Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 46 No. 2 (2025)

Fake News Identification and Classification Using DSSM

Mr. R. Prabhu, Mr. Bharath Arul KG, Mr. Nikhil T, Mr. Ajay A

Assistant Professsor at Dr. Mahalingam College of Engineering and Technology, Pollachi, Student at Dr. Mahalingam College of Engineering and Technology, Pollachi, Student at Dr. Mahalingam College of Engineering and Technology, Pollachi, Student at Dr. Mahalingam College of Engineering and Technology, Pollachi

Abstract:- Internet social networks are increasingly being used to spread information, including false information, for a variety of political and economic purposes. By applying technologies like artificial intelligence (AI) and natural language processing (NLP), scientists can create systems that can automatically identify fake news. The widespread use of social media has significant effects on business, culture, and society—both positive and negative. However, in order to identify false news, models must generalize the data and compare it with real news, which makes it difficult to find. It is suggested that enhanced recurrent neural networks and a deep structured semantic model can be used to identify a technique for identifying and classifying bogus news messages. With 99% accuracy, the suggested method—which does not require topic knowledge—naturally identifies distinctive characteristics linked to bogus news. The suggested system's performance measuring approach is predicated on sensitivity, accuracy, and specificity.

Keywords: *Artificial Intelligence, Natural Language Processing (NLP), Recurrent Neural Networks (RNNs), Deep Structured Semantic Model (DSSM), Fake News Detection, Machine Learning.*

1. Introduction

In the internet age, how people consume and receive news has dramatically altered, thanks primarily to the universal access of the internet and social media. Although these technologies have facilitated it as never before to remain abreast of events, they have also fostered a situation for the mass dissemination of false information. Social media websites such as Twitter, with their enormous user base and real-time sharing of content, have emerged as powerful news sources for most. Yet, the ability of users to easily post unverified content has given rise to fake news—misinformation or completely false information intended to deceive or manipulate public opinion.

Such increased concern has alarming implications for democratic processes, social trust, as well as public safety, leading to a keen need for solid means of detection of fake news.Many academics think that deep learning models and data mining algorithms could help solve the problem of fake news. Since hardware is becoming more affordable and larger datasets are more readily accessible, data mining algorithms have recently started to handle a wide range of classification problems (speech recognition, image identification, etc.) much better. Semantic features would be extremely beneficial in identifying misinformation, but such baseline classifiers' biggest shortcoming is that they fail to detect textual semantics.

A Deep Neural Network, which is composed of several hidden layers between its input and output layers, can mimic complex nonlinear relationships and convey the meaning of text. To improve upon conventional techniques for detecting fake news, Deep Neural Networks are trained using novel techniques. Our method, which is based on deep learning models like DSSM and enhanced RNN, offers a simple way to identify fake news.

2. Literature Survey

[1] Tschiatschek et al. (2018) proposed a crowd-signal-based method for fake news detection that relies on user interactions like likes, shares, and comments to determine credibility. Although effective in

some cases, the system's performance was hampered by susceptibility to manipulation by bots and fake user accounts, thus reducing its reliability in dynamic environments.

[2] Granik and Mesyura (2017) introduced a Naïve Bayes classifier to detect fake news based on word frequencies. While lightweight and easy to implement, this model lacks semantic understanding, making it ineffective against well-written deceptive content.

[3] Buntain and Golbeck (2017) explored deep learning methods, combining Recurrent Neural Networks (RNNs) with contextual embeddings to capture semantic and syntactic relationships in fake news. Their work showed significant improvements over traditional models, especially in handling contextual nuances in news articles.

[4] Liu and Wu (2018) proposed a hybrid method using DSSM and propagation path classification, which leverages both semantic embeddings and network behavior to detect fake news early in its dissemination. This combination effectively enhanced the system's semantic interpretation and contextual accuracy.

[5] Zhou and Zaarani (2018) emphasized a multi-dimensional approach for fake news detection, integrating credibility, content style, domain knowledge, and transmission patterns. Their model improved detection rates but introduced additional complexity in feature engineering and processing.

[6] Dimpas, Po, and Sabellano (2017) applied LSTM-based RNNs for clickbait detection in English and Filipino. By solving the vanishing gradient problem common in standard RNNs, LSTM networks provided improved performance by remembering long-term dependencies within text.

[7] Wang (2017) introduced the LIAR dataset, a benchmark composed of over 12,000 labeled short statements from political fact-checking sources. Alongside datasets like FakeNewsNet, LIAR remains a valuable resource for training and evaluating fake news classifiers.

