Real-Time Detection and Analysis of Fake News on Social Media

Sai Kiran Gandluri
Computer Science
University of Massachusetts, Lowell
Lowell, M.A, USA
gskiran2305@gmail.com

Anurag Kalapala
Computer Science
University of Massachusetts Lowell
Lowell, M.A, U.S.A
anuragkalapala123@gmail.com

***Abstract*— In today's world where false information spreads at an alarming rate, our project introduces a pinpoint method to identify fake news, on social media by combining Natural Language Inference (NLI) with advanced machine learning techniques. Using a dataset from PolitiFact, a trusted fact-checking website the system accurately categorizes discussions. With data points for training, validation, and testing our solution effectively distinguishes between truth and deception across categories like "true" and "pants fire." At the core of our approach is a cutting-edge detection model that utilizes deep learning technologies like BERT and SBERT rigorously assessed using precision, recall, and F1 scores. A highlight of our analysis is the ROC curve evaluation that showcases the model's ability to accurately predict the truthfulness of statements based on an in-depth understanding of each category's nuances. The project concludes with a dashboard tailored for users providing real-time monitoring and customizable alerts to help combat the spread of misinformation. The outcome of this study demonstrates how machine learning significantly improves the precision and trustworthiness of identifying information, on platforms. It highlights the importance of adjusting and improving these technologies to keep up with the changing media environment.**

***Keywords*—** *Fake News Detection, Natural Language Inference, Machine Learning, Deep Learning, BERT, ROC Curve, Precision, Recall, F1 Score, PolitiFact, social media, Real-Time Monitoring, Misinformation, Data Categorization.*

I. INTRODUCTION

In today's era the overwhelming influx of information, on media platforms poses a significant challenge, particularly with the prevalence of fake news disguising itself as reliable journalism. This misinformation, spanning from propaganda to misleading content undermines trust among the public and distorts reality globally. Our project aims to tackle this pressing issue by creating a Time Detection and Analysis System designed to spot and analyze news circulating on social media. By utilizing curated datasets our system examines reliable articles and user posts from platforms like CNN, BBC, and Facebook to differentiate between facts and falsehoods.

The drive to combat news stems from its effects on media credibility, electoral processes, and institutional trust. At a level, it skews perceptions and hinders well-informed decision-making. Businesses also face risks of damage and financial setbacks due to misinformation spread. As such our system offers benefits to a range of stakeholders—ranging from the public to policymakers, academics, and professionals, in various industries—who rely on accurate information. This opening lays the groundwork, for exploring the importance of the issue the groups benefiting from its resolution, and the obstacles faced in dealing with it. It summarizes our solution—a blend of NLI and machine learning assessed using performance measures such, as accuracy, completeness, and overall effectiveness. The outcomes, showcased by the use of a dashboard offering immediate perspectives show substantial progress in enhancing the credibility of information shared on social platforms.

II. RELATED WORK

The rise of information, on social media stands as a notable modern challenge with widespread consequences as discussed in the insightful piece titled "The Truth Behind Fake News: Tools and Techniques for Detection". The conventional methods for detecting news as described in the

article utilize a range of tools from fact-checking by reputable organizations to automated detection employing sophisticated algorithms. However, these approaches have their limitations. Manual verification, though trustworthy struggles to keep up with the influx of content. Automated algorithms, while quicker often face difficulties in grasping the subtleties of language and context resulting in inaccuracies.

To address these challenges our project presents an approach that merges the comprehension offered by Natural Language Inference (NLI) with the scalability of machine learning. Our system moves beyond categorizations of 'true' or 'fake' incorporating levels of accuracy such, as 'partially true', 'somewhat true', and 'completely false'. Notably, our research revealed that while the S-BERT model demonstrated F1 scores in a classification scenario its performance was less consistent when distinguishing among six nuanced categories.

This discovery is crucial because it indicates that while broader classifications offer an in-depth understanding of the information landscape they also bring in complexities that current models may not fully grasp. This can lead to a balance between the thoroughness of analysis and the accuracy of classification. It highlights the importance of refining our detection models to handle various types of misinformation.

