Indian Journal of Natural Sciences www.tnsroindia.org.in ©IJONS

Vol.14 / Issue 82 / Feb / 2024 *International Bimonthly (Print) – Open Access* *ISSN: 0976 – 0997*

RESEARCH ARTICLE

Fake News Detection using Machine Learning and Natural Language Processing

Maitri Patel[1],Krunal Shinde[1],Manav Shah[1] and Bijal Dalwadi[2]

[1]Student at IT Department, BVM Engineering College, V.V. Nagar, Gujarat, India.
[2]Assistant Professor at IT Department, BVM Engineering College, V.V. Nagar, Gujarat, India.

Received: 30 Dec 2023 Revised: 09 Jan 2024 Accepted: 12 Jan 2024

***Address for Correspondence**
Maitri Patel
Student at IT Department,
BVM Engineering College,
V.V. Nagar, Gujarat, India.
Email: maitri.1082002@gmail.com

ABSTRACT

In an era marked by the rapid dissemination of information through social media, the proliferation of fake news poses a critical challenge to the integrity of online information. This research addresses this issue by employing a multi-pronged approach, utilizing logistic regression, decision tree, gradient boosting, LSTM, and BERT models, to discern the veracity of news content. Leveraging a comprehensive dataset sourced from Kaggle, encompassing diverse news articles prevalent on different news papers websites, we train and evaluate these models for their efficacy in distinguishing between genuine and fabricated information. Through NLP, it extracts various linguistic features, including textual patterns, sentiment analysis to build a comprehensive understanding of the content. This research not only contributes to advancing the field of fake news detection but also underscores the necessity for a multifaceted approach in combating misinformation on popular online platforms.

Keywords: Fake news detection · natural language processing · *machine learning* · BERT

INTRODUCTION

In an era characterized by the rapid dissemination of information through digital platforms, the proliferation of fake news has emerged as a critical societal concern. Fake news, often defined as intentionally false or misleading information presented as genuine news, has the potential to misinform, deceive, and manipulate public opinion. Its impact extends across various domains, including politics, health, finance, and beyond. As the volume of online content continues to surge, distinguishing between reliable information and deceptive narratives has become an increasingly challenging endeavor. The phenomenon of fake news is not merely confined to the realm of social media or fringe websites; it has penetrated mainstream news cycles, amplifying its potential to influence public

discourse. Its prevalence has raised pressing questions about the ethics of information dissemination, the role of technology platforms, and the responsibilities of both content creators and consumers. This review paper seeks to comprehensively examine the landscape of fake news, encompassing its origins, dissemination mechanisms, psychological and societal impact, as well as the diverse array of methods and technologies developed to combat its spread. By synthesizing and critically analyzing the existing body of literature, this paper aims to provide a comprehensive overview of the multidimensional challenges posed by fake news, while also highlighting the innovative approaches that have been employed in its detection, mitigation, and prevention.

RELATED WORK

The research in fake news detection has seen intense activity in recent years, with a primary focus on analyzing the dissemination of hoaxes through social media channels. Researchers have explored the integration of convolutional neural networks and linguistically-infused neural networks, leveraging techniques like Long-Short-Term-Memory (LSTM) and incorporating pre-trained vectors. The complexity of the modelis not the sole solution; instead, the right choice of parameters and data proves essential. Despite the progress, the challenges persist due to the diverse variables associated with news statements, including sarcasm, abbreviation, metaphors, etc.One study proposed a method involving recurrent neural networks for stance detection of fake news. This approach captures temporal patterns of user activity, extracts source characteristics, and integrates them to form a classification model. It's worth noting that even simpler network models have demonstrated superior performance, highlighting the importance of parameterselection and data quality. Addressing the fake news problem requires aggressive efforts, given its alarming growth rate. The availability of reliableand extensive datasets is crucial for further progress in this area. In summary, the research landscape in fake news detection is diverse and evolving, encompassing various methodologies from classical machine learning to deep neural network approaches. There is still significant room for development, especially in tackling the intricate nature of news statements.

