Document Summarization for News Articles and Fake News Detection

SHREYA U CHAUDHARY[1], SHASHWAT TANDON[2], ANIKET FAND[3], ATHARVA KHADILKAR[4], URMILA PAWAR[5]

[1, 2, 3, 4, 5] *SCTR'S Pune Institute of Computer Technology*

Abstract- ***The recent surge of social media groups, forums, and pages with people from all walks of life sharing information about all worldly events ranging from festivals to global food crises. Purportedly, a lot of these blocks of information tend to be fabricated owing to reasons as simple as humor. In a country as big as ours where misinformation can cause mayhem of unprecedented scale. To prevent any such mishappening we aim to define a model which can study, learn, and then classify any such news as real or fake. Also, to give the user a more concise representation we seek to design a summarization model which will study the given news and present a distilled version with only the most relevant information intact.***

Indexed Terms- ***Naive Bayes Theorem, k-means clustering, decision tree, support vector machine, social media, Artificial Intelligence, etc.***

I. INTRODUCTION

The events are always changing around the world every minute, and due to advancement of technology we have information overload all around us be it audio, visual, or textual. Within this era of "information is power", misinformation becomes the weapon and can cause great discord in society if not controlled. A large chunk of the population across the globe gets its news from social media platforms. In the pandemic times when fake influences scared the general public about the vaccines which caused even further spread of the virus. Even during the recent war crisis, information about refugees, transport, annexed areas, etc. was spread in a dangerous manner causing mass confusion. Our model plans to eradicate such disarray such that fake news like this is instantly reported. Moreover, to reduce the complexity of the text so that even young learners can take part in knowing about burning issues across the world. We plan to make our model such that it can summarize genuine news and serve the simplified result to the user with the aim to reduce spreading of misinformation and also increase interest among people to take part in it.

II. LITERATURE SURVEY

In Fake News Detection system using Decision Tree algorithm and compare textual property with Support Vector Machine algorithm [1] The DT machine algorithm is used in examining whether a piece of news from social media is fake or not with accuracy that appears to be better than the SVM algorithm machine learning algorithm. There is a statistically significant difference among the study groups with significance value 0.092 for accuracy and 0.825 for precision for Confidence Interval (CI).

The work proposed in Automatic textual Knowledge Extraction based on Paragraph Constitutive Relations[2] the automatic textual knowledge extraction considering paragraph constitutive relations. Bi-directional Long Short-Term Memory (Bi-LSTM) + conditional random field (CRF) model on the existing CoNLL dataset is applied in entity recognition, dependency parsing is well processed in relation extraction. The test experiment has achieved good results and applies this method to ICDM competition.

In Fake News Detection Using Machine Learning Algorithms[3], perform binary classification of various news articles available online with the help of concepts pertaining to Artificial Intelligence, Natural Language Processing and Machine Learning. We aim to provide the user with the ability to classify the news as fake or real and also check the authenticity of the website publishing the news.

Text Summarization Techniques: A Brief Survey[4] , in this review, the main approaches to automatic text summarization are described. We review the different processes for summarization and describe the effectiveness and shortcomings of the different methods.

III. METHODOLOGY

In the proposed approach, there are steps. These are:

- User Interface: It is the area where the user will input the text of the news article that needs to be classified as fake or true. It also displays the summarized news and result whether the news is fake or not.
- Backend: The data is then sent to the backend servers using Axios to send the request from the Frontend and Fast API is used for backend.
- Classification module: The module uses 9 models for classification and then using Latent Dirchilet's Algorithm we perform classification.[7]
- Summarization module: If the news is true then using nltk library we summarize the article.

IV. ALGORITHMS USED

- Gradient boosting Algorithm:[8]

Gradient Boosting uses XGBoost, for fake and true news classification. It first performs data preprocessing, then divides your dataset into two parts: a training set and a test set. The training set will be used to train the Gradient Boosting model, while the test set will be used to evaluate its performance. Using the XGBoost algorithm to train a Gradient Boosting model on the training set. XGBoost is known for its ability to handle imbalanced datasets and its high predictive power. During training, the algorithm iteratively builds an ensemble of weak learners (decision trees) that sequentially minimize the loss function by fitting to the negative gradient of the loss. The GBC several hyperparameters that can be fine-tuned to improve model performance like, earning rate, the number of estimators (trees), the maximum depth of the trees, and the regularization parameters. Once deployed it can classify new, unseen news articles as fake or true.

