Identifying Real vs. Fake COVID-19 News on Social Media:

A Text Mining Approach

SAITEJA NAMANI

INTRODUCTION

Background

Social media has grown to be an integral part of everyday life in the digital age, serving as a vital information source for billions of people globally. Social media platforms now have billions of active users as of 2023; by 2025, forecasts indicate that this number will rise to 4.41 billion. This extraordinary surge highlights how important social media is to the sharing and exchange of knowledge. Its ability to work quickly makes it possible to do a wide range of tasks, including improving brand awareness and speeding up consumer interaction and feedback channels.

Significance of the Study

The purpose of this paper is to create a solid foundation for methodically identifying and classifying content as genuine or fraudulent by applying text mining and machine learning approaches. By reducing the spread of false information, our research hopes to aid in the worldwide effort to combat the COVID-19 epidemic. This work is significant in a number of ways, including helping public health authorities distribute correct information, helping social media platforms stop the spread of false information, and eventually enabling people to make decisions based on verifiable facts. This research establishes a standard for handling any future emergencies that may arise in the digital sphere in addition to having the ability to save lives during the epidemic.

DATASET DESCRIPTION

In the COVID-19 Tweet Analysis Project, we have curated a specialized dataset focused on distinguishing between real and fake news related to the COVID-19 pandemic. This dataset is composed of two primary categories of tweets, characterized as follows:

- Real News Tweets: This category includes tweets sourced from verified and credible sources. These tweets provide useful, accurate, and factual information regarding various aspects of the COVID-19 pandemic.
- Fake News Tweets: The second category comprises tweets, posts, and articles that contain claims or speculations about COVID-19 which have been verified to be untrue. This category includes a range of misinformation, from unfounded medical advice and conspiracy theories to incorrect statistics and fraudulent health warnings.

Collection and Annotations:

We follow a simple guideline during the data collection phase as follows:

- Content is related to the topic of COVID-19.
- Only textual English content is considered. Non-English posts are skipped.

Fake News:

The fake news was dataset was gathered from various public fact-verification websites and social media platforms. Each post was manually cross-referenced with original documents for accuracy. Our sources included a diverse mix of web-based resources like Facebook posts, tweets, news articles, Instagram posts, public statements, press releases, and other popular media content. Additionally, we utilized prominent fact-checking websites such as PolitiFact, Snopes, and Boomlive, which are known for their comprehensive collation and verification of viral claims. These sites provide detailed verdicts on a wide range of topics, including COVID-19, thereby offering a rich repository of factually verified fake content.

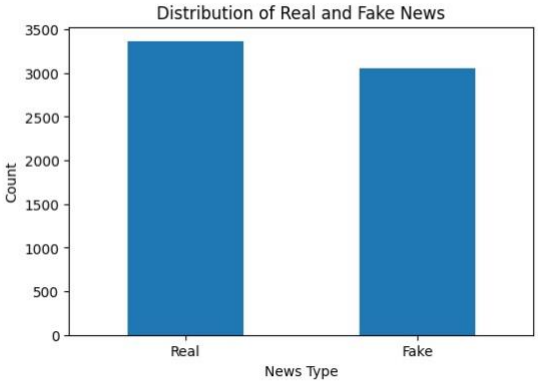
Real News:

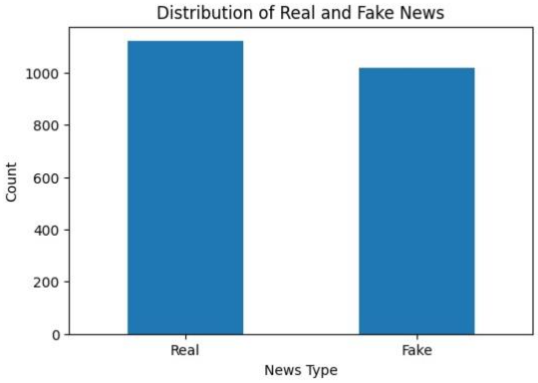
The dataset of real news tweets was collected through the Twitter API, focusing on tweets from official and verified accounts related to the COVID-19 pandemic. These sources encompassed a range of authoritative entities such as government bodies, medical institutions, and news channels. The dataset included tweets from 14 prominent organizations, including the World Health Organization (WHO), the Centers for Disease Control and Prevention (CDC), Covid India Seva, and the Indian Council of Medical Research (ICMR). Each tweet in this dataset had been carefully reviewed by human evaluators to confirm that it contained valuable and factual information about COVID-19, including statistics, dates, vaccine development updates, government policies, and details about hotspots.

