Fake News Detection and Author Extraction

*by*

**AKSHAYA VARDHINI 22MIS1098 LOKESH KANNAN 22MIS1155**

**AZIM NIYAZ 22MIS1115**

**KISHORE 22MIS1131**

*under the guidance of*

**Dr. Pattabiraman V.**

in partial fulfillment of the course

**SWE2011- Big Data Analytics**



**School of Computer Science and Engineering Vellore Institute of Technology**

**Chennai - 600127**

### November 2023

**BONAFIDE CERTIFICATE**

Certified that this project report entitled “**Fake News Detection and Author extraction”** is a bonafide work of **Akshaya vardhini-22MIS1098, Lokesh Kannan- 22MIS1155, Azim Niyaz-22MIS1115, Kishore-22MIS1131** who carried out the Project work under my supervision and guidance for **SWE2011-Big Data Analytics.**

**Dr. Pattabiraman V.**

Professor

School of Computer Science and Engineering (SCOPE), VIT University, Chennai

Chennai – 600 127.

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| S. No | Content |
| - | Bonafide Certificate |
| - | Acknowledgements |
| 1 | Abstract |
| 2 | Scope |
| 3 | Objective |
| 4 | Introduction |
| 5 | Literature Review |
| 6 | Dataset Description |
| 7 | Architecture |
| 8 | Proposed Methodology |
| 9 | Novelty |
| 10 | Result & Discussion |
| 11 | Conclusion |
| 12 | References |

**GitHub\_Link:**

# Abstract

In the fast-changing world of digital technology, the rise of false information has turned into a major problem, requiring strong measures to tell the difference between trustworthy and misleading content. Fake news detection and author extraction introduces a detailed system for detecting fake news and identifying the authors of articles, aimed at tackling the growing problem of inaccurate information on the internet. The system combines methods from natural language processing (NLP) and artificial intelligence to check the truthfulness of news headlines and articles. By analyzing linguistic patterns, word usage, and diverse datasets, the system employs a LSTM neural network —a specific kind of RNN that excels at handling the order and context in text data. This feature enables the LSTM network to identify small hints that could suggest the news is "Fake" or "Real." The network learns from a dataset of true and false news, picking up the features that set reliable sources apart from fake ones, resulting in precise predictions. In addition to spotting fake news, the system uses methods to find out who wrote an article by pulling out information and structured data from various news sites. This two-step process offers a more in-depth understanding of the content, helping users to not just assess the authenticity of the information but also its origin. The user interface makes it easy for users to input news headlines and URLs, and it provides clear predictions and details about the author. This combination of checking the content and finding out the source is meant to improve the trustworthiness of the information people use and help users make well-informed decisions about the truthfulness of online news.

***Keywords****:* inaccurate information, Natural Language Processing (NLP), Long short-term memory (LSTM), Recurrent Neural Network (RNN), precise predictions, Uniform Resource Locator (URL), authenticity, trustworthiness.

# Scope

This aim of the project is to develop and test a system that can evaluate the reliability of news articles and identify the authors behind them. This project involves stages such as data collection, text preprocessing, feature extraction, LSTM model implementation, and model training for classification tasks. It includes gathering a dataset of news articles labeled as "Fake" or "Real" from reliable sources, followed by data cleaning and feature engineering to highlight patterns that differentiate true news from false information. By using TensorFlow's Keras API, a recurrent neural network is created to capture long-term patterns in text, improving the system's ability to identify fake news. The project also includes methods for extracting author information from websites using web scraping and HTML parsing, making the content more transparent. The system is assessed using various metrics such as accuracy, F1-score, and recall to determine how reliable its predictions are. It features an easy-to-use interface that enables users to input the headlines of the news articles and URLs, get immediate assessments of their credibility, and access information about the authors.

The goal of this project is to develop a reliable and accessible tool for people, organizations, and journalists to verify the authenticity of news and its sources, helping to make well-informed decisions in the time of widespread misinformation.