- Light Gradient Boosting Machine:

LightGBM requires categorical features to be encoded as integers. Therefore, if your dataset includes categorical variables, you'll need to perform encoding techniques such as label encoding or one-hot encoding to convert them into numerical form. Using the LightGBM algorithm to train a gradient boosting model on the training set. LightGBM employs a leaf-wise tree growth strategy, which prioritizes the leaf nodes that contribute the most to the loss function during the tree building process. This approach leads to faster training times and potentially better performance. Like GBM, LightGBM has various hyperparameters that can be tuned to optimize the model's performance.

- Ada Boost Classifier:

AdaBoost combines multiple weak classifiers into a strong classifier. Using the AdaBoost algorithm to train a classifier on the training set. AdaBoost starts by training a weak classifier on the original data and assigns higher weights to misclassified samples. It then iteratively trains additional weak classifiers, giving more weight to misclassified samples in each iteration. The final classifier is an ensemble of these weak classifiers. Choose a weak classifier as the base learner for AdaBoost. Common choices include decision trees (often with limited depth or stump size), but other classifiers like logistic regression or support vector machines can also be used. Each weak classifier focuses on a specific aspect of the data, and their combination results in a more accurate model. Later we evaluate the model and tune hyperparameters to improve performance.

- Random Forest Classifier:[11]

Random Forest is a powerful ensemble learning algorithm that can be used for the classification of fake and true news. It combines multiple decision trees to create a robust and accurate classifier. Use the Random Forest algorithm to train a classifier on the training set. Random Forest builds an ensemble of decision trees by randomly selecting subsets of features and training individual trees on different subsets of the data. Each tree independently classifies the input, and the final prediction is made by aggregating the predictions of all the trees. To improve performance, we tune hyperparameters like the

number of trees (n_estimators), the maximum depth of the trees (max_depth), the number of features to consider at each split (max_features), and the minimum number of samples required to split an internal node (min_samples_split). Random Forest provides a feature importance score, which indicates the relative importance of each feature in the classification process. Analyzing feature importance can help you understand which features have the most impact on differentiating between fake and true news.

- Logistic Regression:

Logistic Regression is a popular classification algorithm that can be used for fake and true news classification. It models the relationship between the input features and the binary outcome (fake or true) using a logistic function. Using Logistic Regression to train a classifier on the training set. Logistic Regression models the relationship between the input features and the probability of a news article being fake or true. It applies a logistic function to the linear combination of the features to squash the output into the range [0, 1], representing the probability. We tune the hyperparameters like regularization parameter (C) that can be tuned to control the regularization strength and prevent overfitting. Logistic Regression can provide feature coefficients that indicate the importance of each feature in the classification process. Positive coefficients indicate features that contribute to classifying an article as true, while negative coefficients indicate features that contribute to classifying an article as fake.

- Ridge Classifier:

The Ridge Classifier is a linear classification algorithm that can be used for fake and true news classification. It is a variant of Logistic Regression that incorporates L2 regularization, also known as Ridge regularization. Use the Ridge Classifier algorithm to train a classifier on the training set. The Ridge Classifier applies Ridge regularization to the linear regression problem, helping to prevent overfitting and improve generalization. It estimates the coefficients of the linear model that best separates the fake and true news samples. The main hyperparameter to tune in the Ridge Classifier is the regularization strength (alpha). A higher value of alpha increases the amount of regularization, which can help prevent overfitting. Ridge Classifier can provide feature coefficients that indicate the importance of each feature in the classification process. Positive coefficients indicate features that contribute to classifying an article as true, while negative coefficients indicate features that contribute to classifying an article as fake.