DATA PREPROCESSING

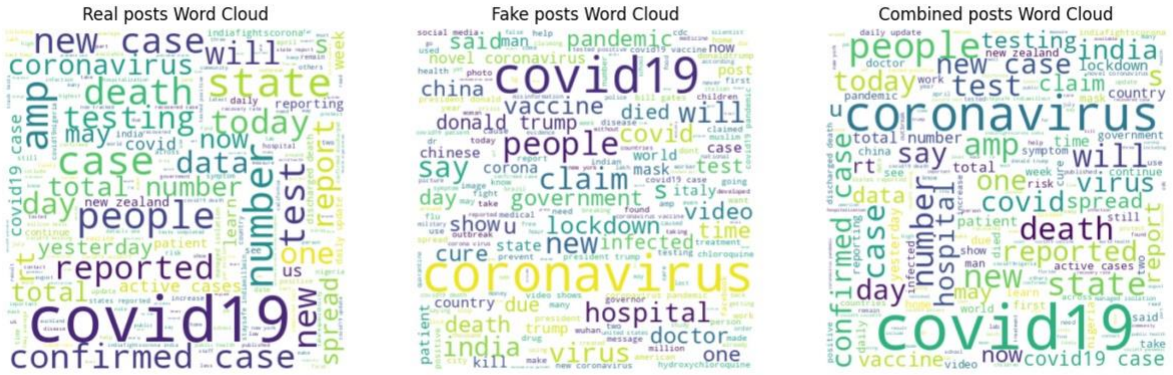
We pre-processed the tweets which included converting to lowercase, removing URLs and user mentions, converting emojis to text, removing punctuation, and tokenization. We didn't remove the Stop Words because tweets may contain words like 'may', and 'might', So it'll become difficult for contextual understanding.

EXPLORATORY DATA ANALYSIS





Word Cloud Visualizations:



Real Posts Word Cloud:

- The most prominent terms are "COVID-19," "case," "new," and "people," suggesting that real posts may focus on reporting new cases and impacts on people.
- Words like "death," "confirmed," "state," and "number" indicate a focus on statistics and data.
- The presence of "testing," "reported," and "data" implies that real posts might contain information related to testing numbers and data reporting.

Fake Posts Word Cloud:

- The word "covid19" is central, but other large terms include "China," "Trump," and "vaccine," which might suggest that fake posts often include conspiracy theories or politically charged content.
- Terms like "lockdown," "government," and "cure" indicate discussions about government measures and potential cures, which are often subjects of misinformation.
- The word "video" stands out, perhaps implying that fake posts may frequently refer to video content, which is a common medium for spreading misinformation.

LDA TOPIC MODELLING

LDA was chosen for its ability to systematically identify and categorize themes in large text datasets without the need for pre-labeled data. This feature is especially helpful for social media content analysis, where the amount and variety of data pose major difficulties. By using LDA, we can distinguish between false information and factual information by gaining a better grasp of the topic structure of tweets about COVID-19. This strategy offers a scalable methodology for real-time social media content monitoring and analysis during existing or future public health emergencies, in addition to being beneficial for immediate analytical needs.

Methodology and Data Preparation:

The tweet corpus was first split into two subsets: "real" and "fake," according to their labels. The TF-IDF vectorization approach was employed to analyze each subset. This method minimizes the influence of often used phrases while efficiently capturing the relevance of words inside the tweets. The algorithm was then able to extract and show the most relevant terms for each subject by applying the LDA model to both subsets with a given number of topics of 10.

