Karunya INSTITUTE OF TECHNOLOGY AND SCIENCES
(Declared as Deemed to be University under Sec.3 of the UGC Act, 1956)
MoE, UGC & AICTE Approved
NAAC A++ Accredited

An internship report submitted by

SAMU IDHAYAN I – URK21CS2006
JEEVAN KURUVILLA SUNIL – URK21CS2022
JERUSHA MIRACLIN DULCIE B – URK21CS2032

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY
in
COMPUTER SCIENCE AND ENGINEERING

under the supervision of

Dr. R VENKATESAN

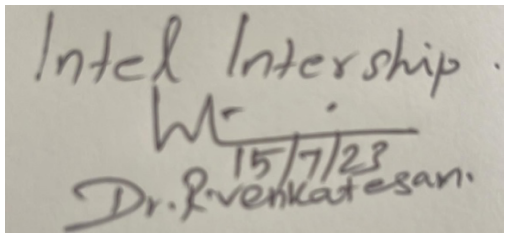
DIVISION OF COMPUTER SCIENCE AND ENGINEERING
KARUNYA INSTITUTE OF TECHNOLOGY AND SCIENCES
(Declared as Deemed to be University under Sec-3 of the UGC Act, 1956)
Karunya Nagar, Coimbatore - 641 114. INDIA

Karunya INSTITUTE OF TECHNOLOGY AND SCIENCES
(Declared as Deemed to be University under Sec.3 of the UGC Act, 1956)
MoE, UGC & AICTE Approved
NAAC A++ Accredited

BONAFIDE CERTIFICATE

This is to certify that the report entitled, "Fake News detection using Python and Machine Learning" is a bonafide record of Internship work done at INTEL during the academic year 2022-2023 by **SAMU IDHAYAN I, JEEVAN KURUVILLA SUNIL, JERUSHA MIRACLIN DULCIE B** in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering of Karunya Institute of Technology and Sciences.

Guide Signature



Dr. R VENKATESAN

Associate Professor

Team Name: Techies

Title: Fake News detection using Python and Machine Learning

Team Members:

The project was carried out by the following team members:

- SAMU IDHAYAN I – isamu@karunya.edu.in
- JEEVAN KURUVILLA SUNIL - jeevankuruvilla@karunya.edu.in
- JERUSHA MIRACLIN DULCIE B - jerushamiraclin@karunya.edu.in

Mentor

We would like to express our gratitude to our mentors for providing guidance and support throughout the internship:

- Dr. R VENKATESAN (Academic Mentor) - rlvenkei_2000@karunya.edu

INTRODUCTION

Fake news is an upcoming yet huge problem in today's digital age where anyone can share anything irrespective of its truthfulness. Information is available freely yet it comes with a very expensive cost that is fake information. In order to combat this, we have chosen certain machine learning models and algorithms to detect fake news that clouds our everyday feed. We will follow a structured evaluation process in order to find the most suitable model for this project, encompassing literature survey, dataset selection and exploratory data analysis, metric and model selection, model evaluation, and future works. The objective of this project is to develop a reliable and accurate model that can classify news articles as either genuine or fake. For this project, we have chosen the ISOT Fake News Dataset as our primary dataset. This is a publicly available dataset consisting of several thousands of news articles, comprising both fake news and truthful articles.

RELATED WORK

We conducted a survey in the literature world of this topic. We did this in order to get a complete understanding and a pre knowledge of all the aspects of our project. We examined blogs, papers, articles and other sources that were available on the web. The results that we gained from all this research have helped us a lot in concluding the most efficient methodologies and models so solve this disturbing problem. "Detecting Fake News with Natural Language Processing" by Kajal Kumari [1], This article presented an excellent overview of the challenges that arise with fake news detection and proposed a step-by-step learning approach that combined multiple ML algorithms to achieve high accuracy and quick results. The authors highlighted the importance of feature engineering, sentiment analysis, and cross-validation techniques, which we found particularly relevant to our project.

