Human vs. Machine Learning Insights: Fake News Detection

Medha Trisal, Rahul Sai Krishna Guduri
The Pennsylvania State University
The College of Information Sciences and Technology
mmt5877@psu.edu, rxg5516@psu.edu

Abstract — As was seen during the COVID-19 pandemic, fake news on social media poses significant challenges to information integrity and public trust. This study leverages a pretrained BERT model, fine-tuned with a custom neural network architecture, to classify news tweets as real or fake. Linguistic features such as sentiment, post length, and readability were examined to assess their impact on both human interpretability and model performance. Hypothesis testing revealed statistical significance between these features and news type, while an ablation study demonstrated their influence on enhancing predictive performance. By bridging human-recognizable features with advanced NLP techniques, this research addresses gaps in understanding the role of these features in fake news detection.

Keywords: COVID-19, Fake news, BERT, NLP

Ⅰ. INTRODUCTION

The abundance of fake news on social media has driven widespread misinformation, with significant consequences for public perception, policy-making, and societal behavior. This challenge became particularly acute during the COVID-19 pandemic, where the rapid spread of misinformation undermined public health efforts, fueled vaccine hesitancy, and intensified societal polarization. The pandemic serves as a pivotal case study due to the wealth of data it provides on the problems of fake news and its impact during a global crisis.

While progress has been made in developing models for fake news detection, less attention has been given to understanding the intrinsic characteristics of fake news, especially during a high-stakes event like a pandemic. This study examines sentiment, post length, and readability — features associated with human interpretation — to uncover patterns that distinguish fake news from real news and to determine their utility in improving machine learning models.

Two key research questions guide this work:

1. During the COVID-19 pandemic, did sentiment, post length, and readability aid in distinguishing fake news from real news?
2. When integrated into a transformer-based BERT model, do human-recognizable features enhance classification performance for fake news detection?

Ⅱ. LITERATURE REVIEW

A. Transformer-Based Models for Fake News Detection

Transformer architectures, such as BERT, have emerged as powerful tools for fake news classification. Gupta et al. [2] and Khan et al. [1] explore the use of individual and ensemble transformer models, demonstrating their high effectiveness in detecting misinformation, whether used individually or as part of ensembles. However, while ensembles offer performance improvements, they are computationally expensive and less practical for many applications. Additionally, no one kind of transformer model greatly outperformed another. Addressing these conclusions, our study employs a single BERT model, striking a balance between effectiveness and computational efficiency.

Khan et al. also provide the dataset used in this study, consisting of 8,560 labeled tweets, split into 6,420 entries for training and 2,140 entries for testing. Each entry includes the raw text of the tweet and its corresponding label as "real" or "fake." Designed specifically for robust modeling, this dataset features a diverse collection of COVID-19-related tweets. It is benchmarked with four machine learning baselines: Decision Tree, Logistic Regression, Gradient Boost, and Support Vector Machine (SVM). The best performance was achieved by the SVM model, with an F1-score of 93.46%, highlighting the dataset's suitability for evaluating fake news detection methods across a variety of modeling approaches.

B. The Role of Social Media in COVID-19 Misinformation

Social media platforms have been central to the rapid dissemination of both information and misinformation. The study by Vosoughi et al. [3] underscores how social media exacerbates the spread of fake news, which in turn influences public attitudes and behaviors during the pandemic. The findings highlight a cycle where misinformation negatively impacts public trust and compliance with health measures, emphasizing the need for robust detection mechanisms at the platform level.

C. Mental Health Decline

Garcia et al. [4] and Rocha et al. [5] found a causal relationship between the spread of fake news and mental health decline. During the pandemic, the misinformation encountered by social media users contributed to increased levels of depression, anxiety, and other mental health conditions, exemplifying the profound impact of fake news on societal well-being.

Ⅲ. METHODOLOGY AND RESULTS

The methodology combines statistical analysis and predictive modeling to investigate the relevance of features like sentiment and post length in distinguishing fake news from real news. The dataset was preprocessed for a BERT-based classifier, with additional features engineered: sentiment and post length. Statistical tests, including a Chi-square test for sentiment and T-tests for post length, were conducted to determine their significance in differentiating news types. A BERT-based model was trained to classify fake and real news, with sentiment and post length integrated as additional features. Performance was evaluated using accuracy, precision, recall, and F1-score. An ablation study examined the impact of adding and removing these features on model performance, assessing their unique contribution. Finally, the results were compared to explore differences in feature relevance for human interpretability versus machine learning predictions.