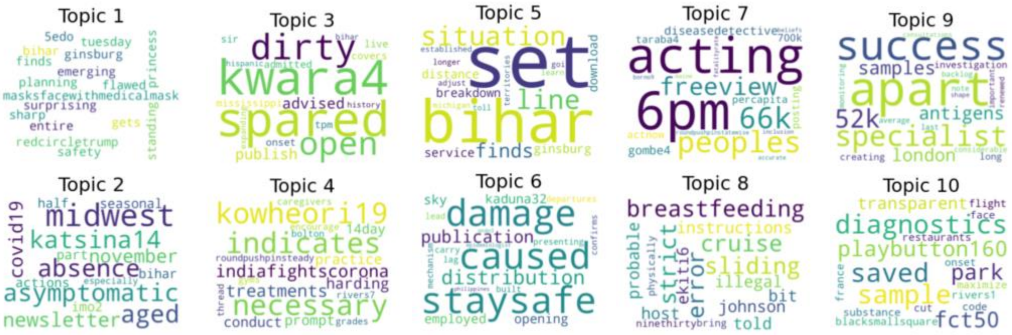
Analysis of Topics in Fake Tweets:

The themes that were taken from the "fake" tweets are diverse. These themes frequently seem fragmented or lack a distinct consistency, which may be a reflection of the various and erratically created character of disinformation. Specific figures, geographical allusions, and medical terminology were mixed in with language that didn't appear to be as pertinent to the situation. This trend points to a trait of disinformation, which is the propensity to blend accurate details with irrelevant or deceptive information to create a message that is less cohesive as a whole.

Analysis of Topics in Real Tweets:

In contrast, more cohesive and contextually relevant subjects were found in the'real' tweets collection. Geographical allusions, healthcare data, and COVID-19 statistics were the main elements of these themes. A consistent and fact-based narrative is often maintained in genuine information distribution, which is exemplified by this attention on factual content and consistency.

Real Tweet Topics Word Cloud:



Insights from Real Tweets Topics:

1. Support and Improvement in Services:

- Pediatric Healthcare Discussions (Topic 0): Enhance communication and resources for pediatric healthcare.

- App Feedback and Technical Queries (Topic 1): Improve tracing apps and provide better technical support.

2. Corporate and Governmental Response:

- Corporate and Organizational Response (Topic 2): Encourage transparency and effective communication from organizations.
- Government Measures and Shortages (Topic 5): Focus on disseminating information about government actions and managing shortages.

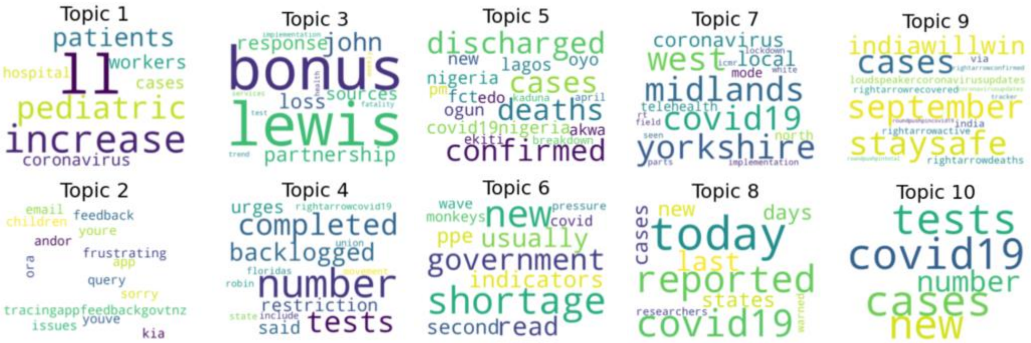
3. Regional and National Information Dissemination:

- Regional COVID-19 Effects (Topic 6): Emphasize region-specific impacts and measures.
- India COVID-19 Awareness and Safety (Topic 8): Boost awareness campaigns in India focusing on safety and preventive measures.

4. Emphasis on Up-to-date Information:

- Recent COVID-19 Updates (Topic 7): Ensure the public receives the latest information and updates on COVID-19.
- General COVID-19 Information (Topic 9): Maintain a stream of general but accurate information about COVID-19.

Fake Tweet Topics Word Cloud:



Insights from Fake Tweets Topics:

1. Need for Fact-Checking and Awareness:

- Mask and Political References (Topic 0): There's a need to debunk myths and clarify facts, especially when political figures are involved.
- Health Guidelines and Errors (Topic 7): Create awareness about accurate health guidelines and address common miscommunications.

