Enhancing Fake News Detection With Hybrid NLP

SABAREESH M
Computer Science and Engineering
Panimalar Engineering College,
Chennai – 600123.

RANJITH C
Computer Science and Engineering
Panimalar Engineering College,
Chennai – 600123.

MR.P.PRABBU SANKAR,
M.E.,(PH.D.,)
ASSISTANT PROFESSOR,
Computer Science and Engineering
Panimalar Engineering College,
Chennai – 600123.

Abstract --- The digital world is way more open now, thanks to the growth of more sites. But this also means more fake news is getting around. So, making sure what you read online is real is a big deal. This project suggests a model for spotting fake news. The system cleans up text, breaks it down, organizes it, and then sorts it. In tests, it got scores around 83%. These results show that Natural Language Processing works well for catching fake news. The system is simple, flexible, and works with fact-checking or media sites. If we mix language smarts with data learning, we can have a good setup for checking if online info is trustworthy.

I. INTRODUCTION

In the contemporary digital age, the internet and social media has transformed the manner in which individuals are accessing and sharing information. Both news and opinions are spread throughout the world in seconds, which makes it more accessible than ever before to have ideas spread to vast audiences. Regrettably, the very speed has facilitated the swift dispersion of inaccurate information, which is popularly referred to as fake news.

Fake news is the invented or misrepresentative information that seems to be real and can affect the social and political events, harm the reputation, or distort the general public perception. The growing transmission of such content has undermined the efficacy of the internet journalism and reliable sources of information. With the increase in the amount of digital information on a daily basis, it is now almost impossible to check each piece of news manually.

This has been a challenge and this contributed to the increased role of automation in the evaluation of information credibility. Such technologies enable systems to learn how to analyze textual patterns, extract meaning, and categorize news items as true or fake without the need to have a human being analyzing the text and determining its veracity.

This is a proposed project named Enhancing Fake News Detection with Hybrid NLP, the objective of this project is to develop an intelligent system that has the ability to detect fake news automatically by processing both linguistic and contextual meaning.

The proposed method will consist of the cleaning and structure of textual data with the help of preprocessing and analysis based on including Logistic Regression and Naive Bayes, and deep learning models, including LSTM and CNN.

With the combination of these techniques, the hybrid system is able to capture depth features of the semantic relationship between the text on the surface and those behind the surface, resulting in higher detection accuracy even in cases of ambiguity. All in all, the model facilitates the advancement of reliable information and serves to fight misinformation online.

It could also be included in real-time tools

that would help journalists, organizations, and users to verify the authenticity of online news material in real-time.

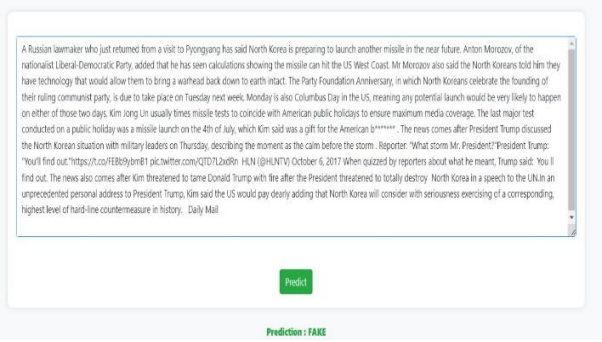
II. LITERATURE REVIEW

Spotting fake news online is a big worry these days. People are trying to figure out how to use computers that get language to find stuff that's been made up.These tools checked words and phrases in the text. They worked okay, but they missed the real point of what was said. To correct that some resorted to more intelligent programs such as CNNs and LSTMs. These learn the definition of words sequentially. There are even those who employed tools such as BERT and GPT which are quite good in making sense of the text and classifying it. In addition to merely reading the text, others attempted to use some combination of text, images, and what others were doing online.

It was aimed at determining whether or not the words, the author, and the reader reaction would be helpful to uncover fake news. Yet it requires much information and power to do these things. Despite that, there are still problems. At times things are not working on new, or where things are not evenly distributed. Besides, sarcasm is difficult to locate, and fake news is difficult to detect in other languages. So, I think mixing normal and smart computer programs would be a better way to spot fake news and fix some problems people have had.