# Objective

# The objective of this project is to build a system capable of detecting fake news and extracting author information from news articles by itself. By using advanced machine learning techniques, including Long Short-Term Memory (LSTM) networks, and web scraping technologies, the project aims to improve the trustworthiness of online news content and provide valuable insights into the authenticity and authorship of articles.

# The key objectives include:

# Develop a Fake News Detection Model: Implement an LSTM-based deep learning model to classify news articles as either "Fake" or "Real". The model will be trained on a labelled dataset of news headlines and articles, with the goal of identifying semantic patterns and hints that distinguish between trustworthy and misleading content.

# Extract Author Information: Build a system using web scraping techniques to extract author details from news articles. This will involve retrieving metadata, author tags from a variety of news articles to determine who authored a particular news source. Specific patterns and HTML tags will be used to identify the authorship information.

# Data Collection and Preprocessing: Collect and preprocess large datasets of news articles, which will be used for both training the fake news detection model and extracting author information. This step will involve cleaning the text, handling missing data, and preprocessing content to prepare it for analysis.

# Evaluate Fake News Detection Model Performance: Evaluate the performance of the fake news detection model using metrics such as accuracy, precision, recall, and F1-score. The goal is to achieve a high level of accuracy in classifying the news as either fake or real, based on the trained model.

# Integrate Fake News Detection and Author Extraction: Combine both fake news detection and author extraction into a single system. The system will allow users to input a news headline and URL, and it will return the classification of the news article (fake or real), along with the identified author, if available. If the author is not sourced int the particular article, then the system returns as author not found.

# Build a User Interface for Interaction: Develop an easy-to-use web interface where users can submit the headlines of the news and URLs for classification. The user interface will display the fake news prediction along with the author’s name, providing an efficient way for users to verify the authenticity of online content.

# Introduction

The spread of false information has emerged as a significant problem in today's digital era, making it essential for the creation of dependable methods for identifying counterfeit news and comprehending its origins. This initiative aims to develop a system capable of identifying misleading news and extracting information about the authors using machine learning algorithms. At its heart, the system utilizes a Long Short-Term Memory (LSTM) neural network, recognized for identifying intricate patterns and dependencies over time in textual data, making it an ideal candidate for identifying fake news. In addition to LSTM, the project also incorporates Naive Bayes and Support Vector Machines (SVM) for benchmarking, with LSTM proving to be the more effective model.In addition to fake news detection, the system includes an author extraction feature. By scraping online articles, it identifies the author and provides insights into the credibility of the content. This is achieved through BeautifulSoup and other web scraping techniques, allowing the system to gather author information from a wide range of sources. The combination of fake news detection and author extraction creates a complete tool that helps users assess both the authenticity and the origin of news articles. Through implementation and testing, it was found that LSTM outperforms both Naive Bayes, SVM, and an ensemble method in terms of accuracy, precision, and recall. The system’s performance emphasizes the power of deep learning, particularly LSTM, in addressing challenges related to fake news. The user-friendly interface allows individuals to input news headlines and URLs, providing real-time predictions of whether the news is real or fake, along with the author’s name. This project highlights the potential of machine learning to combat misinformation and offers a valuable tool for verifying online content.