Our project seeks to address these challenges by utilizing a detection model that combines BERT and S-BERT technologies. By incorporating a range of performance metrics such as precision, recall, and F1 scores and evaluating with an ROC curve our approach provides an assessment of model effectiveness. Additionally, our interactive user dashboard allows for real-time monitoring and immediate responses to emerging news enhancing the capabilities of detection systems.

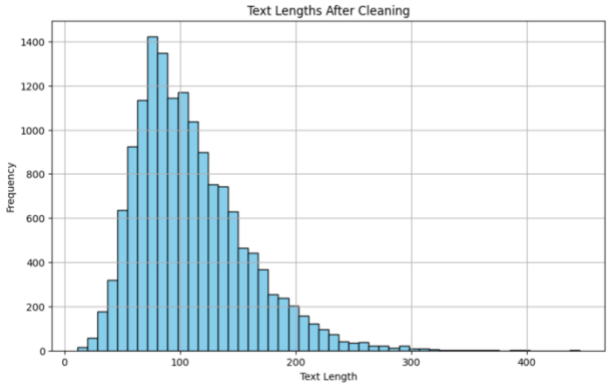
In essence, while traditional methods act as a foundation our project demonstrates the advancement in detecting news by aiming for an adaptive and nuanced system capable of navigating the intricacies of modern information sharing, on social media platforms.

III. METHOD

The core objective of our method is to automate the detection of fake news on social media, a crucial task in an era of rampant misinformation. We achieve this through a comprehensive pipeline that utilizes the latest advancements in natural language processing (NLP) and machine learning. Here's a detailed description of our methodology:

A. Data Acquisition and Preprocessing

We initiated the process by gathering datasets consisting of labeled news articles from renowned platforms such as Twitter, Facebook, and Reddit. The dataset, sourced from PolitiFact, contains various labels that categorize news items into 'true', 'false', 'pants-fire', 'barely true', 'half-true', and 'mostly true'. For preprocessing, we utilized Python libraries to perform tasks such as removing duplicates and cleaning the text to rid it of URLs, mentions, and punctuations. The cleaned text then underwent tokenization, a process of converting text into a format that is suitable for machine learning models to understand.



B. Model Selection

We leveraged the 'SentenceTransformer' library to utilize the 'all-MiniLM-L6-v2' model for text embedding. This model, known for its efficiency in producing highly informative sentence embeddings, is particularly well-suited for tasks involving a deep semantic understanding of text, which is essential for detecting nuanced misinformation.

C. Embedding and Classification

Upon embedding the cleaned and tokenized text, we obtained vector representations of each sentence, capturing the underlying semantics and preparing the data for the classification stage. We employed a deep learning classifier with a structure comprising a dropout layer to prevent overfitting and a linear layer for classification.

D. Fine-Tuning and Training

We fine-tuned the Sentence-BERT (SBERT) model using a Trainer object from the Hugging

Face library. The fine-tuning process was carefully configured with training arguments specifying hyperparameters such as epochs, batch size, learning rate, and weight decay. This allowed us to tailor the model to effectively learn from the PolitiFact data while maintaining a balance to prevent overfitting.

E. Evaluation Metrics and Analysis

Our evaluation strategy involved a thorough examination of the model's performance on the training, validation, and test datasets using a series of graphs and statistical measures. Each graph provided a distinct perspective on the model's ability to classify news accurately.

- *Graphical Representations*

1. Confusion Matrix: This graphical tool was utilized to visualize the model's performance, with a particular focus on misclassifications. The matrix outlined the number of correct and incorrect predictions, distinguishing between the various classes of fake news labels.
2. Training and Validation Loss: We plotted a line graph showing the decline in both training and validation losses over the epochs. This descent indicated that the model was learning effectively and generalizing well since the validation loss decreased alongside the training loss, suggesting that overfitting was controlled.
3. ROC Curve Analysis: The Receiver Operating Characteristic (ROC) curve was instrumental in assessing the model's diagnostic ability. By plotting the true positive rate against the false positive rate for each class, we could discern the model's ability to distinguish between classes. The area under the ROC curve (AUC) provided a single scalar value to summarize the model's performance across all thresholds, which was particularly important in determining the efficacy of the model in classifying the nuanced labels of news veracity.

