Explainability in NLP model: Detection of Covid-19 Twitter Fake News

Wan Yit Yong
debby_yong123@hotmail.com
Technological University Dublin
Dublin, Ireland

Rajesh Jaiswal
rajesh.jaiswal@tudublin.ie
Technological University Dublin
Dublin, Ireland

Fernando Perez Tellez
fernando.pereztellez@tudublin.ie
Technological University Dublin
Dublin, Ireland

ABSTRACT

Fake news has found fertile ground on social media. A global health crisis such as COVID-19 further helps propagate fake news on social media. Much research has been done to develop AI systems that classify news as real or fake. However, there is a growing concern about trust in these AI systems. To this end, we attempt to improve the trustworthiness of AI text classification systems. We use tools to explore data, explain feature extraction techniques, interpret the ML models implemented, and explain the decision-making progress of AI systems. In this study, we compared five ML classifiers for our experiments: Naive Bayes, Support Vector Machines (SVMs), Logistic Regression, Decision Tree, and Random Forest. The models were trained on 10700 tweets containing 5,600 real and 5,100 fake tweets related to COVID-19. In comparison, the SVMs model performance was the best, with a detection accuracy of 0.93 and F1 scores of 0.94 and 0.93 for real and fake news, respectively. Global and local explanations are included to understand the overall model behavior, ensuring transparency and fostering confidence in AI users. We have chosen the SVMs model for the explanation section as it was the best model in this study.

CCS CONCEPTS

• Computing methodologies → Natural language processing; **Machine learning algorithms**; **• Human-centered computing**;

KEYWORDS

Fake News Detection, Natural Language Processing(NLP), Machine Learning Classifier, Explainable AI(XAI), Covid-19 News

ACM Reference Format:
Wan Yit Yong, Rajesh Jaiswal, and Fernando Perez Tellez. 2023. Explainability in NLP model: Detection of Covid-19 Twitter Fake News. In *Human Centered AI Education and Practice Conference 2023 (HCAIep '23), December 14–15, 2023, Dublin, Ireland.* , 7 pages. https://doi.org/10.1145/3633083.3633212

1 INTRODUCTION

Millions of posts about COVID-19 are circulating on social media. While some provide genuine and useful information, much is unverified. However, the fake content, including manipulated videos and photos, accounts for roughly 30-35% of all shared content, causing widespread panic [19]. The use of social media has been on the rise over the past decade. In 2023, social media will have a user base of more than 4.89 billion, and this figure is expected to increase to approximately 5.85 billion by 2027 [8], which is of serious concern to society on fake news spreading.

According to a study by the International Fact-Checking Network (IFCN) between January and April 2020, fake news related to COVID-19 can be classified into several categories, including information related to symptoms, causes, and treatments of the diseases; dissemination of government documents; discussion on the virus's transmission; manipulation of videos and images; statements from politicians; and conspiracy theories that assign blame to specific groups, nations, or communities for the virus's spread [5]. The spread of fake news on social media has led to health crises in some countries. Fake news ('Alcohol is a cure for COVID-19') led to many deaths and hospitalizations around the world [12]. This had a severe impact on the livelihoods of many people[5].

The ability to distinguish between real and fake news is a complex problem because fake news covers many topics, styles, and platforms. However, researchers say there are clues in the language that can give away deceptive news [24]. If we spot these clues, we can make an intelligent tool that's even better than our judgment. So, having good tools is essential to distinguishing real news from fake news, making it easier to find unreliable articles. Machine learning algorithms have proven to be an effective tool [2][3].

Moreover, Explainable Artificial Intelligence(XAI) is essential for Natural Language Processing(NLP) models, providing insight into their decision-making process. Achieving explainability in NLP models is essential for building trust, improving transparency, and ensuring the responsible deployment of AI systems. When people understand how a model arrived at a particular decision, they are more likely to trust it. This is because they are able to see the reasoning behind the decision and evaluate whether it makes sense. XAI techniques can help to make NLP models more transparent and interpretable, thereby increasing trust in their outcomes by unraveling their complex inner workings [31].