2. Regional Focus:

- Midwest Asymptomatic Cases, Regional Health Concerns (Topics 1, 2): Emphasize the importance of localized health data and correct information for these specific regions.

- Health Situation in Bihar (Topic 4): Focus on providing accurate updates and health resources for Bihar.

3. International and Medical Awareness:

- International COVID-19 Response (Topic 3): Share verified information about global COVID-19 responses to prevent misinformation.
- Medical Investigations (Topic 8): Highlight the importance of understanding medical research and the scientific process.

METHODS

All models are used for classifying tweets related to COVID-19. The models are tested on validation and test datasets, focusing on accuracy, precision, recall, and F1 score. Furthermore, an error analysis was performed in order to determine the type of misclassifications.

Support Vector Machine

Support Vector Machine (SVM) model is used for classifying tweets related to COVID-19. Using a linear kernel, the SVM was tested on validation and test datasets, focusing on accuracy, precision, recall, and F1 score. Furthermore, an error analysis was performed in order to determine the type of misclassifications.

Performance Overview:

With an accuracy of 92.68% and an F1 score of 93.06% on the validation set, the SVM demonstrated a high degree of classification accuracy. The model's balanced capacity for prediction was demonstrated by the exceptionally high recall and accuracy. Similar patterns with accuracy and F1 score above 91% were noted in the test dataset, confirming the model's capacity to generalize.

Error Analysis:

The analysis of tweets that were incorrectly identified exposed several trends and the SVM model's shortcomings. When tweets demanded a richer contextual knowledge or used complex wording, the model encountered difficulties. It had trouble, for example, understanding complicated statements and tweets where the context greatly affected the content. Moreover, mistakes were frequently caused by tweets that contained deceptive keywords—words that, while normally connected to a particular class, were employed differently in the context at hand. Additionally, the model appeared to struggle with sarcastic or indirect allusions, indicating a weakness in its capacity to comprehend and analyze complex linguistic patterns.

Example: Tweet: 11 out of 13 people (from the Diamond Princess Cruise ship) who had intially tested negative in tests in Japan were later confirmed to be positive in the United States.

Predicted: 1, Actual: 0

XGBoost

Performance Overview:

The XGBoost model obtained an F1 score of 91.40% and an accuracy of 91.12% on the validation set. These numbers point to a very accurate degree of prediction. In the test dataset, the model's accuracy dropped to 88.60% and its F1 score to 88.98%, indicating a modest decline in performance. Comparing this decline to the model's validation performance indicates a little decline in the model's capacity to generalize to unknown data.

Error Analysis:

Misclassification analysis shows that the XGBoost model encountered issues that were comparable to those of the SVM. It has trouble understanding tweets with complicated sentence patterns, sophisticated wording, and delicate contextual connotations. In tweets where explicit COVID-19-related information or indirect references were necessary for an appropriate categorization, misinterpretations were clearly visible. Additionally, the model appeared to make mistakes in tweets with ambiguous tones or intentions, such sarcasm or hypothetical remarks.

Example: Tweet: Breathlessness excessive fatigue and muscle aches from COVID can last for months. https://t.co/OUhBRirKpE

Predicted: 0, Actual: 1

ADABoost

Performance Overview:

The AdaBoost model obtained an F1 score of 84.77% and an accuracy of 84.27% on the validation dataset. Although these numbers are commendable, they show a somewhat diminished capacity for classification accuracy when juxtaposed with the SVM and XGBoost models. A like pattern could be seen in the test dataset performance, which had an F1 score of 84.12% and an accuracy of 83.74%. The model appears to generalize rather well, but less efficiently than the previously stated models, based on the consistency seen between validation and test results.

Error Analysis:

The AdaBoost model shows difficulties in correctly categorizing tweets with complex language and context through analysis of the misclassified tweets. It has trouble understanding convoluted words, oblique allusions, and nuanced tone or purpose, including sarcasm or hypothetical assertions, much like SVM and XGBoost did. This trend points to a typical restriction in machine learning models' capacity to decipher intricate linguistic subtleties in the absence of sophisticated natural language processing methods.