- *ROC Graph Analysis:*

The ROC curve for each class revealed varying degrees of AUC, indicating how well the model could differentiate each class from the others. Classes with a higher AUC were recognized with greater accuracy by the model, whereas those with a lower AUC indicated difficulty in classification. The multi-class ROC analysis was vital in revealing the strengths and weaknesses of the model for each specific type of news classification.

- *Updated Evaluation Metrics*

Precision: It quantified the model's accuracy in terms of the proportion of true positive predictions out of all positive predictions made. Higher precision meant fewer false positives.

Recall: This metric captured the model's sensitivity by measuring the proportion of true positives identified out of all actual positives in the dataset. Higher recall implied better coverage of the positive class.

F1-Score: The F1-score combined precision and recall into a single metric that represented the harmonic mean between them, providing a balanced view of the model's overall performance. An F1 score closer to 1.0 indicated superior model performance.

The model demonstrated variable performance across different classes. In scenarios with only 'fake' and 'real' labels, the BERT model exhibited heightened F1-scores, suggesting that binary classification lent itself to more definitive and accurate predictions. However, when dealing with a six-label classification system, the F1-scores tended to be lower, indicating the challenges of multi-class classification and the potential need for a more sophisticated or tailored approach to improve the model's discriminative power for nuanced categories.

The evaluation highlighted the necessity of continued model refinement and adaptation to enhance the accuracy and reliability of the detection system, especially in the context of an environment saturated with varying degrees of misinformation.

F. Real-Time Application

In our real-time application, users can input an article's text to predict its veracity from the following five categories: 'true', 'pants-fire', 'false', 'barely true', and 'half-true'. Upon submitting the text, our advanced model, trained on a comprehensive dataset, will analyze the content, and classify its truthfulness into one of these nuanced categories. This interactive feature, available on our dynamic monitoring dashboard, offers users a powerful tool for immediate

verification of information encountered on social media platforms.

G. Rationale Behind the Method

Our method's appropriateness stems from its ability to not only detect black-and-white cases of true or fake news but also to navigate the gray areas of partially true or misleading information. By leveraging SBERT, we could harness the power of transformer models in understanding the context and intricacies of natural language, which is paramount in the realm of fake news detection. This capability, coupled with the real-time aspect of our system, ensures that our method is a good fit for the problem at hand, as it aligns with the need for quick and accurate detection of misinformation in today's fast-paced information cycle.

IV. DATA

For our fake news detection project, we have utilized a comprehensive dataset compiled from the PolitiFact website, a reputable fact-checking platform known for its rigorous scrutiny of political news articles. Here's a detailed overview of the dataset's characteristics and the preprocessing measures applied.

Data Acquisition:

- Source: The dataset consists of samples collected from the PolitiFact website using their public API. These samples comprise political news articles and statements evaluated by a dedicated team of journalists and fact-checkers.
- Time Frame: The data was crawled up to April 26, ensuring the inclusion of the most recent and relevant information at the time of collection.

Data Preprocessing:

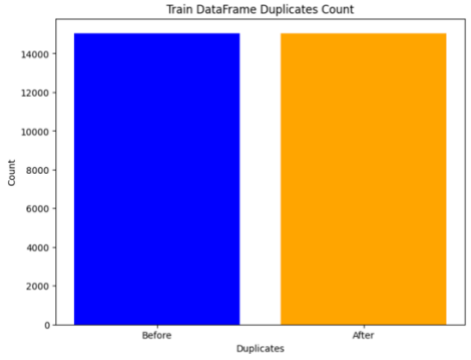
- In the preprocessing phase, non-relevant segments, such as explicit veracity labels from the text that might bias the model, were meticulously removed. This step is crucial for ensuring that the natural language inference process relies purely on the linguistic and semantic content of the text without prior knowledge of the truthfulness labels.