Having said that, this paper aims to explore the use of machine learning techniques for detecting and classifying Twitter fake news related to COVID-19. We will compare the performance of widely used NLP-based ML techniques to classify fake news and show why a particular ML technique performs better than the rest. We will also use a taxonomy of XAI (see section 2.2) to explain how the best-performing model makes the decisions.

The rest of the paper is organized as follows: Section 2 reviews the related work around fake news and XAI. Section 3 explains the methodology followed in this study, which also included the data explanation and the proposed models. The experimental results

HCAIep '23, December 14–15, 2023, Dublin, Ireland

ACM ISBN 979-8-4007-1646-1/23/12
https://doi.org/10.1145/3633083.3633212

are presented along with the discussion in Section 4. In Section 5, the model explanation will be explored. Future work is include in Section 7. Finally, the paper is concluded in Section 6.

2 RELATED WORK

As part of this study, we have reviewed the related work around fake news detection and XAI in Section 2, which includes the various techniques that have been used in the past to identify fake news and the definition and taxonomy of XAI.

2.1 Fake news detection

The problem of identifying fake news has been approached using traditional machine-learning techniques. Reis et al. [21] employed hand-crafted features encompassing both syntactic and semantic attributes through feature engineering. They tackled the problem as a classification task, utilizing established ML classifiers such as K-Nearest Neighbor (KNN), Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVMs), and XGBOOST (XGB), RF and XGB has produced better results. Shu et al. [26] have introduced the TriFN framework, which systematically characterizes the tri-relationship between authors, news stories, and readers. On an early version of the FakeNewsNet dataset, this framework significantly outperformed the baseline ML models and previous state-of-the-art frameworks [25].

The detection of non-factual news has undergone a significant revolution with the introduction of deep learning in the field of text classification. In comparison to earlier methods that relied on manually created features, Karimi et al.'s[11] proposed a multi-source, multi-class fake news detection framework. This framework can automatically extract features using Convolution Neural Network (CNN) based models and combine these features from multiple sources using an attention mechanism. A novel Capture-Score-Integrate (CSI) framework has been proposed by Ruchansky et al. [22] that makes use of the Long Short-Term Memory (LSTM) network to record the temporal spacing of user activity and a doc2vec [15] representation of a tweet in addition to a neural network-based user scoring module to determine whether a tweet is authentic or not. It underlines the need to take into account all three potent traits the content of the article, the user relationship, and the tweet response in order to identify fake news. Zellers et al. [32] develop a neural network model designed to check the accuracy of news based on the text content.

2.2 Taxonomy of XAI

XAI is a rapidly growing field of research that aims to develop AI systems that can provide clear and understandable explanations for their decisions and actions. The need for XAI arises from the increasing use of AI in critical applications such as healthcare, finance, and autonomous vehicles, where the ability to understand and trust the decisions made by AI systems is crucial. XAI is also important for ensuring that AI systems are fair, transparent, and accountable and for addressing ethical concerns related to AI. XAI research has made significant progress in recent years with the development of various methods and tools for providing explanations for AI systems. However, there are still many challenges and open questions that need to be addressed, such as how to balance accuracy and interpretability, how to evaluate the quality of explanations, and how to design AI systems that can provide explanations in real-time.

The paper "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence" [1] provides a comprehensive taxonomy of XAI techniques, which is divided into four axes using a hierarchical categorization system: data explainability, model explainability, post-hoc explainability, and assessment of explanations. The authors discuss the challenges and future directions of XAI, including XAI system design, generalization of XAI, user interactions with XAI, XAI ground truth evaluation, and advanced XAI tools. Overall, this paper provides a valuable resource for researchers and practitioners interested in developing transparent and trustworthy AI systems. We will explore the following XAI taxonomy for our model.

(1) **Data explainability** [1][23]: Data explainability refers to the ability of an AI system to provide clear and understandable explanations for the data used to train the model. This includes information about the sources of the data, the preprocessing steps used to clean and transform the data, and the features used to train the model. Data explainability is important for ensuring that the data used to train the model is accurate, unbiased, and representative of the real-world context in which the model will be used. Techniques for data explainability include data visualization, data profiling, and data lineage analysis. This will be seen in Sections 3.1 to 3.3.