- Linear Discriminant Analysis:

Linear Discriminant Analysis (LDA) is a dimensionality reduction technique and a classification algorithm that can be used for fake and true news classification. LDA seeks to find a linear combination of features that maximizes the separation between different classes while minimizing the within-class variance. Apply LDA to reduce the dimensionality of the feature space. LDA projects the data onto a lower-dimensional space while maximizing the separation between classes. The number of dimensions in the reduced space should be less than the number of classes. Use a classifier such as Logistic Regression, Support Vector Machines, or Naive Bayes on the reduced-dimensional data. The classifier will learn the decision boundary that separates fake and true news based on the transformed features. If you are using a classifier with hyperparameters, such as Logistic Regression or Support Vector Machines, perform hyperparameter tuning to optimize the model's performance.

- K neighbors Classification

K-Nearest Neighbors (KNN) is a simple yet effective classification algorithm that can be used for fake and true news classification. KNN classifies samples based on their proximity to other samples in the feature space. Train a KNN classifier on the training set. KNN assigns a class label to a sample based on the class labels of its nearest neighbors in the feature space. You need to specify the value of K, which is the number of neighbors to consider. The main hyperparameter in KNN is the value of K. Larger values of K provide a smoother decision boundary but may lead to misclassification if the classes have overlapping boundaries. Smaller values of K can capture more local patterns but may be sensitive to noise.

- Extra Trees Classifier

The Extra Trees Classifier is an ensemble learning algorithm that combines multiple decision trees to

create a powerful classifier. It is particularly useful for fake and true news classification tasks. Use the Extra Trees Classifier algorithm to train a classifier on the training set. The Extra Trees Classifier is an extension of the Random Forest algorithm, where each decision tree is trained on a random subset of features and uses random thresholds for node splitting. This randomness helps to reduce overfitting and improve generalization. The Extra Trees Classifier has several hyperparameters that can be tuned to optimize model performance. Key parameters include the number of trees (n_estimators), the maximum depth of the trees (max_depth), the number of features to consider at each split (max_features), and the minimum number of samples required to split an internal node (min_samples_split). The Extra Trees Classifier provides a feature importance score, which indicates the relative importance of each feature in the classification process.

- SVM- Linear Kernel

Support Vector Machines (SVM) with a linear kernel can be used for fake and true news classification. SVMs are powerful and versatile machine learning algorithms that can handle both linearly separable and non-linearly separable data. Train an SVM classifier with a linear kernel on the training set. The linear kernel assumes that the data can be separated by a hyperplane in the input feature space. SVM finds the optimal hyperplane that maximizes the margin between the two classes while minimizing the classification error. SVM has a regularization parameter (C) that controls the trade-off between maximizing the margin and minimizing the training error. Higher values of C allow for fewer training errors but may result in overfitting. Lower values of C increase the margin but may lead to more training errors.

- Naïve Bayes

Naïve Bayes is a probabilistic classification algorithm that can be used for fake and true news classification. It assumes that the presence of each feature is independent of the presence of other features, hence the term "naïve." Train a Naïve Bayes classifier on the training set. Naïve Bayes calculates the probability of a sample belonging to each class (fake or true) based on the presence or absence of features. It assumes that the features are conditionally independent given the class label. Naïve Bayes assumes that the features are independent, which may not always hold in real-world scenarios.

- Dummy Classifier

A Dummy Classifier is a simple baseline model that can be used for fake and true news classification. It is a classifier that makes predictions using simple rules or random strategies, without learning from the data. It serves as a benchmark to compare the performance of more sophisticated models. rain a Dummy Classifier on the training set. Dummy Classifier does not learn from the data but makes predictions based on simple rules or random strategies. It can serve as a baseline to compare the performance of other classification models. The Dummy Classifier does not learn from the data, so it does not require deployment. It serves primarily as a benchmark to compare the performance of other models.

- Quadratic Discriminant Analysis

Quadratic Discriminant Analysis (QDA) is a classification algorithm that assumes the features follow a Gaussian distribution and estimates separate covariance matrices for each class. QDA can be used for fake and true news classification by modeling the distribution of features in each class and making predictions based on these distributions. Train a Quadratic Discriminant Analysis model on the training set. QDA assumes that the feature distributions in each class follow a multivariate Gaussian distribution. It estimates separate covariance matrices for each class, allowing for quadratic decision boundaries. Once the QDA model is trained and evaluated, it can be deployed to classify new, unseen news articles as fake or true.