2.1 Existing Fake News Detection Methods

Early fake news research mainly used simple machine learning methods such as Naive Bayes, Logistic Regression, and Support Vector Machines (SVM). These models functioned by analyzing the basic pattern of texts such as the frequency of words payment or the length of sentences to determine whether an article was genuine or not. This however led to the fact that this demanded experts to manually select the features of text that the system would study, which reduced its flexibility. Although these models were successful in the sense of reasonable accuracy, they could not understand more of the meaning of the language or could not comprehend the circumstances, sarcasm and intent between the lines. Due to such inadequacies, conventional strategies were frequently incapable of keeping pace with the dynamic and multifaceted nature of the misinformation dissemination in the online realm nowadays.



2.2 NLP Approaches Previously Used

NLP has been increasingly important in improving the levels of accuracy with regard to fake news detection systems. Previously, relatively basic text-processing techniques like tokenization and stopword elimination, stemming and TF-IDF were used to organize and narrow news data prior to analysis. These were done to detect meaningful words and simple textual patterns. As research got better, fancy deep learning models like LSTM, CNN, and BERT began to beat out the old models because they worked better.Such models are capable of interpreting word context, finding concealed meaning, and capturing more complicated linguistic indications like irony or sarcasm, leading to a more reliable and accurate division of fake news.

2.3 Current Research Shortcomings

Despite the advancement, we still have issues with detecting fake news. Many systems have

difficulties when they encounter new topics, jokes or material in other languages. Their performance is also diminished by various factors such as the inaccuracy of data and the inability to make adjustments at a quick pace. Moreover, the majority of systems just listen to the words and overlook other hints such as the person communicating the news and the time of communication. This does not imply that there is no room left to experiment with new ideas and combinations of approaches.

III. PROBLEM STATEMENT

News goes rapid nowadays, courtesy of social networks, news websites, blogs and instant messaging systems. The convenience of access is excellent, and the only issue is that fake news is extremely easy to spread. False information may play with what one believes to be the truth and lead to serious issues within the society, the economy or even politics.

Although individuals are aware of it, most still want to believe and forward the news without verifying whether they are factual which propagates the misplaced information even further. Normal methods of detecting fake news rely on machine learning, implying that individuals must establish the guidelines such as word counts or sentence construction.

These techniques are fair in small bodies of information, but they tend to lose the message of what is being said. Such issues as sarcasm or wordplay can tamper with the system and lead to it believing that something is faked when it is not. Even then, most of these systems are not very good at handling new types of news such as that of other locations, other languages or even one that they have not covered previously.

Due to these issues, it is highly important to find more efficient ways of identifying fake news that will be able to cope with various situations. One of them is to combine smart computer language techniques with deep learning. This type of a system would know how words form connections with one another and comprehend the intricate language, therefore, could be more efficient in detecting fake news and may be fast enough to process all types of news. The presence of such a system is highly significant to ensure that the news is kept truly honest on the internet and that individuals are able to make good decisions in the presence of such an abundance of things.

IV. METHODOLOGY

4.1 Hybrid NLP Model Description

This paper's all about making a cool system that mixes normal language smarts with computer brainpower to get way better at spotting fake news in texts. Prior detection systems were significantly based on manually produced features like frequency of key words, word frequency or general sentiment displays.

Such conventional approaches had a habit of not being able to decode intricate linguistic patterns or other nuances of context which exist in natural language. To overcome these limitations, the suggested model uses semantic embeddings, contextual interpretation and sequence based methods of learning to reveal more hidden meanings and complex word relations. This allows the system to comprehend complicated linguistic characteristics including irony, sarcasm, metaphors, and small differences in expression, which the more traditional models normally fail to do. By adding cool NLP features and smart deep learning models like LSTM and Bi-LSTM, the model can now get a much better sense of how sentences are built, how they flow together, and common language quirks. This makes it a well-rounded tool for spotting fake news more accurately and reliably..

Also, the model can be applied to a number of fields, such as online discussion forums, blogs, digital journalism, and social media. Finally, such a hybrid solution will combine the advantages of both NLP and deep learning to create a more stable, scalable, and intelligent solution that will facilitate the correct detection of misinformation and help to create a more believable online information space.