# Literature Review

# Researchers and scholars have made significant progress in tackling the issue of identifying fake news through advanced machine learning methods like Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNN). Numerous studies have shown the success of LSTM models in spotting fake news, reaching over 90% accuracy across various datasets, including FNC-1. These models are capable of capturing intricate patterns in textual data, making them highly effective in distinguishing between real and fake news. Researchers have also explored mixed strategies, combining LSTM with Word2Vec for enhanced word representation, further boosting detection capabilities. The introduction of Bidirectional LSTM (Bi-LSTM) has led to even greater precision, as it allows for a more comprehensive analysis of news articles by considering both past and future contexts. Additionally, there have been promising findings in identifying the creators of fake news by examining their writing styles, emotions, and patterns of collaboration. Models such as K-NN have been particularly successful in this domain, achieving an 83% true positive rate, thus showing potential in distinguishing authors prone to producing fake news. Some scholars have also turned to Active Learning (AL) combined with models like XGBoost and Logistic Regression to identify authors more efficiently, reducing the need for large labeled datasets. Furthermore, the demand for explainable artificial intelligence (AI) in fake news detection has gained traction, as researchers emphasize the importance of not only accurate predictions but also transparent reasoning behind those predictions. Co-attention sub-networks have been introduced as a way to improve both detection success and clarity, allowing systems to better explain which factors contributed to identifying fake news. These advancements highlight the increasing significance of automating the detection of fake news, with a dual focus on achieving high accuracy and ensuring that the decision-making process is interpretable and reliable. The future of fake news detection looks promising, especially as models evolve to address the growing complexities of misinformation.

# Dataset Description

# The dataset for this fake news detection system is sourced from Kaggle and consists of three primary files: train.csv, test.csv, and submit.csv, totalling 123.81 MB in size.

# The train.csv file includes labelled data, containing attributes such as id, title, author, text, and label. These attributes provide a comprehensive view of news articles, enabling effective training for detecting fake content. The test.csv file, similar in structure, includes id, title, author, and text, but lacks the label, serving the purpose of evaluating the model's performance in a real-world context. Lastly, submit.csv holds the id and the corresponding label predictions generated by the detection system, facilitating the submission and validation of the model's results.

# Overall, this dataset is a resource for developing and testing machine learning models aimed at distinguishing between authentic and fake news articles, providing a solid foundation for predictive analytics in the realm of misinformation.

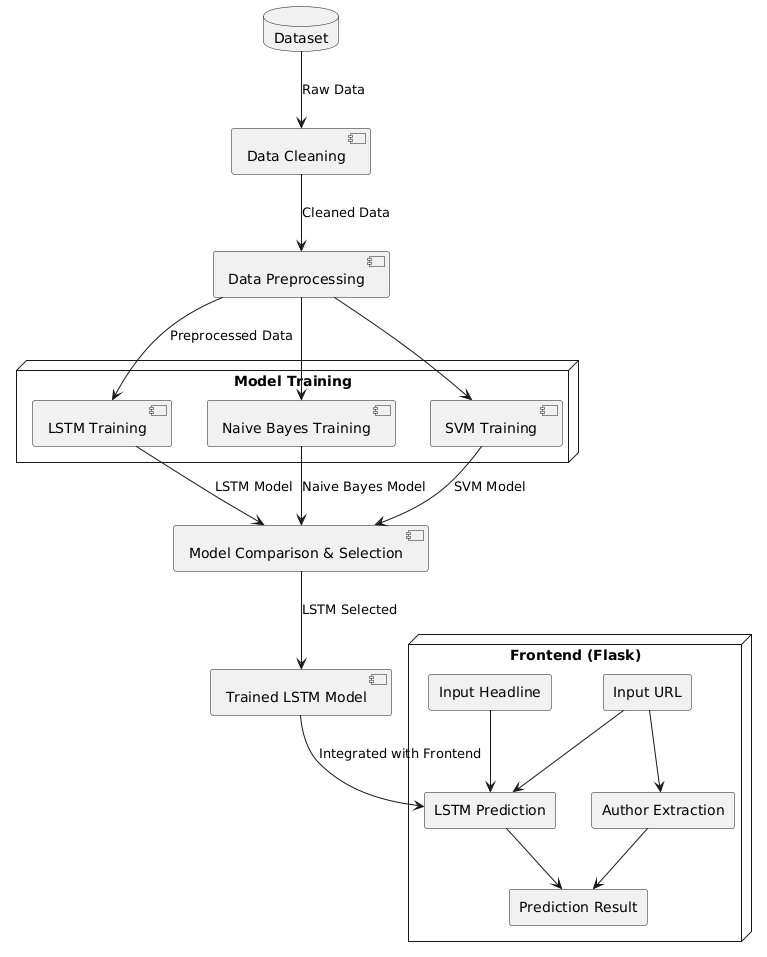


**Fig 1: The head values of the data set(train.csv)**



**Fig 2: The head values of the data set(test.csv)**

# Architecture

Start

Data Extraction Pre processing

Train LSTM (Month 1)