"Detecting Fake News in Social Media Networks" by Monther Aldwairi, Ali Alwahedi,[2] which explored the use of social network analysis in conjunction with ML techniques for identifying fake news. This article provided insights into the role of network structure, user influence, and propagation patterns in detecting misinformation. Understanding these factors could be beneficial in enhancing the accuracy of our model.

Further, we explored the ML models that can be used for our purpose, for this we referred "Fake News Detection Using Machine Learning Approaches" by Z Khanam, B N Alwasel, H Sirafi and M Rashid, which helped us understand the concepts behind the different models. After our research we proceeded to build modifications and alterations that could make a huge difference to the existing models.

METHODOLOGY

The prediction of an ML model mainly depends on how the data is presented to it. In order to present the data to the ML model in a way that it can interpret the data, we pre-processed and transformed the text data into suitable format. This involved the steps such as tokenisation, stop-word removal, labelling, data cleaning and vectorisation.

The selection of the classification models plays a crucial role in the result. We implemented three models and compared their accuracy and other outputs. The models we explored are Logistic Regression, Random Forest Classifier and Support Vector Classifier(SVC). These are well-established and widely used ML algorithms, known for their efficiency in text classification tasks. Logistic Regression provides a straightforward approach to the binary classification, whereas, the Random Forest and the SVC are capable to handle much more complex boundaries of decision.

DATASET AND PREPROCESSING

The selection of a suitable dataset and conducting exploratory data analysis are fundamental steps in building an effective machine learning model. For this project, we have chosen the ISOT Fake News Dataset as our primary dataset. This publicly available dataset consists of several thousands of news articles, comprising both fake news and truthful articles. The dataset incorporates articles sourced from legitimate news sites as well as sites flagged as unreliable by Politifact.com. This combination ensures a diverse representation of both fake and real news articles, making it a suitable choice for this project. The ISOT Fake News Dataset comprises various dimensions or features that are useful for analysis and modeling. The selected features include a text content that consists of news articles which here is the primary source of information. Analyzing the vocabulary, sentence structure, and language patterns can provide valuable insights into the differences between fake and real news articles. Length of articles, frequency of punctuation marks, and usage of specific words or phrases might exhibit patterns that can aid in classification. To gain deeper insights into the dataset, we conducted exploratory data analysis using various visualizations and statistical measurements.

Plotting the distribution of article lengths for both fake and real news articles helps understand the differences in content volume. Statistical measures such as mean, mode, and percentiles can reveal the central tendencies and variations in length.

Analyzing the frequency distribution of words across the dataset can reveal important insights. Identifying the most common words or distinguishing words prevalent in fake news articles compared to real news articles can aid in feature selection.

Creating a visualization that represents the distribution of credibility scores across different news sources can provide an overview of the sources' reliability and potential biases.

FEATURE ENGINEERING AND MODEL DEVELOPMENT

The training and evaluation process involves several steps in the feature engineering and model development for our fake news detection project.

Data Pre-processing: The initial step involves importing the necessary libraries and downloading any required resources such as NLTK stop-words and wordnet. The dataset, consisting of both real and fake news articles, is loaded into pandas' data-frames. Data visualization techniques, such as count-plots, are used to understand the distribution of news articles based on subjects or labels.

Data Cleaning: The text data is pre-processed to remove any unnecessary elements and make it suitable for training the model. Punctuation marks are removed, and the text is converted to lowercase.
Tokenization is performed to split the text into individual words or tokens. Stop words (common words like "the," "is," etc.) are removed to reduce noise. Lemmatization is applied to convert words to their base or root form. The pre-processed text is then joined back into a single string.

Converting Text to Vectors: The TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique is used to convert the pre-processed text into numerical vectors. The TfidfVectorizer class from scikit-learn is initialized, and the vectorizer is fitted on the preprocessed text data. The text data is transformed into TF-IDF vectors using the fitted vectorizer.