DATA

The dataset used in this research paper is sourced from Kaggle and is titled "Fake and Real News Dataset". It is curated by Clément Bisaillon and contains a collection of news articles, categorized into two distinct groups: genuine ("real") news articles and fabricated ("fake") news articles.

Composition The dataset is structured with two main components:

Real News This category comprises news articles from reputable and established news sources, recognized for their credibility and journalistic integrity.

Fake News This category encompasses news articles that have been intentionally created to mislead or deceive readers, often originating from less reputable or unverified sources.

Size The dataset includes a substantial number of articles, providing a diverse and comprehensive corpus for analysis. The specific number of articles in each category can be found in the dataset description on Kaggle.

Attributes Each news article in the dataset is typically represented by several key attributes, including but not **limited to:Title** The headline or title of the news article.

Text The body of the news article containing the main content.

Subject The general category or topic to which the news article pertains (e.g., politics, world news, etc.).

DATA PREPROCESSING TECHNIQUES

Involves cleaning, transforming, and organizing the raw data to make it suitable for training and evaluating the machinelearning algorithms. The main tasks involved in data preprocessing for fake news detection include:

Text Cleaning

Text cleaning involves removing any extraneous characters, symbols, or elements from the text that do not contribute tothe meaning. This may include HTML tags, special characters, or any other noise in the data.

Example Removing HTML tags like <p> or &.

Lowercasing
Lowercasing entails converting all the text to lowercase. This is important as it standardizes the text and ensures thatwords are treated consistently regardless of their capitalization.
Example Converting "Hello World" to "hello world".

Tokenization
Tokenization involves splitting the text into smaller units, usually words or phrases (tokens). These tokens serve as thebuilding blocks for further analysis.
Example Splitting the sentence "Natural Language Processing is amazing!" into tokens: ["Natural", "Language", "Processing","is", "amazing", "!"].

Stop-word Removal
Stop words are common words (e.g., "the", "and", "is") that occur frequently in a language and do not carry significantmeaning. Removing them helps reduce noise and focus on more meaningful content.
Example Removing words like "the", "is", and "and" from a sentence.

Lemmatization / Stemming
Both lemmatization and stemming are techniques used to reduce words to their base or root form. This helps in reducingthe dimensionality of the data and capturing the core meaning of a word.
Example
Stemming Reducing words like "running", "ran", and "runs" to their common root "run".
Lemmatization Reducing words like "better", "best", and "good" to their base form "good".

FEATURE EXTRACTION
Involves converting raw text data into a set of numerical features that machine learning algorithms can process. Effective feature extraction is essential for capturing the relevant information from the text and representing it in a way that facilitates the detection of fake news.

Word Embedding
Word embedding is a technique that represents words as vectors in a continuous vector space. Each word is mapped to a high-dimensional vector where semantically similar words are located closer to each other in the space. This captures semantic relationships between words and is widely used in tasks like sentiment analysis, language translation, and more.
Example: In a word embedding space, words like "king" and "queen" might be closer to each other because of their semantic similarity.

TF-IDF Vectorization
TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure used to evaluate the importance of a word within a collection of documents. It calculates a weight for each word based on how frequently it occurs in a document relative to its frequency in the entire corpus. This helps in identifying words that are distinctive to a specific document.
Example: In a collection of articles about cats, the term "cat" would have a high TF-IDF score because it's likely to appear frequently in each document.

Count Vectorization
Count vectorization, also known as Bag-of-Words (BoW), is a simple technique that converts text data into numerical vectors. It creates a vocabulary of unique words in the corpus and counts the frequency of each word in a given document. Each document is then represented as a vector with the count of each word from the vocabulary

Example Consider the sentences "I love cats" and "I love dogs". In count vectorization, the vectors for these sentences might be [1, 1, 0, 0] and [1, 0, 1, 0] respectively, where the positions correspond to the words ["I", "love", "cats", "dogs"]. These feature extraction techniques are fundamental in converting textual data into a format that machine learning models can understand. They capture different aspects of the text, whether it's semantic relationships (word embedding), importance within a document (TF-IDF), or basic word frequency (count vectorization). The choice of technique often depends on the specific taskand the nature of the text data being analyzed.