(2) **Model explainability** [1][23]: Model explainability refers to the ability of an AI system to provide clear and understandable explanations for the model itself, including its structure, parameters, and decision-making processes. This includes information about the inputs and outputs of the model, the algorithms used to train the model, and the features that are most important for making decisions. Model explainability is important for understanding how the model works and for identifying potential biases or errors in the model. Techniques for model explainability include feature importance analysis, decision tree visualization, and model debugging. We will discuss in Sections 3.4 and 3.5 regarding the feature extraction and the model selection.

(3) **Post-hoc explainability** [1][23]: Post-hoc explainability refers to the ability of an AI system to provide clear and understandable explanations for its decisions and actions after they have been made. This includes information about the factors that influenced the decision, the confidence level of the decision, and the potential consequences of the decision. Post-hoc explainability is important for building trust in the AI system and for identifying potential errors or biases in the decision-making process. Techniques for post-hoc explainability include counterfactual analysis, sensitivity analysis, and local explanation methods. In Section 5, we will have a look at the explanations of how the SVMs model classifies in terms of overall view and individual prediction.

(4) **Assessment of explanations** [1][23]: Assessment of explanations refers to the ability of an AI system to evaluate the quality and effectiveness of its explanations. This includes information about the accuracy, completeness, and relevance

of the explanations, as well as the user's perception of the explanations. Assessment of explanations is important for improving the quality of the explanations and for identifying areas where the AI system can be improved. Techniques for assessment of explanations include user studies, human-in-the-loop evaluation, and metrics for evaluating the quality of explanations. In this paper, the results for the models will be shown in Section 4.

3 METHODOLOGY

This study's main goal is to employ ML approaches to classify COVID-19 fake news. There are 5 different ML models that have been used: Naive Bayes (NB), Support Vector Machines (SVMs), Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF). Models were trained with the CovidFakeNews dataset with the intention of classifying the articles in the news as real or fake. Additionally, we evaluated the performance using evaluation metrics such as accuracy, precision, recall, and F1-score. The total of 10 experiments were run using two feature extraction algorithms, including TF-IDF and Count Vectorizer, combined with different ML algorithms. Fig. 1 below shows the steps of the research methodology followed in this paper. The remainder of this section will introduce each part of the framework in detail.

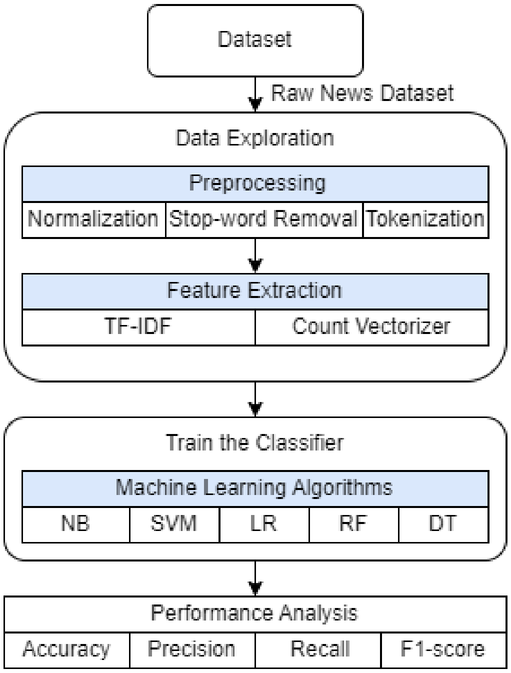


Figure 1: Methodology Framework

3.1 Data Exploration

In this study, we used the CovidFakeNews dataset that was publicly available from Kaggle [27], which was originally provided by Patwa R [18]. It consists of data that has been collected from fact-checking websites and several social media platforms, and the content of each post has been verified manually. The "real" news was collected from verified sources such as news companies that provide useful information, while the "fake" news was collected from tweets, posts, and articles that contain claims and speculations about COVID-19 that have been verified to be false. Real news is collected from verified Twitter users like the World Health Organization (WHO), the Centers for Disease Control and Prevention (CDC), etc. by using the Twitter API. Fake claims have been collected from fact-checking websites such as Politifact, Boomlive, NewsChecker, and some other tools like the IFCN chatbot and Google Fact-Check Explorer.