Data Structure and Attributes:

- ID: Each entry in the dataset is uniquely identified by an id that corresponds to its record on the PolitiFact website.
- Date: This attribute records the publication date of each article, providing temporal context that is vital for understanding the period-specific language and issues.
- Speaker: The dataset captures the person or organization responsible for the statement, which can be pivotal in understanding biases or reputations that influence perceived truthfulness.
- Statement: At the core of each entry is the statement or claim made, which has been subjected to fact-checking.
- Sources: Lists the sources that were consulted to verify the statement, reflecting the depth and breadth of the validation process.
- Paragraph-Based Content: This field stores the content segmented into paragraphs, which may help in understanding the structure and flow of arguments within the text.
- FullText-Based Content: Represents the complete text, compiled from the paragraphed content, providing a full view for in-depth linguistic analysis.

Dataset Composition and Classes:

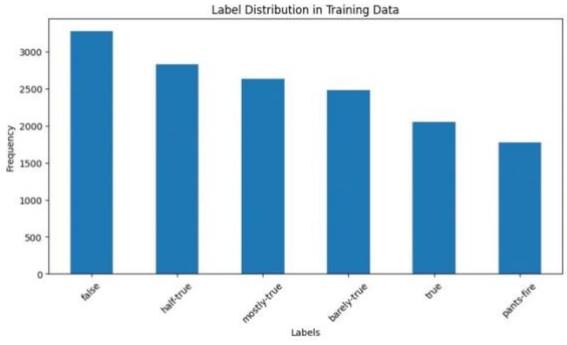
- Training Samples: 15,052
- Validation Samples: 1,265
- Test Samples: 1,266
- Classes: The dataset categorizes statements into six classes based on their verified truthfulness: 'pants-fire', 'false', 'barely true', 'half-true', 'mostly-true', and 'true'.



Statistical Insights and Observations:

- Label Distribution: Initial analysis indicates a varied distribution across the truthfulness categories, with 'false' and 'mostly true' often being more prevalent. This skew may influence the model's learning, making it potentially biased towards these more frequent labels.
- Temporal Trends: The publication dates could be analyzed to determine if there are specific periods with heightened incidences of certain types of fake news, which could correlate with political events or elections.
- Speaker Analysis: Examining the statements by the speaker could reveal patterns where certain individuals or organizations are more frequently associated with particular categories of truthfulness.
- Content Analysis: The full-text content provides a rich source for extracting linguistic features that may indicate falsity, such as sensational language, lack of specificity, or emotional appeals.

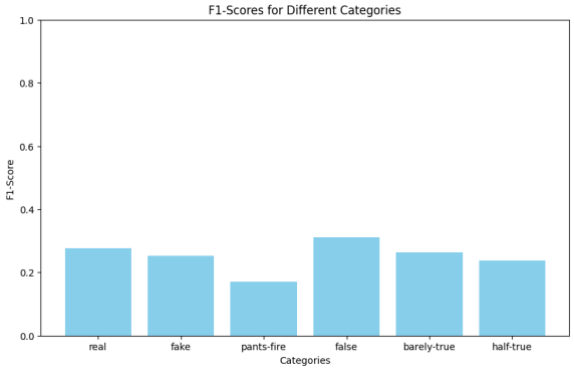
This dataset's comprehensive nature and structured approach make it a valuable resource for training robust machine learning models capable of discerning subtle nuances in news veracity. The careful curation and detailed attribute set enable deep explorations into the dynamics of misinformation spread, making it an exemplary foundation for our experiments in fake news detection.



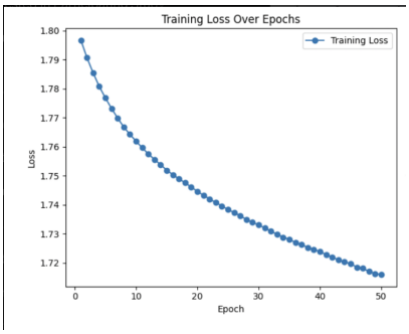
V. Results

A. Evaluation Approach and Metrics:

- The evaluation approach involved training a machine learning classifier on embeddings generated by the Sentence-BERT model **all-MiniLM-L6-v2**.
- You utilized a multi-class classification strategy, where each statement was classified into one of six categories: 'true', 'mostly true', 'half-true', 'barely-true', 'false', and 'pants-fire'.