[8] Wu and Liu (2018) proposed a propagation modeling technique to analyze the spread patterns of fake news across social media platforms. Their method enhances tracking and classification of fake content by examining its dissemination footprint.

[9] Ruchansky et al. (2017) developed CSI, a hybrid model integrating content, social behavior, and temporal patterns for fake news detection. This multifaceted model outperformed traditional content-only classifiers by learning user credibility and posting history.

[10] Shu et al. (2017) presented a comprehensive review of fake news detection models from a data mining perspective, highlighting challenges such as limited labeled data, real-time detection, and evolving misinformation strategies. They also proposed frameworks that integrate user and post-level features.

[11] Castillo et al. (2011) studied information credibility on Twitter, identifying that source reputation and message characteristics play key roles in fake news propagation. Their early work set the foundation for credibility-based models that combine social and linguistic features.

[12] Parikh and Atrey (2018) proposed a media-rich fake news detection approach, emphasizing the role of images and videos in spreading disinformation. Their method integrates multimodal analysis to better understand and classify fake news content in visually driven social media platforms.

3. Objectives

The objective of this project is to develop an advanced fake news detection system that leverages deep learning techniques—specifically a hybrid model combining Deep Structured Semantic Models (DSSM) and enhanced Recurrent Neural Networks (RNN). This system is designed to overcome the limitations of traditional machine learning methods, which often fail to capture the semantic and contextual meaning within text. By integrating DSSM for semantic feature extraction and RNN for sequential learning, the model aims to accurately identify and classify fake news content circulating on social media platforms like Twitter.

This project also seeks to improve the reliability and efficiency of fake news detection by using real-world datasets such as LIAR and FakeNewsNet. The proposed system will be evaluated based on performance metrics including accuracy, sensitivity, and specificity. Ultimately, the goal is to support the development of trustworthy digital environments by providing a robust, automated solution for detecting misinformation, thereby contributing to the fields of artificial intelligence, natural language processing, and societal well-being.

4. Existing System

Because of its fundamental qualities, identifying fake news is a challenging task. News identification strategies use a variety of information types, such as social and news-related data, to counteract this. This multifaceted strategy contributes to increased fake news detection accuracy. The four main components of contemporary false news detection systems are credibility, style, knowledge, and transmission. Systems can more accurately detect bogus news by examining these viewpoints. This all-encompassing method has been successful in identifying false information, as researchers Zhou and Zaarani (2018) have shown.In 2017, researchers, organizations, and entities from around the world teamed up to fight fake news. Among the methods they used were algorithmic information verification and human intervention. These efforts opened the door for the creation of automated techniques that evaluated trustworthiness on social media sites such as Twitter, utilizing human evaluations to increase precision. Researchers have suggested systems that use classification techniques like Naive Bayes, clustering, and decision trees to identify bogus news. These straightforward artificial algorithms have demonstrated promise in spotting false information. Deep Learning techniques, on the other hand, have shown even more efficacy, creating models that take into account textual and visual datasets in addition to semantic elements in text that baseline classifiers might overlook.

The application of Deep Learning methods has greatly enhanced the efficacy of fake news detection. By integrating visual and textual information, as well as semantic information, these models are able to identify fake news more effectively. With ongoing research in this field, it is expected that even more efficient means of detecting fake news will be developed, and efforts to stem the tide of misinformation will be aided.

Drawbacks of the Existing System

Despite the advancements in fake news detection systems, there are several challenges and drawbacks associated with these methods. One major issue is the quality and availability of data. Many fake news detection systems rely heavily on large datasets for training, but these datasets are not always readily available or diverse enough. Inadequate data quality can lead to inaccurate or biased models, which may not generalize well to unseen news articles or social media posts. Additionally, fake news constantly evolves, and attackers can adapt their strategies to evade detection. This requires continuous updates to models and datasets, which can be resource-intensive and difficult to maintain.

Another challenge is the difficulty in detecting new or subtle misinformation. Fake news often involves subtle manipulation, such as presenting partial truths or using clickbait headlines, which can be difficult for automated systems to identify. While deep learning models have proven effective, they may still miss these nuanced cases. Furthermore, the interpretability of deep learning models is a concern. These models are often seen as "black boxes," meaning it can be difficult to understand how decisions are being made, which can be problematic in high-stakes situations where the rationale behind a detection decision needs to be explained.