3.2 Data Statistic

The original dataset contains a total of 10,700 Twitter tweets, which include all the fake and real news related to COVID-19. The dataset is class-wise balanced, with 52.34% (5600) of the samples consisting of real news and 47.66% (5100) of the data being fake samples. The vocabulary size of unique words is 37,503 with 5141 duplicate words appearing in both fake and real news (see Table 1).

Attribute	Fake	Real	Combined
Unique Words	19728	22916	37503
Avg words per post	21.65	31.97	27.05
Avg chars per post	143.26	218.37	182.57

Table 1: Numeric Features of the dataset

3.3 Data Preprocessing

In the data preprocessing step, the text was cleaned and prepared to be used for NLP. The steps included stop-word removal and tokenization.

- **Text Normalization**[7]: Text normalization involves cleaning and standardizing text into a uniform format. This included special characters such as '@', '&', URLs, and digits. Besides that, all text was converted to lowercase.
- **Stop-words Removal**[7]: Stop words are frequently occurring words that do not affect their meaning in the text. For example, the words 'a', 'the', 'is', etc. will be removed.
- **Tokenization**[7]: The tokenization step involves breaking down a text or a sentence into its individual components, which are typically words or sub-word units. These individual components are known as tokens.

3.4 Feature Extraction

Before training the machine learning algorithms, feature extraction is required, as the words are the features to be used to train the model. Two techniques were used in this study, which are Term Frequency-Inverse Document Frequency (TF-IDF) and Count Vectorizer. After the data preprocessing part, each feature was performed on the dataset.

- **Term Frequency-Inverse Document Frequency (TF-IDF) Vectorization**[10]: TF-IDF is a statistical measure that evaluates the importance of each word in a document. This method calculates a weight for each word in the document based on its frequency and how many documents it appears in. The weight of a term is proportional to its frequency in the document but inversely proportional to its frequency across all documents in the corpus. This approach ensures that more weight is given to words that are rare in the corpus and frequent in the document.

- **Count Vectorization**[10]: The Count Vectorizer method is used to convert a collection of text documents to a matrix of token counts. This approach creates a vector for each text document with a length equal to the total number of distinct words in the corpus. Each element of the vector represents the frequency of the corresponding word in the document. Count vectorizer ignores the weights of the words and only counts their frequency in the document.

3.5 Proposed Models

(1) **Naive Bayes (NB)**[6][30]: Naive Bayes is a simple and effective algorithm for text classification. It works well with small datasets and can handle a large number of features. NB assumes that all features are conditionally independent, which may not always be the case in real-world applications. However, it has shown good performance in many NLP tasks, including sentiment analysis and spam filtering.

(2) **Support Vector Machines (SVMs)**[6][29]: SVMs are known for their ability to handle high-dimensional data and perform well on text classification tasks. They work by finding the hyperplane that maximally separates the data into different classes. SVMs have been widely used in various NLP tasks and have shown good performance in sentiment analysis, text classification, and named entity recognition.

(3) **Logistic Regression (LR)**[6][14]: Logistic Regression is a popular algorithm for binary classification tasks. It works by finding the best decision boundary that separates the data into different classes. It is simple to implement and can handle a large number of features. LR is often used as a baseline model in NLP tasks.

(4) **Decision Tree (DT)**[6][13]: Decision Trees are simple yet powerful algorithms for classification tasks. They work by constructing a tree-like model of decisions and their possible consequences. DT are easy to interpret and can handle both categorical and numerical data. However, they can be prone to overfitting if not properly tuned.

(5) **Random Forest (RF)**[6][4]: Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve performance and reduce overfitting. It works by building a collection of decision trees and aggregating their predictions. RF has shown good performance in many NLP tasks, including sentiment analysis and text classification.

Each algorithm has its own strengths and weaknesses, and these algorithms were chosen based on their popularity and proven performance in text classification tasks. For example, NB can work well with small datasets and has a low computational cost, while SVMs can handle large datasets and complex decision boundaries. LR can be used as a baseline model, while DT and RF can capture nonlinear relationships between the features and the class label [20].