- Decision Tree Classifier

Decision Tree Classifier is a popular and intuitive machine learning algorithm that can be used for fake and true news classification. It builds a tree-like model of decisions based on the features in the dataset. Train a Decision Tree Classifier on the training set. The Decision Tree algorithm builds a tree-like model of decisions by recursively splitting the data based on the values of the features. It selects the best feature to split based on criteria like information gain, Gini index, or entropy. he Decision Tree Classifier has various

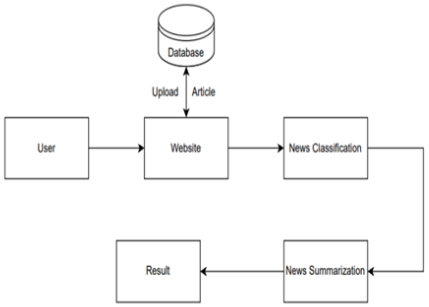
hyperparameters that can be tuned to optimize the model's performance. Important parameters include the maximum depth of the tree (max_depth), the minimum number of samples required to split an internal node (min_samples_split), and the minimum number of samples required to be at a leaf node (min_samples_leaf). ree Classifier provides a feature importance score, indicating the relative importance of each feature in the classification process.

V. SYSTEM ARCHITECTURE

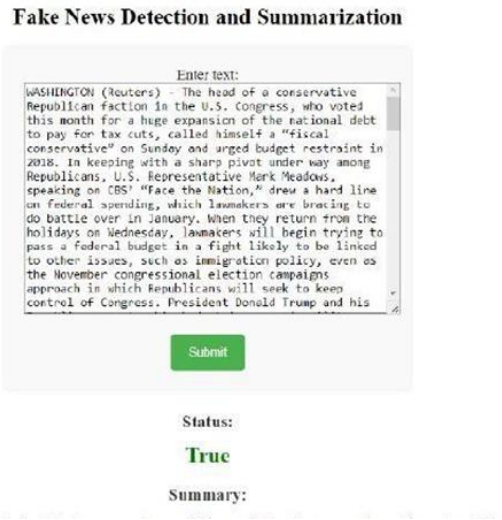
- Data Collection: The system starts by collecting a large dataset of news articles from various sources. These articles should cover a wide range of topics and include both legitimate and fake news articles. The dataset can be obtained from public sources or curated manually.
- Preprocessing The collected news articles undergo preprocessing to clean and prepare the text for further analysis. This includes steps like removing HTML tags, punctuation, and special characters, handling encoding issues, and converting the text to lowercase. Additionally, the articles may be split into paragraphs or sentences for easier processing.
- Fake News Classification: The fake news classification module takes the preprocessed news articles as input and aims to determine whether an article is genuine or fake. This can be achieved using machine learning or deep learning techniques. The module typically involves the following steps:

1. Feature Extraction: Extracting relevant features from the preprocessed articles, such as word frequencies, TF-IDF scores, or word embeddings.
2. Model Training: Model Training: Training a classification model, such as a Support Vector Machine (SVM), Random Forest, or a deep learning model like a Recurrent Neural Network (RNN) or Convolutional Neural Network (CNN), using the labeled dataset of genuine and fake news articles.
3. Model Evaluation: Assessing the performance of the trained model using evaluation metrics like accuracy, precision, recall, and F1-score.

- Summarization The summarization module aims to generate concise summaries of the news articles, capturing the key information and main points. There are various approaches to automatic text summarization, including:
- Extractive Summarization: Identifying and selecting important sentences or phrases from the original article to form a summary.[9]