Example: Tweet: 11 out of 13 people (from the Diamond Princess Cruise ship) who had intially tested negative in tests in Japan were later confirmed to be positive in the United States.

Predicted: 1, Actual: 0

BERT

A pre-trained BERT model (bert-base-uncased) is loaded. This model is configured for sequence classification with two output labels.

Optimizer and Loss Function:

An AdamW optimizer with a learning rate of 5e-5 and epsilon of 1e-8 is used. AdamW is an optimizer with weight decay fix, commonly used for training BERT. The CrossEntropyLoss function is used, suitable for binary classification tasks.

Performance Overview:

After undergoing four epochs of training, the accuracy of the BERT model showed a notable improvement. It started off with an initial training accuracy of 89.48% and by the fourth epoch, it had impressively attained 99.57%. This quick learning curve demonstrates the strong contextual comprehension skills of BERT. The model had a great capacity to generalize, as evidenced by its constant accuracy of around 95.48% throughout all epochs during validation. With a validation F1 score of 95.84%, its effectiveness was further confirmed.

Error Analysis:

Even with its great accuracy, there were some tweets that the model misclassified. These misclassifications all have one thing in common: the nuanced and intricate use of words. BERT encountered difficulties with:

Ambiguous Statements: Tweets that, in order to be accurately categorized, required a better comprehension of the context or outside expertise.

Subtle Misinformation: Tweets that contained information that was subtly misleading or that needed to be verified by fact-checking.

Statements with intricate sentence patterns or subtle vocabulary are examples of complicated linguistic constructs.

Example: Tweet: you can help protect others from covid19 by social distancing amp wearing a cloth face covering that fits snugly and reaches above your nose below your chin and completely covers your mouth and nostrils watch this video featuring

True Label: 1, Predicted Label: 0

However, as seen by its greater accuracy and F1 scores, BERT demonstrated a stronger understanding of context and linguistic subtleties in comparison to more conventional models like SVM, XGBoost, and AdaBoost.

BERT + BILSTM

We combined a pre-trained BERT model with a bidirectional LSTM for text classification. BERT extracts contextual features from text, while the LSTM processes these features, capturing sequential dependencies. A linear layer maps LSTM outputs to a two-dimensional vector for binary classification. Finally, a SoftMax layer transforms these logits into class probabilities, ensuring to classify fake or real labelled posts.

Performance Overview:

After six training epochs, the model showed a tendency toward increased accuracy and decreasing training loss, both of which are signs of successful learning. The model's great capacity to comprehend and categorize the dataset effectively was demonstrated by the highest accuracy of 97.89% attained in the last epoch, with a little training loss.

This achievement was probably made possible by the combination of BERT and BiLSTM, which improved the model's comprehension of the context and word order in tweets:

Contextual Understanding: It is possible that BERT's contextual analysis, which is based on deep learning, assisted in correctly deciphering the meaning of tweets, particularly those with intricate linguistic patterns.

Sequential Data Handling: Understanding the flow and connection of concepts in tweets is crucial for correct categorization, and BiLSTM's proficiency in capturing sequence information in the text likely helped.

Error Analysis:

The BERT + BiLSTM model, despite its advanced capabilities in natural language processing, exhibited certain limitations in accurately classifying COVID-19 related tweets. This analysis focuses on the types of errors and their implications.

Ambiguity and Misinformation: A significant challenge was handling ambiguous content and subtle misinformation. The model occasionally failed to distinguish between factual information and statements that were misleading or lacked scientific backing

Complex Scientific Language: Tweets containing specialized scientific data or referencing specific studies posed a challenge due to their complex terminology and detailed content.

Sarcasm and Hypothetical Scenarios: The model also showed difficulty in interpreting tweets with sarcastic tones or hypothetical scenarios.

Example: Tweet: why mouthwash could help fight covid along with bad breath

True Label: 1, Predicted Label: 0

ETHICAL CONSIDERATIONS

We recognize the value of upholding ethical norms in the course of the COVID-19 Tweet Analysis Project. Several ethical issues are brought up by this study, which involves analyzing social media content—in particular, tweets on the COVID-19 pandemic—which we have made an effort to appropriately address.