3.6 Evaluation Metrics

(1) **Accuracy**[17]: Accuracy measures the overall correctness of your model's predictions. It calculates the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances. While accuracy is a useful metric, it can be misleading if the dataset is imbalanced (i.e., one class greatly outnumbers the other).

(2) **Precision**[17]: Precision quantifies the model's ability to correctly classify positive instances among all instances it predicts as positive. In other words, it measures how many of the predicted positive cases were actually positive. High precision indicates fewer false-positives.

(3) **Recall**[17]: Recall assesses the model's ability to identify all positive instances from the actual positives correctly. It measures the ratio of true positives to the total number of actual positives. A high recall indicates fewer false negatives.

(4) **F1-score**[17]: The F1-score is the harmonic mean of precision and recall. It balances both precision and recall, providing a single metric to evaluate a model's overall performance. It is particularly useful when there is an imbalance between the two classes because it considers both false positives and false negatives.

3.7 Training Strategy

For each ML algorithm, we used the default hyperparameters provided in the scikit-learn library. The dataset is randomly split into 70% for training and 30% for testing, ensuring that the classes were balanced in both sets. After that, the models will be trained on the training set and evaluated on the test set for each combination of ML algorithm and vectorizer to determine the optimal combination that produces the highest performance metrics by using the evaluation metrics mentioned above.

4 RESULTS

The performance of five different machine learning algorithms—Naive Bayes (NB), Support Vector Machines (SVMs), Logistic Regression (LR), Decision Tree Classifier (DT), and Random Forest Classifier (RF)—was evaluated using both Count and TF-IDF vectorizers. The evaluation was based on five performance metrics: Accuracy, Precision, Recall, F1-score, and confusion matrix.

From Table 2, the SVMs model achieved the highest accuracy of 92.32% within the TF-IDF, with precision, recall, and F1-score all above 0.92. The RF model also performed well, with an accuracy of 91.16%. On the other hand, the table shows that the best test model with Count vectorizer is the LR with an accuracy of 92.07%, closely followed by the SVMs model with 91.39% accuracy.

Overall, the SVMs model with TF-IDF vectorizer had the highest accuracy and F1-score of 92.3%, followed closely by the LR model with Count vectorizer, which had the highest accuracy and F1-score of 92.0%. SVMs and RF are known for their ability to handle high-dimensional data and can handle the complexities of text data, as well as LR, due to their simplicity and interpretability. In comparison, the DT reported significantly inferior performance in both vectorizers, with an accuracy below 85% in both cases, which is the lowest performance, due to the fact that the DT is prone to overfitting and can be sensitive to small changes in the data. For all the models, the respective precision and recall are close to each other.

The results suggest that the choice of vectorizer can have a significant impact on the performance of the machine learning models [20]. In this case, the TF-IDF vectorizer generally performed better than the Count vectorizer. Additionally, the SVMs and LR models showed consistently high performance across both vectorizers, indicating their robustness for this dataset [20].

In summary, decision trees are not inherently the best choice for text classification tasks, particularly when dealing with high-dimensional and sparse text data, especially with this dataset [16]. Other models, like SVMs and RF, which can handle the complexities of text data and mitigate issues like overfitting, may yield better results.

Vectorizer	Model	Accuracy	Precision	Recall	F1-score
TF-IDF	NB	0.905	0.909	0.903	0.905
	SVMs	**0.923**	0.922	0.923	0.923
	LR	0.910	0.910	0.911	0.910
	DT	0.844	0.843	0.843	0.843
	RF	0.911	0.911	0.912	0.911
Count	NB	0.906	0.907	0.906	0.906
	SVMs	0.913	0.913	0.914	0.913
	LR	**0.920**	0.920	0.921	0.920
	DT	0.840	0.840	0.839	0.839
	RF	0.912	0.912	0.912	0.912

Table 2: Experimental Results

Figure 2 shows the confusion matrix of SVMs model with TF-IDF vectorization. This indicated that the model shows high performance in distinguishing between the two classes as the number of wrong predict sample in each class is considerably low compared to the size of the testing set.