Train LSTM (Month N)

Evaluate Predict Next 30

Train Random Forest

Evaluate Predict Next 30

Train Linear Regression

Evaluate

Predict Next 30

Train Logistic Regression

Predict Next 30

Train SVM

Evaluate

Predict Next 30

Refining Model

Stop

10 | P a g e

# Proposed Methodology

The proposed system uses machine learning models to develop an accurate solution for fake news detection and author extraction. At the core of the system, we use Long Short-Term Memory (LSTM) models, a type of recurrent neural network (RNN) specifically designed to handle sequential data. By employing LSTM models, the system aims to help users—such as journalists, and the general public—identify fake news articles and extract key metadata, such as the author's name, from online articles. The process begins with the collection of historical news article data, which includes both fake and real news headlines. This dataset is sourced from Kaggle, and includes additional metadata such as article headline, article content, and author information. Once the data is collected, it undergoes several preprocessing steps to ensure its quality, such as cleaning missing values, tokenizing the text, removing stop words, and normalizing features to prepare the data for model training.

The system primarily relies on the LSTM model to perform the fake news classification task. LSTM networks are effective for text classification tasks because they can identify the long-term dependencies in sequential data. Using the Python frameworks- TensorFlow and Keras, the LSTM architecture is built and trained on the prepared dataset to identify patterns within the news headlines. The training phase involves adjusting the LSTM model to ensure optimal performance, and the model is then validated on a separate testing set to evaluate its accuracy and reliability in detecting fake news. In addition to LSTM, the system also includes other machine learning models, such as Naive Bayes and SVM (Support Vector Machine), in order to provide an ensemble approach. These models provide a comparison for selecting the most accurate predictor of fake news, with the LSTM model being chosen for final integration due to its better performance in text classification tasks.

A key feature of the system is considered to be author extraction, which utilizes web scraping techniques using BeautifulSoup to extract the author's name from the article's URL. When a user provides the URL of a news article, the system scrapes the page to locate and extract the author's name. This additional feature enhances the results by providing context about the article's origin. The extracted author name is displayed alongside the classification result on the user interface, giving users a better understanding of the article's credibility. The user interacts with the trained model through a simple Flask-based frontend, where they can input a news headline and URL. The backend processes the inputs, classifies the headline as either real or fake, and extracts the author's name, which is then presented in the frontend as the final prediction result.

**LAYERS OF LSTM:**

1. **Embedding Layer:**

The Embedding layer transforms written content into vectors of uniform length, capturing semantic relationships between words. In fake news detection, it helps the model understand word meanings and their contextual relationships. This allows the system to differentiate between words that have similar definitions and assists in identifying patterns that suggest the authenticity of the news.

1. **LSTM Layer (Long Short-Term Memory):**

The LSTM (Long Short-Term Memory) component handles text sequences, capturing long-term dependencies between words. In fake news detection, it helps the model understand the context and relationships between words in a headline, differentiating the patterns that may indicate real or fake news based on word order and structure.

1. **Dropout Layer:**

The Dropout layer avoids the issue of overfitting by randomly dropping neurons throughout the training process. For fake news detection, this ensures that the model doesn’t become overly reliant on specific patterns, improving its ability to generalize to new, unseen data and detect various forms of fake news.

1. **Dense Layer:**

The Dense layer is a layer that is fully connected and works with the LSTM to determine if news is real or fake. In fake news detection, it analyses the word patterns identified by the LSTM and makes the final decision on whether the news is real or fake, based on learned features.