Accuracy and Non-Bias: The project was based on objective analysis, with bias minimized via the use of techniques such as LDA. We were open about our methodology, which encouraged replication and peer review—two essential components of maintaining scientific integrity and reliability in our results.

Mitigating Harm: We were well aware of the risk of incorrectly identifying people or interpreting data, which might result in prejudice or stigma. As a result, we avoided negative outcomes by doing our research and reporting with an emphasis on responsible and instructive communication.

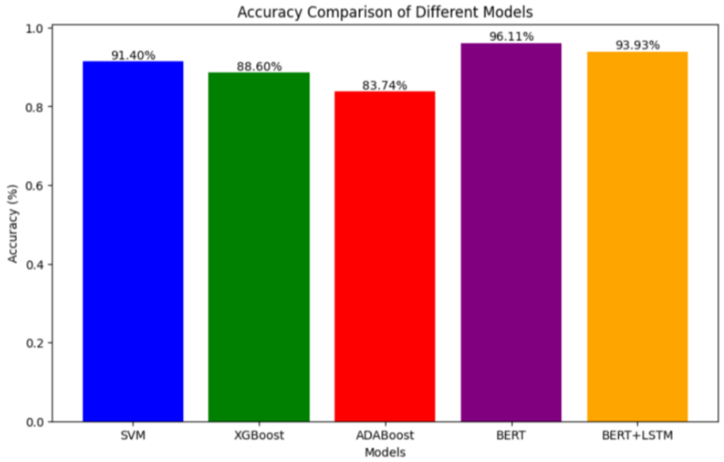
Ethical Use of Data: Regarding data protection and privacy, we followed all applicable laws and regulations during the data gathering and analysis process. Even though the data was made up of tweets that were accessible to the general public, we handled the data with the utmost care to ensure privacy and confidentiality, acknowledging the implied agreement of the public while yet being cautious and respectful.

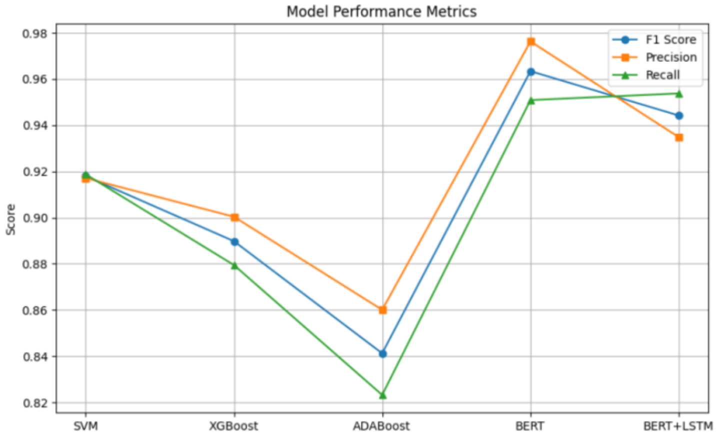
CONCLUSION

Model Accuracy Tabel

Model	Accuracy	F1-Score	Precision	Recall
SVM	91.40	91.80	91.71	91.87
XGBoost	88.60	88.97	90.03	87.95
ADA Boost	83.73	84.12	86.00	82.32
BERT	96.10	96.33	97.62	95.08
BERT+BILSTM	93.92	94.42	93.48	95.37

Performance Metrics Plots





Among the evaluated models, BERT and BERT + BiLSTM stand out as the most effective for classifying COVID-19 related tweets. Their advanced deep learning architectures allow them to understand the context and complexities of human language better than traditional machine learning models. BERT's transformer-based approach excels in contextual understanding, while the addition of BiLSTM in BERT + BiLSTM further enhances its ability to process sequential information, making it slightly more adept for this specific task.

Although they are somewhat successful, the classic machine learning models (SVM, XGBoost, and AdaBoost) do not have the advanced language understanding that deep learning models like BERT do. They work better with datasets when language complexity is not as important.

References

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8608495/
2. https://ieeexplore.ieee.org/document/9803414
3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10197436/
4. https://arxiv.org/ftp/arxiv/papers/2203/2203.09936.pdf
5. https://chat.openai.com/