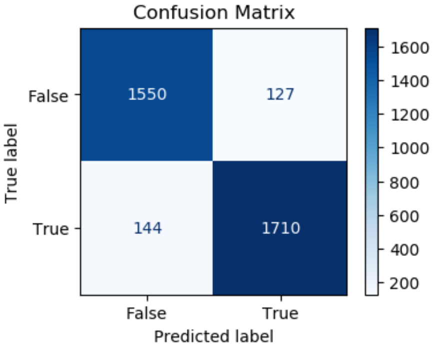


Figure 2: Confusion Matrix

5 XAI

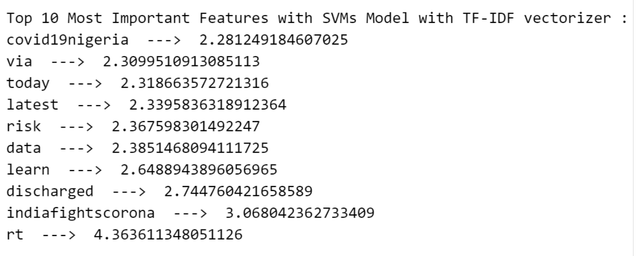
In this section, we chose the SVMs model with TF-IDF vectorizer for XAI; the reason for this is because it has the best result compared to all the models (see Table 2). The model will be discussed from both global and local perspectives. Global explanations provide a comprehensive overview of how a model functions across its entire dataset, identifying feature importance and performance metrics[28]. This broad view provides stakeholders with an understanding of the model's overall capabilities and limitations. On the other hand, local explanations focus on individual predictions, offering insights into why a specific decision was made[28]. By examining local explanations, users can gain context for individual model outputs, enhancing interpretability and trust. Both global and local explanations play pivotal roles in ensuring that machine learning models are not only accurate but also interpretable, fair, and accountable.

5.1 Global Explanation

In the context of Explainable Artificial Intelligence (XAI), a global explanation refers to an overall understanding or summary of how a machine learning model behaves across its entire input space. It provides insights into the model's decision-making process at a broader level, often highlighting general patterns, trends, or feature importance across various instances in the dataset[9].

Here, we provide a global explanation of feature importance based on the analysis of an SVMs model trained on text data using a TF-IDF vectorizer. Feature importance reflects the contribution of each term or word to the model's decision-making process. The higher the numerical value assigned to a feature, the more significant its impact on the model's predictions. It is essential to note that the importance scores provided for each feature are relative and signify their contribution to the SVMs model's predictions. These insights can guide further analysis and understanding of the dataset and may be valuable for decision-making in the context of the analyzed text data.

In figure 3, we noticed that 'rt' is the most significant feature with a higher importance score. This is because 'rt' refers to 'Russia Today' and is a Russia-state-controlled news company funded by the government, which implies that content from 'rt' will play an important role in the model's decision-making process. Besides that, the terms for 'risk', 'data', and 'restrictions' also show very high weights, as these terms actually bring a sense of seriousness when showing up in the news.



```
Top 10 Most Important Features with SVMs Model with TF-IDF vectorizer :
covid19nigeria  --->  2.281249184607025
via  --->  2.3099510913085113
today  --->  2.318663572721316
latest  --->  2.3395836318912364
risk  --->  2.367598301492247
data  --->  2.3851468094111725
learn  --->  2.6488943896056965
discharged  --->  2.744760421658589
indiafightscorona  --->  3.068042362733409
rt  --->  4.363611348051126
```

Figure 3: Global Explanation

5.2 Local Explanation

While global explanations provide an overarching view of a model's behavior, local explanations dive into the intricacies of a single instance. Local explanations aim to shed light on which features or factors influenced a particular model output, providing transparency and interpretability for individual predictions. They are particularly valuable for making machine learning models more accountable and for gaining insights into the reasoning behind specific outcomes[9].

In this section, we will give more details on the model's prediction for each sentence, which refers to the local explanation. We will show two false COVID-19 news articles, and the model will demonstrate how it predicts each word. The graphs were produced using Local Interpretable Model-Agnostic Explanations (LIME) that will be shown in three parts: the prediction probabilities for overall results, the weight for each word to help the model calculate the probabilities for both true and false predictions, and text with highlighted words. The blue highlighted word means that the definition of the word is more likely to appear in false articles, while the orange highlighted word will appear in true articles.