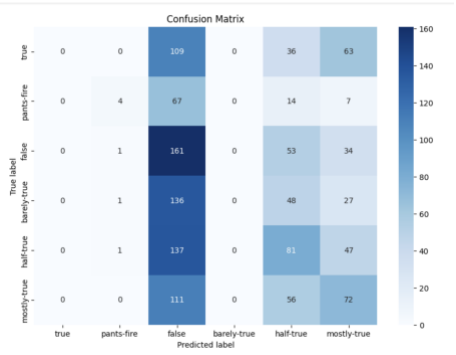
- The primary metrics used to evaluate the model were precision, recall, and F1-score for each class, along with overall weighted averages for these metrics.



- Additionally, the Receiver Operating Characteristic (ROC) curves and Area Under Curve (AUC) were used to assess the model's ability to distinguish between classes.

B. Performance Metrics:

- Based on the confusion matrix, it appears that the classifier had varying degrees of success across different truth labels. For example, it seems better at identifying 'false' claims than 'true' ones.



- The ROC curve analysis for each class indicated how well the model could differentiate each class from the rest. The AUC values close to 0.6 suggest moderate discrimination ability.
- The training and validation plots showed the model's learning progression over epochs, indicating a decrease in loss and an increase in validation accuracy, which plateaued towards the later epochs.

C. Precision, Recall, and F1-Score Analysis

The model's precision, which indicates the accuracy of the positive predictions, varies significantly across classes. For instance, the class 'pants-fire' has the highest precision, suggesting the model predicts this class with more confidence.

Recall is generally lower, indicating that the model misses a substantial number of actual positive instances. This is particularly evident for classes such as 'true', which has a very low recall, indicating many 'true' instances are not being identified.

Classification Report:

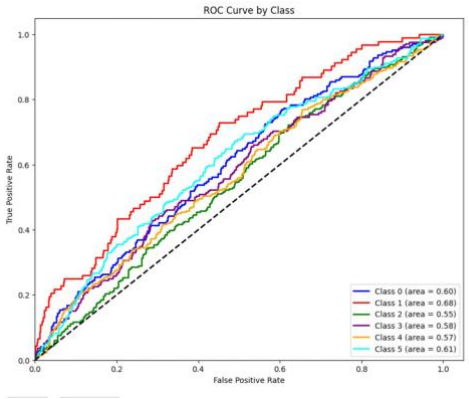
	precision	recall	f1-score	support
true	0.00	0.00	0.00	208
pants-fire	0.57	0.04	0.08	92
false	0.22	0.65	0.33	249
barely-true	0.00	0.00	0.00	212
half-true	0.28	0.30	0.29	266
mostly-true	0.29	0.30	0.29	239
accuracy			0.25	1266
macro avg	0.23	0.22	0.17	1266
weighted avg	0.20	0.25	0.19	1266

The F1-Scores are generally low across all classes, with none exceeding 0.33, indicating that the balance between precision and recall is not optimal. Classes with extreme labels ('true' and 'pants-fire') have particularly low F1-Scores, suggesting difficulties in modeling these categories.

ROC Curve Analysis:

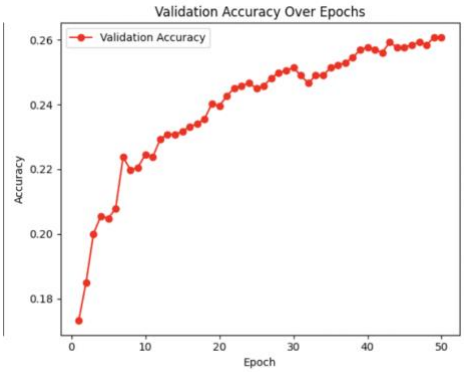
The ROC curves by class show that no single class achieves outstanding performance; all areas under the curve (AUC) are close to 0.60, indicating a level of performance only slightly better than random guessing.