A. Data Preparation and Preprocessing

These training and testing datasets included raw text (“tweet”) and corresponding labels (“real” or “fake”). The datasets were loaded into Pandas DataFrames, and preprocessing was applied to standardize and clean the text data. Preprocessing steps removed special characters, hashtags, and mentions, converted text to lowercase, and ensured the removal of excess whitespace. This standardization ensured that input data was consistent and devoid of noise that could hinder model performance. To enable numerical processing by the model, labels were encoded into binary numerical representations (`real` as 1 and `fake` as 0) using `LabelEncoder`. This allowed the classifier to process the target variable effectively.

B. Train-Validation Split

The training data was further split into training (96%) and validation subsets (4%) to internally validate the model during training. Stratified sampling was used to ensure that the class distribution in the split datasets reflected that of the original data, preventing class imbalance from biasing the model.

C. Tokenization and Data Transformation

The text data was tokenized using the pretrained tokenizer BertTokenizerFast from the Hugging Face library. The tokenizer transformed text into sequences of token IDs, with special tokens added for sentence boundaries. Padding and truncation ensured that all sequences conformed to the maximum length of 200 tokens. This tokenization prepared the text for input into the BERT model, which requires a specific input structure.

The tokenized data was split into input IDs, attention masks, and corresponding labels, then converted into PyTorch tensors. This transformation facilitated efficient data batching and device transfer for training and inference.

D. Model Architecture

The classification model was built on top of the bert-base-uncased pretrained BERT architecture. BERT was used as a frozen feature extractor, leveraging its ability to encode contextual word embeddings from large-scale text corpora. Freezing BERT's parameters reduced computational overhead and ensured stability during training. A custom neural network was added on top of BERT's pooled output:

- A fully connected layer of reduced dimensionality from 768 (BERT’s pooled output size) to 512 neurons.
- A ReLU activation introduced non-linearity.
- A dropout layer (rate of 0.1) mitigated overfitting.
- A second fully connected layer mapped the features to two output classes.
- A LogSoftmax activation provided probabilities for each class.

This architecture was designed to balance expressiveness and computational efficiency while capitalizing on BERT's pretrained capabilities.

E. Training and Validation Phases

The training process was optimized using the AdamW optimizer with a learning rate of 1e-5. A weighted negative log-likelihood loss function accounted for class imbalances by assigning higher penalties to misclassified instances of the minority class. Class weights were computed based on the training data distribution.

Training was conducted over 15 epochs, with each epoch iterating through the training data in batches of 32. Gradients were clipped to a maximum norm of 1.0 to stabilize updates during backpropagation. After each epoch, the model was evaluated on the validation dataset to compute the validation loss, providing an internal measure of generalization performance.

The model's weights were saved whenever validation loss improved, ensuring that the best-performing model was preserved. This process highlighted the model's gradual improvement: training loss decreased from 0.673 to 0.413, and validation loss decreased from 0.632 to 0.384 across epochs. The validation data, though seen during model development, was distinct from the training data and allowed for iterative refinement.

F. Testing Phase

To assess the model's performance on unseen data, a separate test dataset was used. This dataset, withheld during both training and validation phases, provided an unbiased evaluation of the model's generalization ability. The saved best-performing model was loaded, and predictions were made on the test set. The resulting confusion matrix revealed high accuracy and balanced precision, recall, and F1-scores across the two classes, indicating robust performance. See Fig. 1.

	precision	recall	f1-score	support
0	0.80	0.88	0.84	1020
1	0.88	0.80	0.84	1120
accuracy			0.84	2140
macro avg	0.84	0.84	0.84	2140
weighted avg	0.84	0.84	0.84	2140

Fig. 1. Confusion Matrix. Class 0 - Real News, Class 1 - Fake News. Please note that due to sampling, these values are subject to change when code is re-ran. However, they will remain approximately the same.

G. Hypothesis Testing

To investigate additional patterns in the data, sentiment analysis, tweet length, and readability scores were computed for each tweet in the test set:

- Sentiment analysis was performed using TextBlob, classifying tweets as positive, neutral, or negative.
- Tweet length was calculated as the number of words.
- Readability was assessed using the Flesch Reading Ease score via textstat.