For the first example, the SVMs model predicts the sentence as fake news, which is a correct prediction (see Figure 4). The figure shows how each word influences the prediction with each weight. Besides that, we observe that there is a more blue highlighted word in the text with the highlighted words section. As we notice, the words 'alcohol' and 'chlorine' are considered false sentiments; this is because, during the COVID pandemic, there was news that drinking alcohol would cure or prevent Coronavirus, which is considered false news.

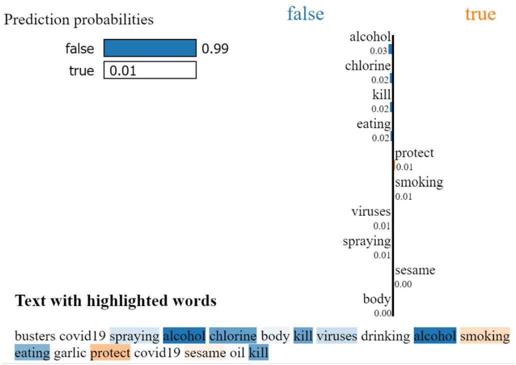


Figure 4: Local Explanation of True Prediction

The model had made an incorrect prediction for the articles in the second example (see Figure 5). This is due to words like 'covid', 'hospitalized', and 'disease' usually appearing in true news, which contains more serious cases; for example, news like 'The number of people reporting to be hospitalized with COVID-19 continues to drop. 57% of COVID-19 hospitalizations are currently in the South, while the Northeast has fallen to 5%.'

6 CONCLUSION

In conclusion, this is a simulation of a comparative study of the performance of five ML methods for the detection of fake news. It is important to note that certain replications can be expected and these performances are only based on a particular dataset. Moreover, this research serves as a foundational exploration, contributing to the taxonomy of XAI. Shifting focus to the realm of COVID-19-related misinformation on Twitter, our investigation delves into various NLP-based machine learning techniques. With a dual purpose of identifying superior methods for fake news classification and elucidating the rationale behind their effectiveness, this study provides insights into the landscape of combating misinformation in the context of a global health crisis.

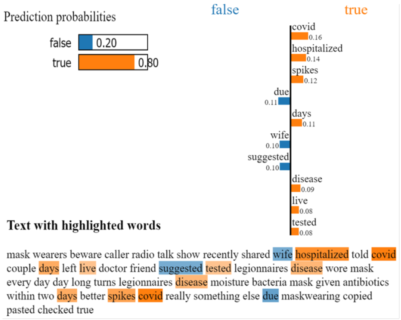


Figure 5: Local Explanation of False Prediction

Through a comparative analysis of performance metrics, SVMs emerged as the superior choice due to their adept handling of complex data, surpassing other models. While acknowledging the potential of more intricate models like deep learning to capture latent patterns, our selected machine learning techniques proved sufficient for our research scope and dataset. Emphasizing transparency, we employed a four-point XAI taxonomy, addressing data explainability, model explainability, post-hoc explainability, and assessment of explanations. This approach with SVMs enhances trust in AI decision-making by providing a clear account of the process, aligning with a human-centered approach and ensuring thoughtful AI system design.

In summary, the detection of COVID-19 fake news using explainable NLP models is an important area of research that has the potential to improve the accuracy and transparency of fake news detection. By continuing to explore ways to improve prediction accuracy and develop XAI tools, we can develop more effective and trustworthy AI systems that benefit society.

7 FUTURE WORK

While this study focuses on explaining the best-performing algorithm in XAI section, it is vital to acknowledge certain limitations in this paper. Future research could broaden its scope by comparatively assessing multiple NLP algorithms, explaining both global and local methods for every methods used, to see which methods work better for making NLP models easier to understand. This would provide valuable insights into the trade-offs between model performance and interpretability, contributing to a more comprehensive understanding of explainability for us.

REFERENCES

[1] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion* 99 (2023), 101805. https://doi.org/10.1016/j.inffus.2023.101805

[2] Malak Aljabri, Amal A Alahmadi, Rami Mustafa A Mohammad, Menna Aboulnour, Dorieh M Alomari, and Sultan H Almotiri. 2022. Classification of firewall log data using multiclass machine learning models. *Electronics* 11, 12 (2022), 1851.