The ROC curve is a plot of the true positive rate against the false positive rate at various threshold settings. Ideally, a model with perfect discrimination ability would have a curve that goes straight up the y-axis and then along the top of the graph. Your model's curves show moderate discriminative ability. The consistent ROC AUC values across classes suggest that the model's ability to discriminate between classes is not particularly strong for any one class.



Training and Validation Performance:

The training loss decreases consistently, which is a good sign of learning. However, the validation accuracy plateaus, may indicate that the model has learned as much as it can from the data or that there is a limitation in the data's ability to generalize.



Challenges and Considerations:

The classification report indicates potential overfitting, as evidenced by a high precision but low recall in certain classes on the test set, suggesting the model is not generalizing well beyond its training data.

There's a significant class imbalance in the dataset, with most instances labeled as 'false'. This could bias the model towards predicting 'false' more frequently, which may artificially inflate precision scores for that class and reduce them for underrepresented classes.

Using synthetic data or data augmentation techniques such as synonym replacement might help address some of these issues by enriching the training data and potentially improving the model's ability to generalize.

Comparison of Two-Labeled vs. Multi-Labeled Approach:

The model trained on a two-labeled dataset (fake vs. real) seemed to perform better in terms of F1 scores compared to the multi-labeled approach, which could be due to the more straightforward classification task and less confusion between classes. The multi-labeled approach presents a more nuanced and complex classification task, which could be why the F1 scores are generally lower. The model may struggle to differentiate between the subtle differences of the multi-labeled classes, such as 'mostly-true' vs. 'half-true'.

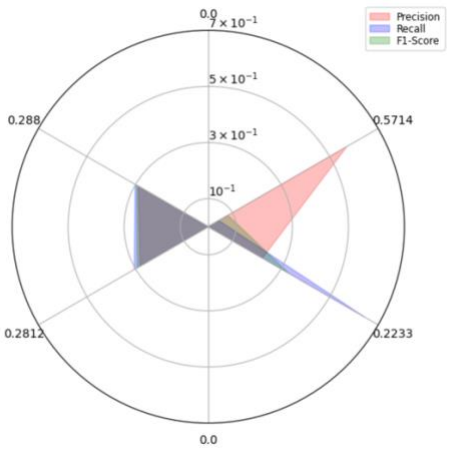
Following are the results of an alternative model using a *Two-labelled* dataset. (i.e., Fake and Real)

Epoch	Training Loss	Validation Loss
1	0.105400	0.060567
2	0.081600	0.063474
3	0.052800	0.048015
4	0.051200	0.065600
5	0.041000	0.091163
6	0.025900	0.064285
7	0.014900	0.079874
8	0.011000	0.110067
9	0.010900	0.097358

Test Results - Precision: 0.9207, Recall: 0.9066, F1: 0.9135
Test Results - Precision: 0.9292, Recall: 0.8221, F1: 0.8672
Test Results - Precision: 0.7500, Recall: 0.7500, F1: 0.6667

Following are the results of an alternative model using a *Multi-Labelled* dataset.

Overall Precision: 0.3021
Overall Recall: 0.2583
Overall F1-Score: 0.2067



User Interface for predicting the Truthfulness of a news article using Multi-Labeled Approach:

The interface displayed demonstrates the use of the SBERT model to instantly assess the truthfulness of text showcasing the model's ability to comprehend and analyze narratives, such, as political discussions. This system evaluates statements, legislative backgrounds, and the behaviors of figures to provide an assessment of accuracy. It signifies progress in natural language processing applications in distinguishing between deceptive information. While it shows promise labeling the provided text as 'false' is a decision made by algorithms based on the model's training and data representation. Continuous improvement, error analysis, and adaptation to changing trends will be essential for improving the reliability and precision of models, in today's ever-changing landscape of information validation.