There is also the issue of bias in algorithmic models. Machine learning models, including those used for fake news detection, can inherit biases from the data they are trained on, leading to unfair or skewed results. This can cause certain sources or types of news to be overrepresented or underrepresented in detection accuracy. Despite advancements in automation, human intervention is still often required, especially for complex or nuanced cases. This makes the process less scalable and can lead to delays in the detection of fake news.

Moreover, fake news creators can use adversarial tactics to bypass detection systems, such as intentionally altering the phrasing or structure of articles to confuse algorithms. This makes it difficult for the system to correctly identify misinformation. Lastly, many models focus on structured textual information, which can overlook non-verbal cues, such as emotional appeal or image manipulation, that may be critical in detecting fake news. These challenges highlight the ongoing need for research and refinement in fake news detection techniques to address these limitations and make the systems more robust and reliable.

Proposed System Methodology

We suggested a trustworthy method in light of the shortcomings of the current one. The LIAR dataset can be used to identify fake news using the suggested method. Figure 4.1 displays a block diagram of the suggested system's general functionality. This research uses the Twitter dataset to discuss how well DSSM and enhanced RNN models work together to detect fake news. The DSSM under this hybrid model is composed of multi-layer nonlinear projection and word hashing. To obtain the long-range temporal associations, the output from the multi-layer nonlinear projection layer is subsequently sent to the enhanced RNN layer..

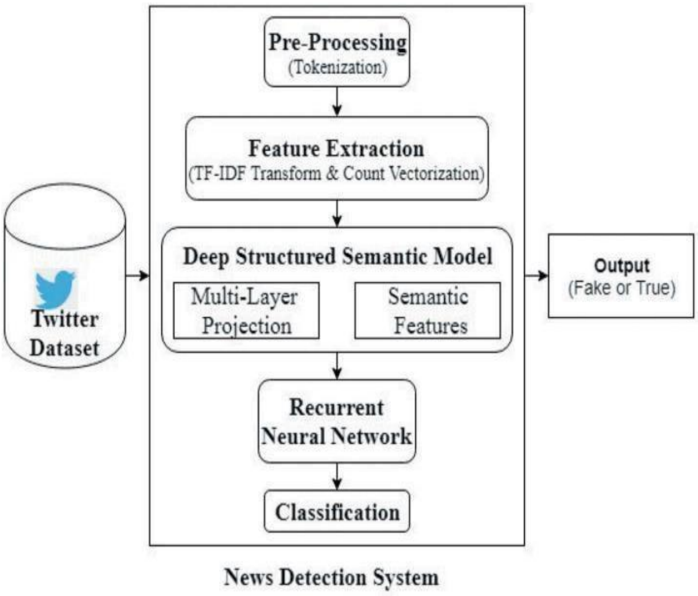


Fig 4.1 Block diagram of proposed system

Methodology

The proposed approach uses the Twitter dataset's DSSM and improved RNN model to detect and classify bogus news. The input layer of this model receives the input strings from the Twitter dataset. A bag of words is created by preprocessing these strings. These word bags are then supplied as input to the DSSM model. Three processes are performed in the DSSM model: word hashing comes first, followed by multi-layer nonlinear mapping and semantic feature construction. While the upgraded RNN layers in the DSSM architecture serve as the classification mechanism, these early levels serve as the feature extraction method. The following describes the overall arrangement of each layer that determines the neural network's architecture:

Input Layer

The input supplied to the input layer is the LIAR dataset. The titles and text of numerous news stories make up this dataset.

Preprocessing and Feature Extraction

Preprocessing transforms the data into a format that can be processed more quickly and easily for the user's intended use. The input Twitter dataset contains textual data that has been tokenized and preprocessed, and TF-IDF transformation is used to extract extra features (Buntain and Golbeck 2017). and vectorization of counts.

Deep Semantic Structural Model

The DSSM output is a concept vector in a low-dimensional semantic feature space, while the DSSM input (raw text features) is a high dimensional term vector, such as raw term counts in a query or a document without normalization. Figure 4.2 above shows the common DSSM architecture developed here to project the raw text features to the semantic space features. The goal of the word hashing technique described here is to compress the bag-of-words' term vectors. The method, which is specifically innovative for this purpose, is based on the letter n-gram.

Improved RNN

LSTM, as noted by Dimpas, Po, and Sabellano (2017), is an improved version of the Recurrent Neural Network (RNN). To solve the vanishing and exploding gradients problem, LSTM provides memory blocks as alternatives to the basic RNN units. LSTMs are much better than traditional RNNs in terms of dealing with long-term dependencies. This means that LSTMs are able to remember and link past knowledge to the current.