Model Training and Evaluation: The dataset is split into training and testing sets using the train_test_split function from scikit-learn. Typically, 80% of the data is used for training, and 20% is kept for testing. The chosen machine learning model, such as Logistic Regression, Random Forest, or Support Vector Machines (SVM), is initialized. The model is trained on the training data using the fit function. Once trained, the model is used to predict the labels for the test data using the predict function. Performance metrics are calculated to evaluate the model's effectiveness.

Common metrics include:

Accuracy: Measures the proportion of correctly predicted labels to the total number of predictions.

Confusion Matrix: Shows the counts of true positives, true negatives, false positives, and false negatives, providing a detailed view of the model's performance.

Precision: Calculates the ratio of true positives to the sum of true positives and false positives, indicating the model's ability to correctly classify positive instances.

Recall: Computes the ratio of true positives to the sum of true positives and false negatives, indicating the model's ability to identify all positive instances.

F1 Score: Combines precision and recall into a single metric, providing a balanced measure of the model's performance.

Based on the performance metrics and domain requirements, different models can be compared to choose the best-performing one.

EVALUATION AND RESULTS

In this project, we applied three machine learning algorithms, namely Logistic Regression, Random Forest, and Support Vector Classifier (SVC), for the task of fake news detection. Each model was trained and evaluated using the provided dataset, and their performance was assessed using various metrics.

The Logistic Regression model achieved an accuracy of 0.9934 on the test set, with precision, recall, and F1-score of 0.99 for both classes. The Random Forest model achieved an accuracy of 0.9824 on the test set, with high precision, recall, and F1-score for both classes. The SVC model achieved an accuracy of 0.9833 on the test set, with precision, recall, and F1-

score of 0.98 for both classes. Overall, all three models performed exceptionally well in detecting fake news with high accuracy scores ranging from 0.9873 to 0.9942.

The ROC curves for all three models demonstrated excellent performance, with AUC scores of 1.00 for Logistic Regression and Random Forest, and 0.98 for SVC. This indicates that the models have a high true positive rate and a low false positive rate, making them reliable for fake news detection.