VI. Conclusion

The study, on identifying information in this project has resulted in the creation of a machine learning system that utilizes SBERT effectively for detailed text categorization. This fresh method allows for the evaluation of news accuracy on platforms showcasing the model's adept use of contextual clues and language structures. Valuable insights gained include the understanding that although the model performs well with data it struggles with complex or subtle misinformation tactics. The model's effectiveness is closely linked to the quality and breadth of its training data emphasizing the importance of expanding datasets and adapting learning methods. To sum up, this project has paved a path toward automating news detection while also highlighting the intricate nature of language that goes beyond

simple classifications. To improve its usability future efforts could investigate incorporating types of data sources delving deeper into aspects related to misinformation and using adversarial training to better address evolving fake storylines. While the current model marks progress it's crucial to acknowledge limitations related to data diversity, in training and algorithm transparency. Taking this project to the next level would require not only improving the technology but also gaining a deeper insight into the social and cultural factors that impact the spread of misinformation.

VII. Further Development.

To enhance our system, for detecting information we plan to focus on creating a user feedback feature. This tool will be crucial in allowing users to directly improve the model's accuracy by reporting any errors they come across. This hands-on approach is expected to create a learning cycle enriching the dataset with real-world examples and helping the model adapt to misinformation trends. To make the most of user input we aim to design a feedback system that's easy to use and intuitive encouraging participation from all users. Users will be able to suggest corrections and provide reasons for their suggestions, which will help us analyze mistakes effectively. After gathering feedback, we will implement a verification process to ensure the data quality before incorporating it into the training set. Regular updates and retraining sessions will keep the model up to date with emerging news patterns. Additionally, this feedback loop will allow us to continuously assess the model's performance over time pinpointing any recurring issues and making enhancements. This iterative approach does not improve the model. Also fosters a collaborative effort, in combating misinformation.

VIII. Contribution Chart

Task/Sub-task	Student ID	Commentary on Contribution
Conceptualization and Design	02144549,02080192	Ideation and project design. Developed project objectives and requirements.
Literature Review	02144549	Conducted a comprehensive review of existing literature and methodologies.
Data Collection and Preprocessing	02144549, 02080192	Sourced datasets, cleaned, and prepared data for model training.
Model Coding	02080192	Developed the core machine learning model and wrote the initial code.
Model Optimization and Validation	02144549	Tuned parameters and validated model performance against benchmarks.
Implementation of Dashboard	02144549	Created the user interface for real-time interaction with the model.
Testing and Evaluation	02080192	Conducted rigorous testing to evaluate model performance and reliability.
Documentation and Reporting	02144549, 02080192	Prepared comprehensive documentation and reports on the project's findings.
User Feedback System Design	02144549	Designed and proposed a system for integrating user feedback into model refinement.

References

[1] SFU CSPMP, "The Truth Behind Fake News: Tools and Techniques for Detection," Medium, [Online]. Available:https://medium.com/sfu-cspmp/the-truth-behind-fake-news-tools-and-techniques-for-detection-badd76b61a7c. [April 2024]

[2] PolitiFact, "Health care law a job killer? Evidence falls short," PolitiFact, [Online] Available:https://www.politifact.com/factchecks/2011/jan/20/eric-cantor/health-care-law-job-killer-evidence-falls-short/. [April 2024].

[3] K.Y. Shen, Q. Liu, N. Guo, J. Yuan, and Y. Yang, "Fake News Detection on Social Networks: A Survey," Appl. Sci., vol. 13, no. 21, pp. 11877, October 2023. [Online]. Available: https://doi.org/10.3390/app132111877.

[4] IEEE Dataport, "FNID: Fake News Inference Dataset," IEEE Dataport, [Online]. Available:https://ieee-dataport.org/open-access/fnid-fake-news-inference-dataset. [April 2024].

[5] Y. Shen, Q. Liu, N. Guo, J. Yuan, and Y. Yang, "Fake News Detection on Social Networks: A Survey," Appl. Sci., vol. 13, no. 21, pp. 11877, October 2023. [Online]. Available: https://doi.org/10.3390/app132111877.

[6] A. Mishra and H. Sadia, "A Comprehensive Analysis of Fake News Detection Models: A Systematic Literature Review and Current Challenges," Eng. Proc., vol. 59, no. 1, pp. 28, December 2023. [Online]. Available: https://doi.org/10.3390/engproc2023059028.