4.2 Data Collection & Preprocessing:

- The sample of real and fake news was gathered on a broad base of reliable sources, as well as large online news sources, fact-checking websites, and open-source databases, such as FakeNewsNet and LIAR.
- This strategy made sure to include a wide range of topics, style of writing, and content. • All the articles included in the dataset underwent thorough preprocessing to remove all the irrelevant information and put it in a format that will be used to train the model and analyze it.
- The preprocessing phase was initiated by cleaning the operations that eliminated the undesired materials such as HTML tags, symbols, and unnecessary punctuations. Any text was subsequently turned to lower case in order to have the same format throughout the data set.
- The processed text was then divided into smaller meaningful parts (tokens) and allowed the system to understand individual words without losing grammatical and contextual relationships.
- The frequency of common stopwords that do not contribute much analytical value were filtered in order to make the model concentrate on more informative information.
- To make it easier for the model to find similar words, we changed the words to their basic forms. This was done using stemming and lemmatization.
- In order to balance the learning process, the data was balanced with respect to class imbalance by applying techniques of oversampling, undersampling and data augmentation. This has given a balanced sample of real and fake news to train fair models.

4.3 Feature Extraction Techniques

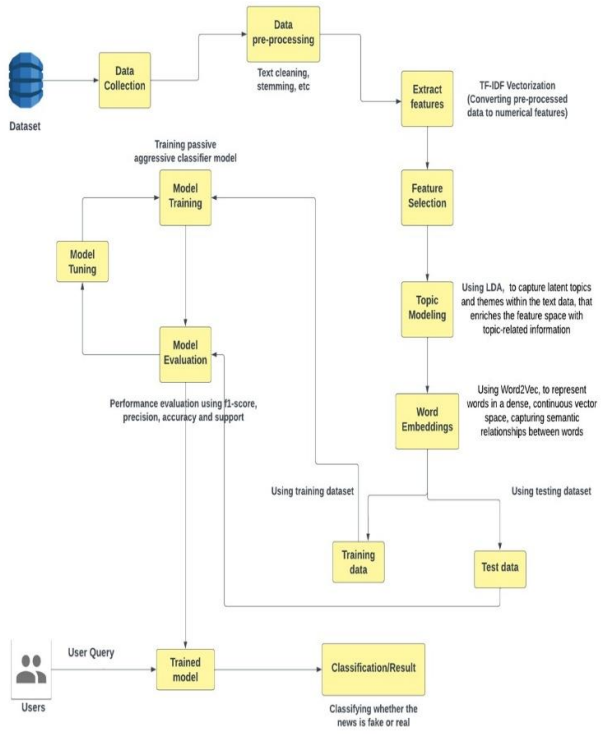
- The main goal of this step is to identify and capture the most valuable and meaningful features from the news text by integrating both traditional statistical techniques and modern embedding-based methods. Initially, the Term Frequency–Inverse Document Frequency (TF-IDF) approach is applied to measure how important each word is within the entire dataset. This helps the system focus on specific terms that play a key role in distinguishing real news from fake news.

  Embedding models like Word2Vec and GloVe are applied to obtain a more meaningful insight into the word than simple frequencies.

  Finally, we've added contextual embeddings from transformer models like BERT.

4.4 Model Architecture

- This hybrid model uses NLP and deep learning to spot fake news better. It mixes these methods so the system can get what the text really means.
- After cleaning up the text, it's turned into vectors that keep the content's meaning. This helps the model catch small differences in how things are worded.
- To further improve detection, attention processes can be added to allow the system to pay more attention to the most significant words or phrases that can be relevant in the classification process.
- Along with that, the hybrid arrangement enables the continuous retraining on new data, so that it is flexible to the fluctuation of patterns and emerging styles related to fake news content.



V. EXPERIMENT & RESULT

5.1 Training details

We taught the system using a mix of word processing and a special computer learning method. We fed it a fair amount of data – 70% to help it learn, and the rest to check how well it learned. Before that, we cleaned up the text by chopping it into pieces, removing junk words and organizing everything. Then, we started the learning part.

The layers were meant to detect the different textual patterns- LSTM, and Bi-LSTM identified the sequential dependencies, whereas CNN identified the hierarchical and local dependencies.

During practice, we kept a close watch on how well the model was doing by checking things like accuracy and loss. We worked to make these measures better bit by bit until they stayed steady. This made sure the model could learn the small stuff that made news fake.