[3] Malak Aljabri, Sumayh S Aljameel, Rami Mustafa A Mohammad, Sultan H Almotiri, Samiha Mirza, Fatima M Anis, Menna Aboulnour, Dorieh M Alomari,

Dina H Alhamed, and Hanan S Altamimi. 2021. Intelligent techniques for detecting network attacks: review and research directions. *Sensors* 21, 21 (2021), 7070.

[4] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[5] J Scott Brennen, Felix M Simon, Philip N Howard, and Rasmus Kleis Nielsen. 2020. *Types, sources, and claims of COVID-19 misinformation.* Ph. D. Dissertation. University of Oxford.

[6] DeepAI. 2019. Machine Learning. https://deepai.org/machine-learning-glossary-and-terms/machine-learning

[7] Deepanshi. 2023. Text preprocessing in NLP with python codes. https://www.analyticsvidhya.com/blog/2021/06/text-preprocessing-in-nlp-with-python-codes/

[8] S Dixon. 2023. Number of worldwide social network users 2027. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/. Accessed: 2023-11-7.

[9] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[10] Abhishek Jha. 2023. Vectorization techniques in NLP [guide]. https://neptune.ai/blog/vectorization-techniques-in-nlp-guide

[11] Hamid Karimi, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. 2018. Multi-source multi-class fake news detection. In *Proceedings of the 27th international conference on computational linguistics.* 1546–1557.

[12] Nasser Karimi and Jon Gambrell. 2022. Hundreds die of poisoning in Iran as fake news suggests methanol cure for virus. *Times of Israel.[online] Available: https://www. timesofisrael. com/hundreds-die-of-poisoning-in-iran-as-fake-news-suggests-methanol-cure-for-virus* (2022).

[13] Sotiris B Kotsiantis. 2013. Decision trees: a recent overview. *Artificial Intelligence Review* 39 (2013), 261–283.

[14] Michael P LaValley. 2008. Logistic regression. *Circulation* 117, 18 (2008), 2395–2399.

[15] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning.* PMLR, 1188–1196.

[16] Shikun Lyu and Dan Chia-Tien Lo. 2020. Fake News Detection by Decision Tree. In *2020 SoutheastCon.* 1–2. https://doi.org/10.1109/SoutheastCon44009.2020.9249688

[17] Sahil Mankad. 2020. A Tour of Evaluation Metrics for Machine Learning. https://www.analyticsvidhya.com/blog/2020/11/a-tour-of-evaluation-metrics-for-machine-learning/

[18] Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an infodemic: Covid-19 fake news dataset. (2021), 21–29.

[19] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science* 31, 7 (2020), 770–780.

[20] Tomas Pranckevičius and Virginijus Marcinkevičius. 2017. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing* 5, 2 (2017), 221.

[21] Julio CS Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Supervised learning for fake news detection. *IEEE Intelligent Systems* 34, 2 (2019), 76–81.

[22] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.* 797–806.

[23] Gesina Schwalbe and Bettina Finzel. 2023. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery* (2023), 1–59.

[24] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36.

[25] Kai Shu, Suhang Wang, and Huan Liu. 2017. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709* 8 (2017).

[26] Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining.* 312–320.

[27] Saikat Dutta Sourya Dipta Das, Ayan Basak. 2021. COVID19 Fake News Dataset NLP. https://doi.org/10.34740/KAGGLE/DSV/2016658

[28] Ilse van der Linden, Hinda Haned, and Evangelos Kanoulas. 2019. Global aggregations of local explanations for black box models. *arXiv preprint arXiv:1907.03039* (2019).

[29] Lipo Wang. 2005. *Support vector machines: theory and applications.* Vol. 177. Springer Science & Business Media.

[30] Geoffrey I Webb, Eamonn Keogh, and Risto Miikkulainen. 2010. Naïve Bayes. *Encyclopedia of machine learning* 15, 1 (2010), 713–714.

[31] Swathi Y and Manoj Challa. 2023. A Comparative Analysis of Explainable AI Techniques for Enhanced Model Interpretability. In *2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN).* 229–234. https://doi.org/10.1109/ICPCSN58827.2023.00043

[32] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. (2019).