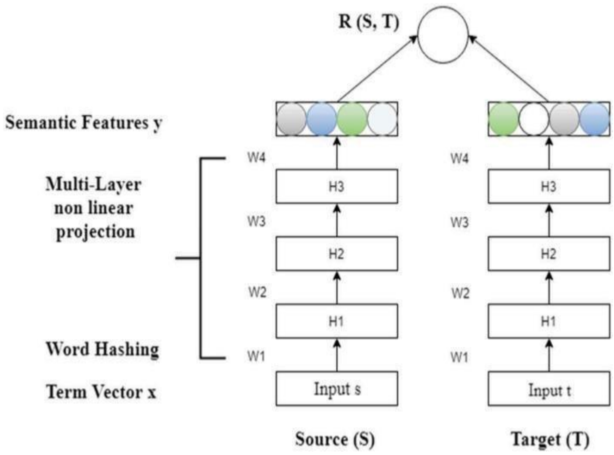


Fig 4.2 DSSM architecture

Output Layer

They received the output in binary format. Output values 0 (predicted false news yes) or 1 (predicted fake news no) are provided by the output layer.

The proposed system's flow

Based on the suggested system flow shown in Figure 4. 3, this system will help identify false news in the Twitter dataset. The suggested system is divided into two stages: Instruction and Evaluation

(1) Steps of Training phase

• Put the training dataset here.

• Use stemming and tokenization techniques to preprocess this data.

• Use the TF-IDF transform to extract features.

The DSSM-LSTM model is generated, and an object file is obtained.

(2) Steps of Testing phase

•Insert the test dataset's text query.

• Use a stemming and tokenization strategy to preprocess each text query.

• Use the TF-IDF transform to extract features, then transmit them to the Twitter API.

• The obj file receives the generated Twitter data.

• To determine the positive and negative incidence of the tweet header, classified tweet content from the obj file is used.

• The number of positive and negative counts is used to confirm news. (If the news is favorable, then it is negative, and vice versa.)

5. Deep Structured Semantic Model (DSSM)

Overview of DSSM in Fake News Detection

DSSM transforms raw news text into semantic representations that capture the meaning of words and sentences.

The transformed features are then fed into an enhanced RNN model (LSTM) to classify news as real or fake.

This approach helps overcome the limitations of traditional NLP techniques by focusing on deep contextual meaning rather than just surface-level text patterns.

5. Advantages of the Proposed System

- The suggested system has several strengths for identifying spurious news. By using the integration of DSSM and enhanced RNN (LSTM) models, the system combines semantic feature extraction and long-term dependency learning for enhancing accuracy in identifying spurious news. Word hashing and multi-layer nonlinear projection in the DSSM model effectively compress term vectors of bag-of-words' terms for efficient representation of text features and semantic meaning, which results in more accurate classification. The enhanced RNN model (LSTM) is good at overcoming the vanishing and exploding gradient problems, thereby suitable for learning long-term dependencies in the news content, important for detecting fake news from contextual understanding.
- Data preprocessing activities such as tokenization, stemming, and TF-IDF conversion enhance the quality of the data so that the system can extract useful features and discard redundant data, resulting in better fake news detection. The system uses data such as Twitter and the LIAR dataset, enabling it to learn a diverse range of patterns of fake news across various platforms, enhancing its generalization capacity. The modular design of the system—separating training and testing—is scalable. As more data are input into the system, it can keep improving its performance in identifying fake news.
- The output layer simplifies the outputs into binary classification (true or false news), hence the system becomes easy to interpret and integrate into a wide range of real-world applications. With emphasis on deep semantic understanding by means of DSSM, the system preserves the underlying semantic meaning of news content, which lessens dependency on surface-level features, that can be weak for identifying more subtle fake news. The interoperability with the

Twitter API and auto-feature extraction techniques enables the system to automatically verify news content as genuine, minimizing manual effort and accelerating the fake news detection process. These benefits make the suggested system a more efficient and powerful method for fake news detection than conventional techniques.

6. Results and Discussion

The recommended system is run on the Windows operating system and is written in Python using the PyCharm IDE. The following are the hardware requirements: Pentium 4.0 or higher processor, 3 GB or more of RAM, and 40 GB or more of storage Operating system type: 64-bit/32-bit. Preprocessed and standardized Twitter data sets are fed into a bag of words. The DSSM-LSTM model is used to further classify bags of words. In order to assess the system performance for the false news detection challenge, the results are retrieved in the form of a confusion matrix.