Statistical hypothesis tests were performed to investigate the relationships between sentiment, tweet length, readability, and news type. See Fig. 2.

```
Chi-square Test for Sentiment and News Type
Chi-square statistic: 179.5984866174071, p-value: 1.0015765595295493e-39

T-Test for Length and News Type
T-statistic: 18.08898692931097, p-value: 7.251436124138891e-68

T-Test for Readability and News Type
T-statistic: -1.1734112312163167, p-value: 0.24076164290346008
```

Fig. 2 Hypothesis Test Results. Please note that due to sampling, these values are subject to change when code is re-ran. However, they will remain approximately the same.

With a significance level of 5% ($\alpha = 0.05$), the chi-square test for independence evaluated the null hypothesis that sentiment and news type are independent. The results ($p < 0.05$) strongly rejected the null hypothesis, indicating a significant relation between sentiment and news type. The independent t-tests compared tweet length and readability between real and fake news. For tweet length, the null hypothesis — that the mean lengths of real and fake news tweets are equal— was rejected, revealing a significant difference. In contrast, for readability scores, the null hypothesis could not be rejected, suggesting no significant difference in readability between real and fake news.

H. Ablation Study

This part of the methodology extends step one by incorporating statistically significant features—sentiment and tweet length—into the model to evaluate their impact on classification performance. These features were computed for all datasets as before, and the pretrained bert-base-uncased model was augmented to include them as input features. The extended input was processed through two fully connected layers, with ReLU activation, dropout, and LogSoftmax for final class probabilities.Training, validation, and testing were conducted the same as with the original model. The new confusion matrix is shown in Fig. 3.

	precision	recall	f1-score	support
0	0.79	0.86	0.83	1020
1	0.86	0.79	0.83	1120
accuracy			0.83	2140
macro avg	0.83	0.83	0.83	2140
weighted avg	0.83	0.83	0.83	2140

Fig. 3 Confusion Matrix. Class 0 - Real News, Class 1 - Fake News. Please note that due to sampling, these values are subject to change when code is re-ran. However, they will remain approximately the same.

Ⅳ. MODEL PERFORMANCE RESULTS

The performance results of the initial BERT model and the BERT model with integrated linguistic features provide valuable insights. This section focuses on analyzing the models' classification performance, evaluating the role of the added features.

A. Performance of the Initial BERT Model

The initial BERT model demonstrated strong classification performance, achieving an overall accuracy of 84% and a balanced F1-score of 0.84. Precision and recall were evenly distributed across both classes, with precision for fake news reaching 0.88 and recall for real news at 0.88. These results affirm the capability of BERT's pretrained architecture to effectively capture complex linguistic and contextual patterns inherent in fake and real news. Moreover, the relatively high recall for real news suggests that the model is particularly adept at identifying authentic information. Conversely, its strong precision for fake news highlights its ability to minimize false positives, ensuring that when fake news is flagged, it is likely to be correct. These balanced metrics are crucial for applications in misinformation detection, where both under-classifying fake news and over-classifying real news can have significant consequences.

B. Impact of Integrating Linguistic Features

The inclusion of these features did not improve the model's performance; in fact, the overall accuracy and F1-score decreased slightly to 83% and 0.83, respectively. Precision and recall for both classes remained similar, with slight decreases across the board. This outcome underscores the following insights:

1. Feature Redundancy in Transformer Models
    - BERT's architecture is highly adept at extracting nuanced linguistic and contextual information directly from the text, including patterns related to sentiment and length. As a result, the additional engineered features may have introduced redundancy rather than providing new information, contributing to the marginal decline in performance.

2. Potential for Overfitting
    - Adding features such as sentiment and post length increased the model's complexity, potentially leading to overfitting. While these features are relevant for human interpretation, they may have reduced the model's ability to generalize to unseen data.

3. Challenges in Feature Integration
    - Transformer models are optimized to learn directly from raw text data. The integration of external features like sentiment and post length requires careful consideration to ensure that the added information complements, rather than overlaps with, the model's existing capabilities.

C. Comparison Between Models

The comparison between the initial BERT model and the model with integrated features highlights a critical distinction between machine-driven and human-interpretable approaches to fake news detection. While linguistic features like sentiment and post length are important for human analysis and provide explanatory power, they may not enhance the predictive capacity of advanced transformer models like BERT. Furthermore, the initial model's performance demonstrates that pretrained transformers are capable of identifying the subtle patterns in fake news text without requiring additional feature engineering. The slight decline in the second model's performance reinforces the need to carefully evaluate the integration of additional features.