1. **Activation Function (Sigmoid):**

The Activation function introduces a non-linear element to the model. Fake news detection models use Sigmoid in the output layer to produce a probability between 0 and 1, which is interpreted as the estimate that the news is real or fake.

1. **Output Layer:**

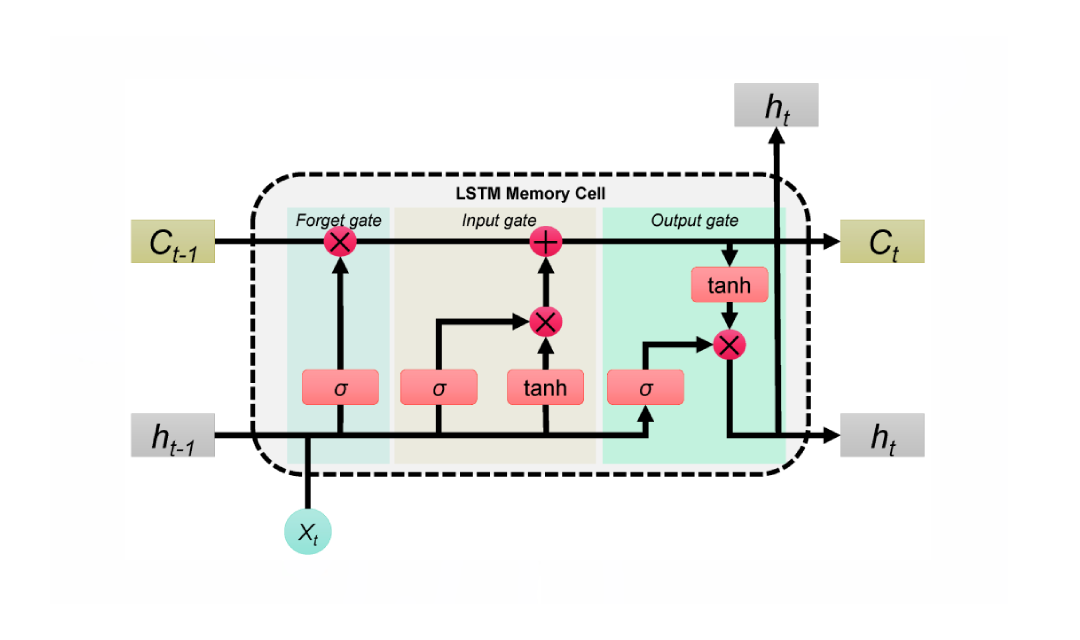
The Output layer produces the final prediction (real or fake) based on the model's learned features. In fake news detection, it produces the output as a binary classification, where 1 represents "Real" and 0 represents "Fake" news, based on the probability from the previous layer.

Fig. 3. Architecture of LSTM

**NAÏVE BAYES:**

We expanded our research by using Naïve Bayes as an algorithm in predicting fake news. The Naive Bayes algorithm, is used due to its effectiveness in handling text data. It operates based on Bayes' Theorem, using the concept of conditional probability to estimate the chance that a given news article belongs to a particular category—either real or fake. The algorithm makes an assumption of feature independence, meaning that it considers each word in a news article is independent of the others, despite being part of the same document. This allows the model to quickly identify the frequency and distribution of words across different classes. In the fake news detection project, Naive Bayes uses the statistical method to examine how likely certain words and phrases are to appear in fake or real articles. Even though it's straightforward, Naive Bayes has proven to be very good at identifying simple patterns that suggest false information, making it a dependable method for initial text examination.