MODEL DESCRIPTION

LSTM

This architecture is based on LSTM cells which are a type of recurrent neurons that have probed to give very interestingresults in problems related to sequence modeling as they have the capability to "remember" information from the past. The LSTM units are composed of several gates in charge of maintaining a hidden cell state which allows them tomitigate the vanishing gradient problem and, therefore, gives them the ability to remember more distant information in the past than vanilla recurrent units. This feature it's important in the context of NLP since the words from the past often influence the current ones. More exactly, the architecture uses bidirectional LSTM layers, in which the sequence (ie. the text) is fed forwards and backwards. This decision is based on the intuition that in language, future words modify the meaning of the ones in the present. For example, polysemous words such as bank, mouse or book show that its context is needed in order to model their meaning. Later in the network, these representations are merged and classified in one of the two possible categories(true or fake).

BERT

In recent years, a huge number of improvements have been made in the field of NLP thanks to deep learning. Most of therecent ones are based on a special type or architecture known as "transformer".

Transformers

It's main goal is, given an input sequence, to transform it into another. The architecture uses "attention mechanisms", which are responsible of determining the most relevant parts of the input sequence. This way, better language representations are created because longer relationships in the sequence can be captured, usually further longer than with LSTM neurons despite of being more computationally efficient as the operations applied to the input the are simpler. A transformer is based on the idea of having two pieces: an encoder and a decoder. The encoder creates a representation of a given input in a different dimensional space and then, the decoder takes that representation and generates other sequence. This strategy is called "encoder-decoder" and is widely used in tasks like text summarization or machine translation. A diagram of the transformer architecture is shown in the figure 2, where the left part corresponds to theencoder block and the right one to the decoder. Each transformer block uses a "self-attention" system which is in charged of choosing the most relevant parts ofthe input sequence. This system works by operating three matrices: Q, K and V (Query, Key and Value, respectively), which represent an abstraction to calculate the attention matrix, Z (equation 1). These three matrices are learnt through in the training phase of the network After obtaining those matrices the Z matrix can be calculated as shown in equation 1, where dk is the chosen dimension of each key vector (i.e the number of columns in K). Inthe original work, this value corresponds to dk = 64.

$$Z = softmax(\frac{QK^T}{\sqrt{d_k}}V) \quad [2]$$

Also, in the figure 2, several "Multi-Head Attention" blocks can be seen. These simply repeat the attention operationexplained above n times, obtaining n attention matrices, which are concatenated and multiplied by other matrix, WO, in order to obtain an output that is fed to the normalization block ("Add & Norm", en the figure). Besides theMulti-Headed system, the researchers also proposed the use of skip-connections, in such a way an identity signalcould be transmited to deeper layers, improving the learning process.

BERT Based

BERT is a language model created by researchers at Google which is based on transformers. Roughly, it iscomposed of several stacked transformer encoder blocks. The strategy to follow with BERT falls under transfer learning. BERT is provided already pretrained on a large text corpora (books, Wikipedia, etc.) with the aim that the final user performs a fine-tunning phase to adapt the model to his specific problem. Google provides several pretrained models. In their work they present variants of the architecture: "BASE" and "LARGE" which differ in their size since the first one uses 12 blocks and the second 24 blocks. Due to computational power constraints, in the current work the "BASE" version has been used. This is also the approach followed in the original publication. The adaptation included in this work consists on adding an extra layer to the model provided by Google.This layer is a fully connected layer with sigmoid as activation function plus a softmax function on top to allow the interpretation of the result as a probability. For simplicity in the implementation and the possible future in which the model is required to classify articles ina broader set of categories, the number of neurons in this last layer can be changed as a function of the number of classes in the classification problem. For this binary classification problem (true/false) the number of output neurons is two. BERT input data format is different from the ones used for the other two architectures since it is based only on text strings. The word tokenization and separation processes are already included in the input data function for this model. The word tokenization follows a strategy called WordPiece. This considers the words as combinations of some more basic tokens joined together. For example, doing would be formed by joining do anding. By separating the tokens like that, the available lexicon is largely increased, minimizing the potential number of OOV errors (Out Of Vocabulary). As BERT admits only one input vector, the title and the article body were concatenated before feeding in to the model.

