of Contextual as Well as Non-Contextual Embeddings. arXiv.org: https://arxiv.org/abs/2411.17661v2. Accessed 03 Dec 2024.
41. Shahmohammadi, H., Heitmeier, M., Shafaei-Bajestan, E., Lensch, H. P. A. & Baayen, R. H. Language with vision: A study on grounded word and sentence embeddings. *Behav. Res. Methods* **56**(6), 5622–5646. https://doi.org/10.3758/s13428-023-02294-z (2024).
42. Hayat, M. K. et al. Towards Deep Learning Prospects: Insights for Social Media Analytics. *IEEE Access* **7**, 36958–36979. https://doi.org/10.1109/ACCESS.2019.2905101 (2019).
43. Yin, Z. & Wang, S. Enhancing bibliographic reference parsing with contrastive learning and prompt learning. *Eng. Appl. Artif. Intell.* **133**, 108548. https://doi.org/10.1016/j.engappai.2024.108548 (2024).
44. Zheng, W. et al. PAL-BERT: An improved question answering model. *CMES - Comput. Model. Eng. Sci.* **139**(3), 2729–2745 (2024). https://doi.org/10.32604/cmes.2023.046692.
45. Liting Jing, A. & Xiaoyan, F. B. A patent text-based product conceptual design decision-making approach considering the fusion of incomplete evaluation semantic and scheme beliefs. *Appl. Soft Comput.* **157**, 111492. https://doi.org/10.1016/j.asoc.2024.111492 (2024).
46. Liu, S. et al. The scales of Justitia: A comprehensive survey on safety evaluation of LLMs. *IEEE Trans. Knowl. Data Eng.* https://doi.org/10.48550/arXiv.2506.11094 (2025).
47. Ahmadi, H. A. & Chowanda, A. Clickbait classification model on online news with semantic similarity calculation between news title and content. *Build. Inform. Technol. Sci. BITS* **4**(4), 4 (2023). https://doi.org/10.47065/bits.v4i4.3030.
48. Sirusstara, J. et al. 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)*, IPOH. 1–6 (IEEE, 2022). https://doi.org/10.1109/AiDAS56890.2022.9918678.
49. Mollah, M. A. & Khan Sami, S. Exploration of transformer ensemble and auto regressive approaches to enhance performance of Clickbait title detection. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, Cox's Bazar, Bangladesh. 1–4 (IEEE, 2023) . https://doi.org/10.1109/ICCIT60459.2023.10441032.

Author contributions

Fawaz Khaled Alarfaj, Amara Muqadas, Hikmat Ullah Khan, and Anam Naz contributed equally to the conception, design, execution, and manuscript preparation. All authors read and approved the final manuscript.

Funding

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU254280].

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.K.A. or H.U.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.