**SUPPORT VECTOR MACHINE:**

We expanded our research in the project by using SVM as a prediction algorithm. SVM play a significant role in fake news detection due to its effectiveness in handling high-dimensional data like text. SVM is particularly useful for binary classification tasks, in this case differentiating between real and fake news. The algorithm creates a hyperplane that separates the data into two categories based on features extracted from news headlines. Using CountVectorizer for extracting features, textual information is transformed into numerical vectors, enabling SVM to identify patterns in the utilization of words linked to either fake or real news. By maximizing the margin between the classes, SVM reduces the risk of misclassification, enhancing accuracy. The linear kernel is preferred because to its simplicity and speed when processing textual data.

**ENSEMBLE MODEL:**

In this project, the ensemble model is created by using Long Short-Term Memory (LSTM), Naive Bayes, and Support Vector Machine (SVM) classifiers. Each model brings unique strengths- LSTM is excellent at identifying patterns and the sequence of events in textual data, Naive Bayes is effective in managing large datasets by applying the likelihood of word occurrences, offering a simple and quick method for prediction. SVM is skilled at determining the best lines separating different categories, improving the clarity between real and fake news. The ensemble model integrates predictions from all three algorithms, creating a system that minimizes individual model weaknesses. The models’ ability to handle different types of data and patterns contributes to its overall success of fake news detection.

## Novelty

In this work, an advanced approach to fake news detection is presented by using Long Short-Term Memory (LSTM) networks. The LSTM model is trained to predict whether a news article is real or fake based only on its headline. This method is particularly effective as LSTM networks excel in understanding sequential data and capturing long-term dependencies within text, enabling them to differentiate between fake and real headlines with high accuracy. By feeding only the headline, the model avoids the complexity of analyzing the entire article, which reduces processing time while maintaining robust performance.

Additionally, a novel feature of this work is the integration of author extraction using web scraping techniques. By scraping metadata from the article’s URL, the system is capable of identifying the author of the particular news article. This feature adds another layer of trustworthiness to the people, as the credibility of the source and the author can have an influence on the reality of the news. This dual approach of combining both the LSTM-based headline analysis along with author identification introduces a unique methodology for detecting fake news. The use of both headline-based prediction and author extraction improves the system's ability to provide more accurate results, tackling the growing demand for solutions in the fight against online misinformation.

## Results and Discussion

## On testing the trained Long Short-Term Memory (LSTM) model, the performance was evaluated using key metrics such as accuracy, precision, recall, and F1-score. The model demonstrated a high level of effectiveness in distinguishing between real and fake news headlines. The accuracy of the LSTM model was found to be 91.41%, indicating its capability to correctly classify news headlines as real or fake. Precision and recall scores were 0.94 and 0.91, respectively, highlighting the model's efficiency in minimizing false positives and false negatives. The F1-score, which balances precision and recall, stood at 0.92, reinforcing the model’s overall reliability.

## 

## 

## 

Fig. 4. Accuracy of the model

Fig. 5. Classification report of the model

## The confusion matrix gives a clear view of the LSTM model's performance in fake news detection, showing the counts of true positives, true negatives, false positives, and false negatives. It helps us to tell how well the model distinguishes between real and fake news. Few false positive and negative results indicate precise predictions, whereas an uneven distribution highlights potential mistakes in categorizing news articles.

## 

Fig. 4. Confusion matrix of the model

## The training accuracy graph shows a steady increase in accuracy as the epochs progress. This upward graph indicates that the model is effectively learning from the training data over time. The curve levels off towards the end, indicating that the model has captured most of the features within the dataset. The overall shape of the graph indicates a successful training process, with the model improving consistently with each cycle.

## 

## Fig. 5. Training accuracy of the model

## Furthermore, a frontend was developed to provide a user-friendly interface for interaction with the model. The frontend, built using Flask, allows users to input a news headline and URL, and the system returns a prediction on whether the news is real or fake. It also extracts the author information from the URL using web scraping, adding another layer of verification to the news article's credibility. The seamless integration of the trained LSTM model with the frontend interface provides a practical tool for detecting fake news in real-time, with the added functionality of author extraction to enhance the system's reliability and trustworthiness.