VI. RESULTS



True News Result

Fake News Result

	Model	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC
gtK	Gradient Boosting Cla.ssifiet	0.8645	0.9364	0.85%	0.8627	0.8569	0.7283	0.7286
lightgbm	light Gradient	0.861	0.9332	0.8483	0.8582	0.853	0.7212	0.7215
ada	Ada Boost Classifier	0.8581	0.9306	0.8441	01560 0.8499		0.7154	0.7157
d	Rarpdom Forest Classifier	(18570	0.9273	0.8405	0.8564	1482	α7130	0.713.4
IT	logistic Regression	MM 1	0.9188	0.8576	0.8-343	1457	0.7019	0.7023
ridge	Ridge Classiher	0.8495	o.ano	0.8594	0.8306	1447	0.6989	0.6994
lda	lineav Discriminant Analysis	0.8493	0.9156	0.9591	1303	1444	0.6983	0.6989
knn	K Neighbors C1assiher	0.8464	0.9048	0.8348	0.8415	0.8379	0.692	0.6923
et	Extra Trees Classifier	0.8443	0.9198	0.8249	0.8445	0.8344	0ώ815	0.6879
svm	SVM-Linear Kernel	0.8435	0	0.8582	0.8214	0.8392	0.687	6881
nb	Naive Bayes	0.8187	0.8929	0.8615	0.7803	1186	0.6382	0.6419
(II	Decision Tree Classifier	0.8118	0.8109	0.7918	0.8088	1001	0.6224	0.6227
durnmy	Dummy Classifier	0.5244	0.5	0	0	0	0	0
%la	Quadratic Discriminant Analysis	0.5204	0.3911	1663	0.4838	0.3644	0.0374	0%

Model accuracy comparison

CONCLUSION

In this project, we have successfully developed a model that leverages various Natural Language Processing (NLP) algorithms to analyze news articles from diverse datasets and predict their authenticity. Our model demonstrates promising accuracy in distinguishing between fake and true news articles. Additionally, when an article is deemed true, the model provides a concise summary of the news as an output. By utilizing powerful NLP techniques, our model addresses the critical challenge of combating misinformation and fake news. The ability to automatically detect and summarize authentic news articles can greatly assist individuals, media organizations, and fact-checkers in making informed decisions and disseminating reliable information.

LIMITATIONS

- The model can process only English articles only.
- The model works for specific domains such as political news.
- The data will be labeled as genuine or fake and genuine news will be summarized.

FUTURE WORK

Although our model has achieved commendable results, there are several avenues for future work and enhancements. Here are some directions to explore: [10]

Dataset Expansion: Expanding the dataset used for training and evaluation can further improve the model's performance. Including diverse sources and domains will make the model more robust and adaptable to different types of news articles. [10]

- Fine-tuning and Transfer Learning: Investigating the application of transfer learning techniques, such as fine-tuning pre-trained languages models like BERT or GPT-3, can potentially enhance the model's accuracy and generalization capabilities.
- Multi-Modal Approaches: Incorporating additional modalities, such as images or videos associated with news articles, can provide richer context and help improve the model's accuracy in detecting fake news.
- Real-Time Monitoring: Developing a real-time monitoring system that continuously analyzes incoming news articles and provides instantaneous predictions and summaries can be valuable for newsrooms and social media platforms.

SUB REFERENCES

[1] N. L. S. R. Krishna and M. Adimoolam, "Fake News Detection system using Decision Tree algorithm and compare textual property with Support Vector Machine algorithm," 2022 International Conference on Business Analytics for Technology and Security (ICBATS), Dubai, United Arab Emirates,2022.

[2] Z. Peng, H. Song, B. Kang, O. B. Moctard, M. He and X. Zheng, "Automatic textual Knowledge Extraction based on Paragraph

Constitutive Relations," *2019 6th International Conference on Systems and Informatics (ICSAI)*, Shanghai, China, 2019

[3] Sharma, Uma & Saran, Sidarth & Patil, Shankar. (2020). Fake News Detection Using Machine Learning Algorithms. 2320-2882.

[4] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). Text Summarization Techniques: A Brief Survey. ArXiv. /abs/1707.02268

REFERENCES

[1] https://arxiv.org/abs/1707.02268

[2] https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2

[3] https://www.simplilearn.com/gradient-boosting-algorithm-in-python-article#:~:text.

[4] https://app.diagrams.net/-flow chart

[5] www.google.com

[6] https://www.javatpoint.com/machine-learning-random-forest-algorithm