The confusion matrix is shown in Figand performance metrics like sensitivity, specificity, and accuracy are calculated using it.

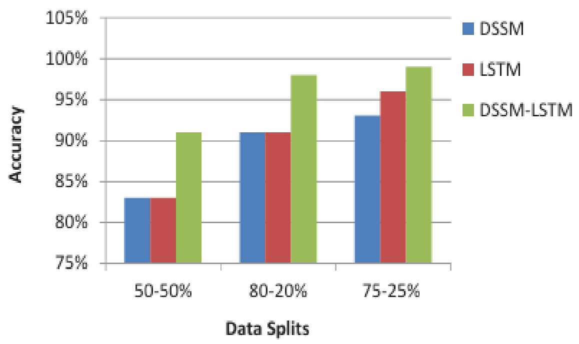
$$\boldsymbol{Accuracy} = (\boldsymbol{TP}+\boldsymbol{TN})/(\boldsymbol{TP}+\boldsymbol{FN}+\boldsymbol{FP}+\boldsymbol{TN})$$

Specificity = $\boldsymbol{TN}/(\boldsymbol{FP}+\boldsymbol{TN})$

Sensitivity= $\boldsymbol{TP}/\boldsymbol{TP}+\boldsymbol{FN}$

We have 99% accuracy, 99% specificity, and 100% sensitivity based on the confusion matrix and performance measure equation. The outcomes of the system are contrasted with those obtained by applying both algorithms separately. Based on accuracy, the classifiers for three distinct data splits—50–50%, 80–20%, and 75–25%—are compared in the following table.

The following graph in Figure 6 shows the overall comparison of hybrid DSSM-LSTM model with individual DSSM and LSTM based on accuracy.



7. Conclusion

In an effort to improve the accuracy of false news classification, this study proposed a deep learning-based fake news detection model that combines DSSM and enhanced RNN. When compared to traditional methods, the proposed method provides a more dependable framework for identifying fake news by effectively understanding textual semantics and contextual linkages.

According to trial results, the algorithm successfully and accurately detects bogus news, indicating that it may find practical use in thwarting misleading material on social networking sites. The model's capacity to extract semantic information and learn from textual patterns makes it a useful tool for identifying fake news in massive datasets. Notwithstanding the model's performance, issues including dataset bias, changing disinformation strategies, and the requirement for real-time processing still need to be investigated. These issues will be resolved and contribute to the creation of even more potent disinformation detection systems.

References

[1] Buntain, C., and J. Golbeck. 2017. Automatically identifying fake news in popular Twitter threads. 2017 IEEE International Conference on Smart Cloud (Smart Cloud 2017), New York City, NY, USA, November 3-5, 2017.

[2] Campan, A., A. Cuzzocrea, and T. M. Truta. 2017. Fighting fake news spread in online social networks: Actual trends and future research directions. 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, December 11-14, 2017.

[3] Castillo, C., M. Mendoza, and B. Poblete. 2011. Information credibility on Twitter. Proceedings of the 20th International Conference on World Wide Web, ACM, Hyderabad, India, March 28 - April 1, 2011.

[4] Dimpas, P. K., R. V. Po, and M. J. Sabellano. 2017. Filipino and English clickbait detection using a long short- term memory recurrent neural network. 2017 International Conference on Asian Language Processing (IALP 2017), Singapore, December 5-7, 2017.

[5] Granik, M., and V. Mesyura. 2017. Fake news detection using Naive Bayes classifier. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kyiv, Ukraine, May 29 – June 2, 2017.

[6] Gupta, A., and R. Kaushal. 2015. Improving spam detection in online social networks. 2015 International Conference on Cognitive Computing and Information Processing (CCIP), Noida, India, March 3-4, 2015.

[7] Liu, Y., and Y.-F. B. Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. Thirty-Second AAAI Conference on Artificial Intelligence, Hilton New Orleans Riverside, New Orleans, Louisiana, USA, February 2–7, 2018.

[8] Parikh, S. B., and P. K. Atrey. 2018. Media-rich fake news detection: A survey. 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Miami, Florida, USA, April 10-12, 2018.

[9] Ruchansky, N., S. Seo, and Y. Liu. 2017. CSI: A hybrid deep model for fake news detection. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM 2017), Singapore, November 6-10, 2017.

[10] Shu, K., A. Sliva, S. Wang, J. Tang, and H. Liu. 2017. Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter,19(1). doi:10.1145/3137597.3137600.