Supervised Machine learning algorithms: Decision tree, Random Forest, Gradient Boosting, Logistic Regression

Decision Tree

A Decision Tree is a flowchart-like structure where each internal node represents a feature (or attribute), each branch represents a decision rule, and each leaf node represents an outcome. It's a popular algorithm for classification tasks. The tree is constructed by recursively partitioning the data based on the values of features, aiming to minimize impurity or maximize information gain at each step

Strengths

1. Easy to interpret and visualize.
2. Can handle both numerical and categorical data.
3. Requires relatively little data preprocessing. Weaknesses:
4. Tends to overfit with complex trees, which may lead to poor generalization.
5. Sensitive to small changes in the data.
6. Limited ability to capture complex relationships.

Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance and reduce overfitting. It builds multiple decision trees using bootstrapped samples of the data and random subsets of features for each tree. The final prediction is determined by aggregating the outputs of all individual trees.

Strengths

1. Reduces overfitting and improves generalization.
2. Handles large datasets with high dimensionality well.
3. Provides feature importance rankings. Weaknesses:
4. May be computationally expensive, especially with large number of trees.
5. Can be challenging to interpret compared to a single decision tree.

Gradient Boosting

Gradient Boosting is an ensemble learning technique that builds a series of weak learners (usually decision trees) sequentially, with each one correcting the errors of the previous one. It minimizes a loss function, typically using

gradient descent, to iteratively improve predictions. Common implementations include XGBoost, LightGBM, and AdaBoost.

Strengths

1. Often provides state-of-the-art performance on a wide range of problems.
2. Handles mixed data types and missing values naturally.
3. Robust to outliers and noisy data.

Weaknesses

1. Can be sensitive to hyperparameters and require tuning.
2. May be prone to overfitting, particularly with deep trees.

Logistic Regression

Despite its name, Logistic Regression is a classification algorithm used for binary and multi-class classification problems. It models the probability of a sample belonging to a particular class using the logistic function. It estimates coefficients for each feature to make predictions.

Strengths

1. Simple and computationally efficient.
2. Provides probabilities for classification.
3. Easy to interpret and explain. Weaknesses:
4. Assumes a linear relationship between features and the log-odds of the response variable.
5. May not perform well with highly non-linear relationships.

COMPARISON USING ACCURACY

PROBLEMS

1. While using LSTM we get low accuracy and over fitting problem
2. In other supervised learning model we face problem related to features

So, we use BERT model for our system because it gives good accuracy and work fine

GUI

RESULT

Using BERT

CONCLUSION

In conclusion, the pervasive influence of fake news in today's information landscape demands a concerted and multidisciplinary response. This review has delved into the multifaceted nature of fake news, exploring its origins, dissemination mechanisms, and far-reaching societal implications. The prevalence of deceptive narratives, often masquerading as legitimate news, poses a significant challenge to the integrity of information consumption and public discourse. As discussed, a wide array of approaches has been developed to combat the spread of fake news, ranging from traditional fact-checking to cutting-edge machine learning algorithms. These efforts underscore the urgency with which stakeholders across academia, industry, and government are addressing this pressing issue. While significant strides have been made, it is evident that the battle against fake news is an ongoing one, requiring continuous adaptation and innovation. The development of robust datasets, the refinement of detection techniques, and the cultivation of media literacy are integral components of a comprehensive strategy. Moreover, collaboration between technology platforms, media organizations, and researchers is crucial in fortifying the defenses against misinformation. Ethical considerations, such as preserving free speech while curbing the spread of false information, must remain at the forefront of these efforts. In conclusion, the fight against fake news is a collective endeavor that demands the concerted efforts of researchers, policymakers, and the wider public. By fostering a culture of critical thinking, leveraging advanced technologies, and upholding the principles of journalistic integrity, we can work towards a more informed and resilient society.