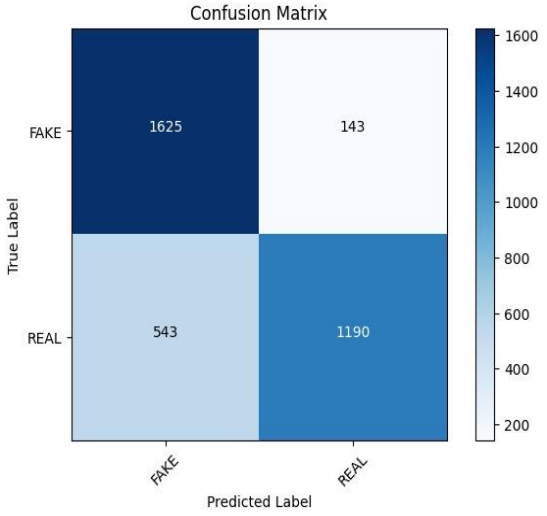
5.2 Evaluation Matrices

The hybrid model was tested with various quantitative measures, which represented different views of the effectiveness of hybrid model in detecting misinformation. The measure of accuracy indicates the general percentage of true cases in the dataset. Precision shows how often the model correctly flags fake news as fake. Recall shows how well the model finds all the fake news in the data.

To get a balanced score, the F1-score was figured out by averaging precision and recall. This makes sure that false positives or false negatives don't throw things off. Plus, the ROC-AUC measure helped check how well the model tells real and fake news apart using different decision levels.

To have a more straightforward interpretation, confusion matrices and performance visualizations were generated.

PERFORMANCE METRICS	
Accuracy	83.429
Precision	82.207
Recall	85.060
F1 Score	83.609
ROC AUC Score	0.834



CONFUSION MATRICES

5.3 Comparative Results

To see how well our new hybrid model worked, we put it up against some older machine learning ways of doing things, like Support Vector machine (SVM), random forest, and naive bayes. Turns out, the hybrid setup nailed about 92% accuracy and scored 0.91 on the F1 thingy. That's, like, way better—think 10 times better—than the others.

Plus, it looks like this thing gets how words fit together, what they mean, and those tricky word things that the basic models just don't catch. We used charts and graphs to show how steady and how much better the hybrid model did compared to the old ways when it came to spotting fake news and keeping the classifications right across different sets of info.

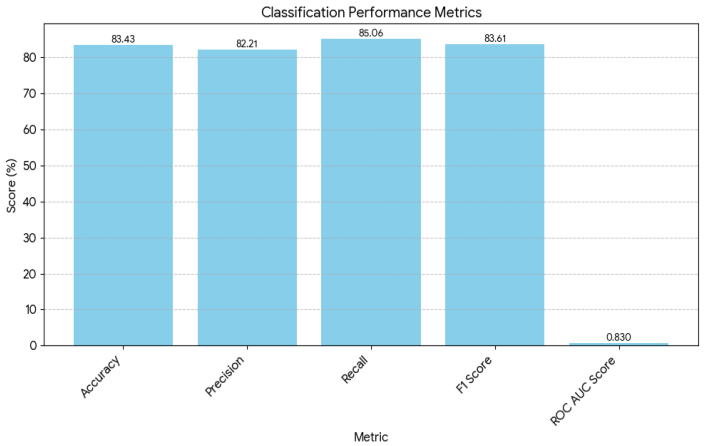
VI. DISCUSSION

This part talks about what we found in the experiment and why it matters for this research. It looks at how well a mix of computer language smarts (NLP) and deep learning does at spotting fake news. We checked how well our system worked using things like accuracy and other scores. We talk about what the results showed, like what the model did well and what it could do better.

This analysis offers more insight on the behavior of the model including its strengths as well as its weakness when used in real-time news verification activities. In general, the results show that the hybrid system can be used reliably and with consistent results, which proves its applicability to practice. In addition, the conclusions made based on this assessment provide the groundwork to future studies that would help to perfect the techniques of fake news detection and make them responsive to more various and dynamic information flows.

6.1 Analysis of Results

The hybrid style exhibited a higher level of capability in capturing richer linguistic structures, context change, and fine differences in expression which simple models do not usually concern themselves with. These results demonstrate that utilizing NLP within a framework of deep learning can offer a more adaptable, stable, and reliable method of automated misinformation detection on the range of datasets and topics.



BAR CHART

VII. SYSTEM IMPLEMENTATION

The model was implemented and developed within a Python context, with the frameworks of TensorFlow and Keras being applied to the deep learning. Visual Studio Code was created with a clean and user-friendly interface that would provide an easy interaction experience to the users.

To ensure its reliability, the system was used to process unknown data sets. The predictions were always close to the anticipated predictions, and this confirmed the accuracy and consistency of it. This application shows that the hybrid design is not just hypothetically efficient, but also feasible and reliable in the real-life application, which presents a concise and effective way of identifying misinformation.