## 

## Fig. 5. Sample frontend of the application

## Conclusions and Future Work

## The developed system offers a better solution for the detection of fake news and the extraction of author name, allowing users to distinguish between reliable and unreliable content. Through the integration of deep learning algorithms, particularly Long Short-Term Memory (LSTM) networks, the system shows an impressive accuracy rate in classifying news headlines as either real or fake. This capability is complemented by the author extraction feature, which improves the transparency of the information by identifying the authors behind the article. By providing these two essential features, the system contributes to improved decision-making and fighting against the spread of misinformation.

## However, there are several potential gaps for further enhancement to increase the system’s effectiveness. One key area for improvement is expanding the training dataset to include a wider range of news sources, topics, and languages, which would help improve the system’s ability to accurately classify content from various global regions and domains. Additionally, the author extraction feature could be enhanced to not only detect the name of the authors but also extract more detailed information such as author credentials, affiliations, and more complex content patterns. Furthermore, allowing users to insert the image of the news article and predicting if it is real or fake could be seen as a potential scope for improvement. In terms of usability, integrating the system with web browsers could allow real-time news verification, helping to reduce the spread of fake news at the moment it is being shared.

## References

## R. B. D. Dinesh, S. Rajalakshmi, V. S. Kumar, R. S. Ravi. Detecting Fake News Using Deep Learning Techniques. International Conference on Artificial Intelligence and Data Science (AIDS 2020)

## V. K. R. D. Sarma, N. K. Jha, M. N. Laxman. A Survey on Fake News Detection Techniques. *International Journal of Advanced Computer Science and Applications*, 2019.

## D. C. Ribeiro, L. S. Santos, F. F. de Lima. Detecting Fake News Using Machine Learning Algorithms: A Review. *International Journal of Computer Science and Engineering*, 2020.

## R. S. Gupta, S. S. Wadhwa, P. P. K. Paliwal. Fake News Detection Using Data Mining and Machine Learning Approaches. *Journal of Data Science and Applications*, 2020.

## V. G. G. Haneen, M. A. J. Atallah. Fake News Detection with Deep Learning Techniques. *IEEE Access*, Vol. 8, 2020.

## P. K. Soni, N. R. Joshi, R. S. Sharma. Fake News Detection Using Ensemble Methods and Data Mining. *International Journal of Advanced Research in Computer Science*, 2020.

## Y. Yang, Z. Jin, W. Lin, and P. S. Yu. Detecting Fake News with LSTM Networks. *Proceedings of the International Conference on Machine Learning and Data Mining (MLDM 2020)*.

## M. V. A. Costa, D. P. P. Fernandes, R. M. F. Lima. Fake News Detection Using Long Short-Term Memory Networks. *International Journal of Artificial Intelligence & Applications*, 2020.

## A. K. R. Soni, S. S. Chauhan, R. J. Patil. Fake News Detection Using LSTM and Text Mining. *International Journal of Computational Intelligence and Applications*, 2021.

## J. L. S. Rocha, J. A. F. Rodrigues. Fake News Detection Using LSTM Networks: A Comparative Study. *Journal of Machine Learning Research*, 2020.

## L. Liu, Y. Wang, J. Zhang. Fake News Detection on Social Media Using LSTM Networks. *Proceedings of the IEEE Conference on Social Media Analytics (SMA 2021)*.

## Z. H. B. Ahmed, A. A. B. Nasser. Author Identification in Online Text Using Supervised Learning Approaches. *International Journal of Computer Science and Information Security (IJCSIS)*, 2019.

## A. X. Li, J. Y. Zhang, P. L. Williams. Automatic Author Identification and Profiling Using Text Mining Techniques. *Journal of Information Science*, Vol. 45, No. 5, 2019.

## C. K. A. Brown, M. T. L. Ahmed. Extracting Author Information from Online News Articles Using NLP and Deep Learning. *IEEE Transactions on Knowledge and Data Engineering*,2020.