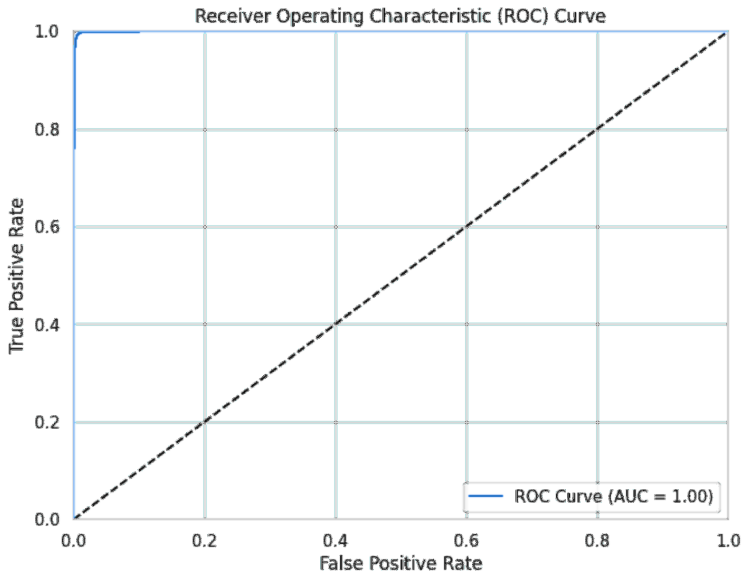


Figure 1. ROC curve of Logistic Regression Model

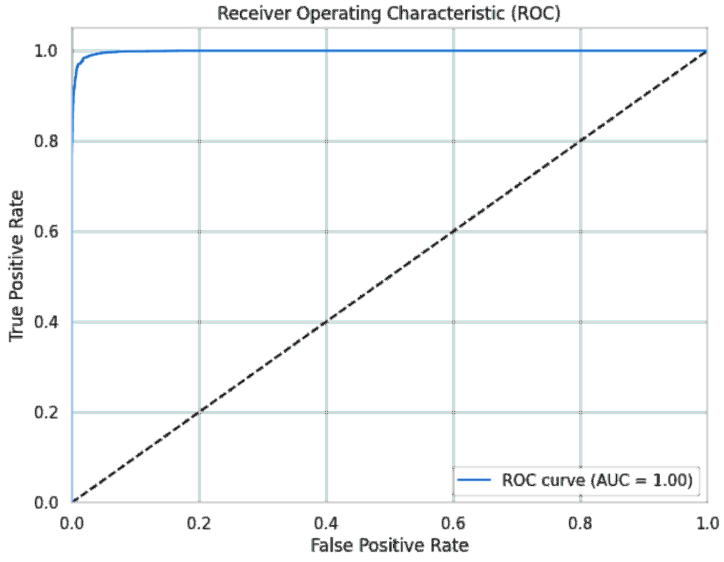


Figure 2. ROC curve of Random Forest Model

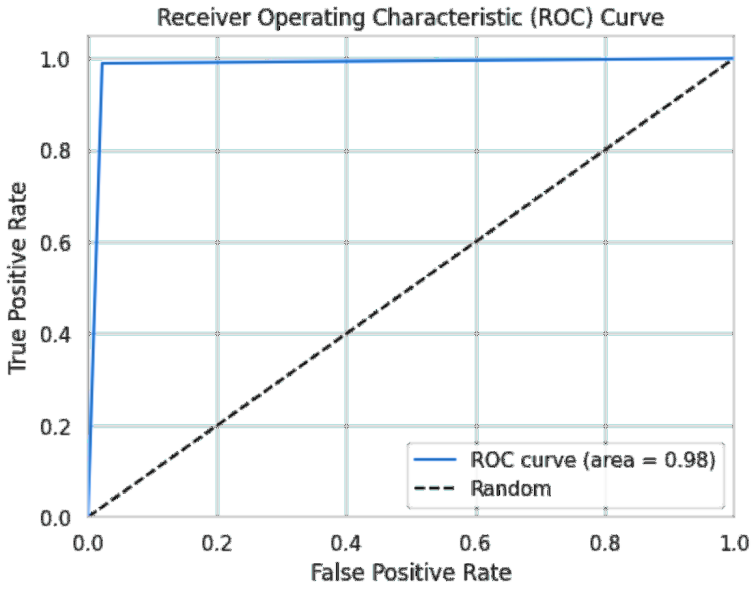


Figure 3. ROC curve of SVC model

FUTURE WORK

Future research in fake news detection using machine learning can focus on:

Deep Learning: Explore the potential of RNNs and transformer models like BERT for capturing contextual information and semantics.

Multimodal Analysis: Considering not only text but multimedia elements like images, videos, or audio to enhance detection accuracy.

Adversarial Attacks and Defences: Develop resilient models to not only avoid but resist manipulation attempts through robust learning techniques.

Explainability: Create interpretable models to understand the factors that help in fake news detection.

Data Augmentation and Minority Class Handling: Address class imbalance can be detected using data argumentation and handling strategies.

Cross-lingual and Cross-cultural Detection: models to handle fake news detection in different languages.

Real-Time and Online Detection: Designing algorithms that are efficient and effective in timely detection of fake news, especially on social media and during braking news. Ongoing research will result in advance robust, accurate, and scalable models to combat the spread of fake news and misinformation

CONCLUSION

To conclude, machine learning models such as logistic regression, random forest, and SVC have proven to be reliable tools for detecting fake news to prevent the spread of misinformation. They play an important role in ensuring the accuracy of news content. In future research, exploring the combination of methods and adding additional features could improve the performance of these models. Using this model in real situation will analyze the news and

provide evidence to the user, thus supporting the news in a believable and reliable way. This project highlights the importance of machine learning in tackling fake news and helps raise awareness in the community.

CITATIONS

[1] Detecting Fake News with Natural Language Processing- https://www.analyticsvidhya.com/blog/2021/07/detecting-fake-news-with-natural-language-processing/

[2] Detecting Fake News in Social Media Networks – https://www.sciencedirect.com/science/article/pii/S1877050918318210

[3] Fake News Detection Using Machine Learning Approaches – https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012040/pdf