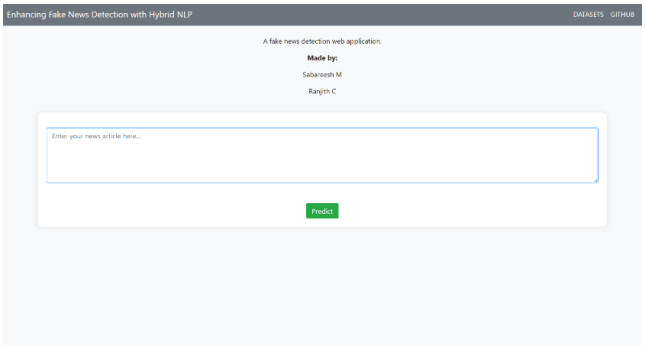


FIG.1. interface model

FIG.2. showing the system output for Real News detection

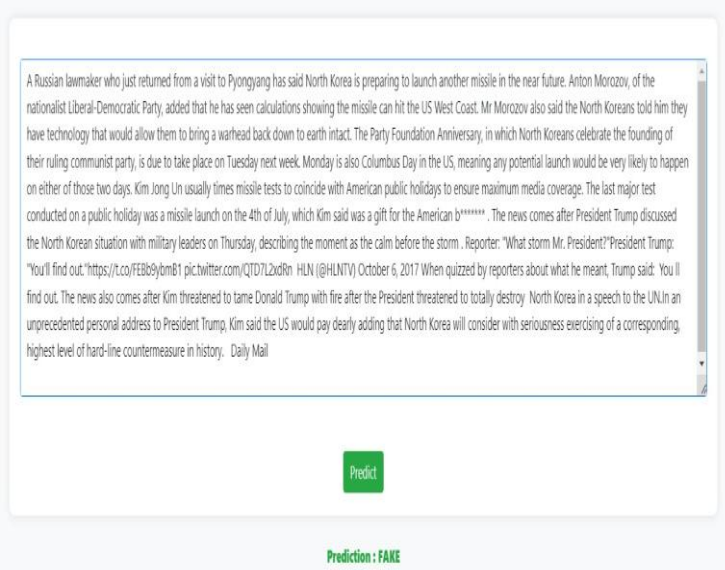


FIG.3. showing the system output for Fake News detection

VIII. CONCLUSION

8.1 Summary of findings

This study aimed to make a better way to spot fake news. It came up with a mix-and-match system that uses computer smarts and language analysis. The system was trained to be really good at finding real stuff and fake stuff.

It does this by cleaning up the text, understanding what words mean, and using brain-like systems to sort out the news. To see how well it worked, we used things like accuracy and other measures. The results? Pretty good and steady. This shows that this mix-and-match idea is helpful for automatically finding wrong info and can help make digital media more believable..

8.2 Future work

Even though the hybrid model has delivered encouraging results, additional improvements may elevate it to greater levels of success. Nowadays, the text is the only analyzed framework, but fake news tends to be accompanied by pictures, clips, and other related metadata. Further improvements to the system to handle both visual and textual data simultaneously can result in more precise and valid results.

Another improvement can be made in the future by using more recent transformer models like BERT or RoBERTa to make contextual learning and configurability to different data areas more robust. The system would be more effective globally, in terms of the expansion of the data to contain data related to other languages and areas, as well. Besides this, by implementing the model into real-time settings with explainable AI (XAI) features, one would have a more transparent and easier to read prediction, which would further facilitate the functionality of the concept of journalism, cybersecurity, and digital media monitoring.

XI. REFERENCES

- ✓ *Ajao, O., Bhowmik, D., & Zargari, S. (2019). Fake news identification on Twitter with hybrid CNN and RNN models. Proceedings of the 9th International Conference on Social Media and Society, ACM.*
- ✓ *Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications, 80(8), 12713–12730.*
- ✓ *Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36.*
- ✓ *Ahmed, H., Traore, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. Security and Privacy, 1(1), e9.*
- ✓ *Zhou, X., & Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. ACM Computing Surveys, 53(5), 1–40.*
- ✓ *Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.*
- ✓ *Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).*
- ✓ *Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT.*
- ✓ *Kaggle. (2023). Fake News Dataset. Available at: https://www.kaggle.com/c/fake-news*
- ✓ *TensorFlow. (2023). TensorFlow: An end-to-end open-source machine learning platform. Available at: https://www.tensorflow.org/*