REFERENCE

1. FAKE NEWS DETECTION USING DEEP LEARNINGBy:- Álvaro Ibrain Rodríguez, Lara Lloret Iglesias
2. Dataset – https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset
3. Analysis of Classifiers for Fake News Detection By:-Vasu Agarwala, H. Parveen Sultanaa , Srijan Malhotraa, AmitrajitSarkar
4. Code- https://www.kaggle.com/code/sadikaljarif/fake-news-detection-using-bert

Table 1: Accuracy Comparison

Sr. No.	Classifiers	Accuracy
1	LSTM	55%
2	BERT	80%
3	Logistic regression	88%
4	Decision tree	78%
5	Gradient boosting	77%
6	Random forest	80%

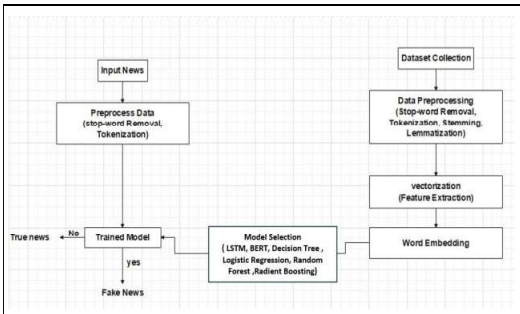


Figure 1:Flow of the System

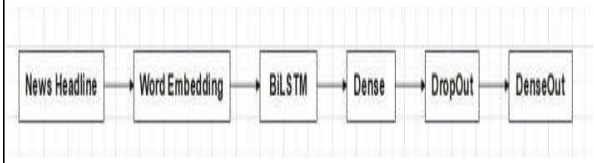


Figure 2:LSTMArchitecture

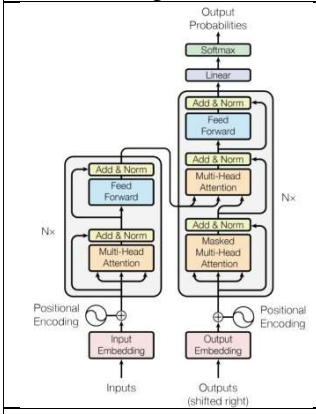


Figure 3: Transformer block diagram

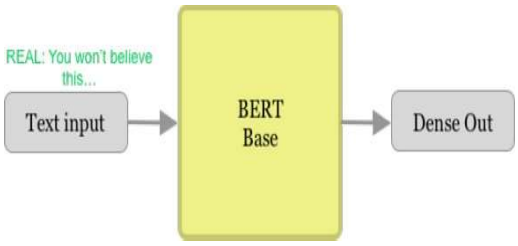


Figure 4: BERT based architecture diagram

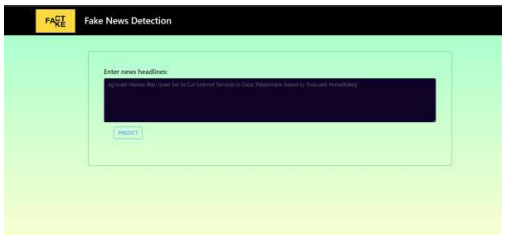


Figure 5: GUI

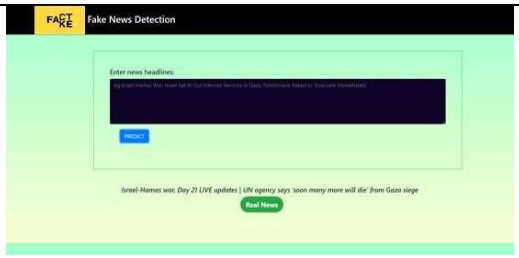
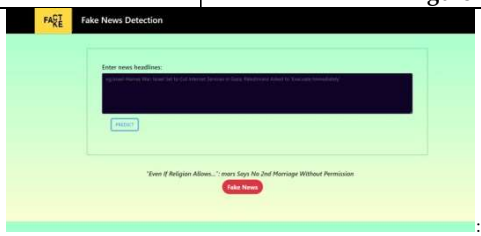


Figure 6 : Real News



:

Figure 7 : Fake News