V. DISCUSSION: USING INSIGHTS TO ADDRESS FAKE NEWS ON SOCIAL MEDIA

The findings of this study, alongside evidence from prior research, provide valuable insights into strategies that social media companies, policymakers, and other stakeholders can adopt to combat the spread of fake news. By combining human-interpretable features with transformer-based models like BERT, this paper highlights a dual approach to understanding and preventing misinformation. This discussion outlines actionable strategies that leverage these insights to create effective detection systems, inform user engagement policies, and mitigate the societal impact of fake news.

A. Using Human-Interpretable Features for Warnings

One key takeaway from this study is that sentiment and post length are markers that can aid human interpretation of fake news. Social media platforms can integrate these insights into their user interfaces to provide real-time feedback and warnings about potentially misleading content. For example:

1. Sentiment Warnings
    - Posts with exaggerated emotional tones, which are often characteristic of fake news, could trigger warnings such as, "This content may be emotionally charged. Verify the source before sharing."

    - These alerts can prompt users to think critically about the information, reducing impulsive sharing behavior that propagates misinformation.

2. Post Length Indicators
    - Short, overly simplified, or extremely verbose posts could be flagged with indicators suggesting further verification is needed.

    - Length-based indicators can guide users to evaluate whether the content lacks substantive information, a trait commonly associated with fake news.

By providing these human-interpretable cues, platforms can empower users to make more informed decisions, complementing automated detection systems.

B. Improving Fake News Detection Systems

The integration of human-recognizable features, even if not always enhancing model performance directly, remains valuable. Social media companies can use these features in the following ways:

1. Hybrid Detection Models
    - Hybrid systems can use transformer models for classification and algorithms to detect sentiment and post length to explain their model's classifications.

    - For instance, platforms can develop models that flag content based on both advanced NLP patterns and simpler, interpretable features,

enabling moderators to prioritize flagged content with clear rationales.

2. Dynamic Feature Updates:
    - Based on this study's findings, sentiment and post length were statistically significant features. Misinformation evolves over time, with new linguistic patterns emerging. Regularly updating detection systems to incorporate insights from evolving trends—such as shifts in sentiment styles or changes in post length strategies—could keep models adaptive and effective. Figuring out how to implement this would require more research that goes beyond the extent of this paper, but ideas are proposed in the next section.

C. Education and Awareness Campaigns

In addition to detection systems, social media platforms can use the insights from this paper to design educational campaigns that help users recognize common traits of fake news. For example:

1. Promoting Media Literacy
    - Workshops or interactive tools can teach users to identify emotionally charged content or overly simplified posts as potential red flags for misinformation.
    - Providing users with examples of real versus fake news, highlighting sentiment and length differences, can help build critical thinking skills.

2. Transparency in Moderation
    - If implementing a Hybrid Detection Model, platforms can share details about how features like sentiment and post length are used to detect fake news in certain posts, fostering trust among users. Transparent communication about why content is flagged or removed can reduce backlash and improve public perception of moderation efforts.

D. Collaboration with Researchers and Policymakers

This study highlights the growing need for collaboration between social media companies, researchers, and policymakers to create a multi-faceted approach to combating fake news. Strategies include:

1. Research Partnerships
    - Platforms can partner with academic researchers to develop improved datasets and refine detection models, incorporating insights like those presented in this paper.
    - Joint initiatives can focus on understanding the societal impact of misinformation and designing systems that align with both technical and ethical standards.

2. Policy Development
    - Policymakers can use these insights to draft regulations that encourage platforms to adopt transparent and effective detection practices. For instance, requiring platforms to disclose how linguistic features are used in their moderation algorithms could ensure accountability.

3. Crisis-Specific Interventions
    - During events like pandemics or elections, platforms can prioritize deploying systems trained on domain-specific data, as demonstrated in this study. Quick adaptation to emerging misinformation trends is critical to mitigating its impact during crises.

. VI. FUTURE SCOPE

This section describes three ideas for future research that correspond to the submissions of Quiz #1, Quiz # 2, and Quiz # 3 for the DS 340 capstone.

A. Quiz 1: Exploring Fake News Detection Using Multimodal Data

1) Idea: An alternative approach to fake news detection involves integrating multimodal data, such as text, images, and videos, instead of relying solely on textual analysis. Techniques like multimodal deep learning, which combines natural language processing with computer vision models (e.g., CNNs for image analysis), could evaluate how visual content correlates with text-based misinformation. For instance, studies could examine the role of emotionally charged images accompanying fake news and investigate whether image sentiment and complexity complement textual features like sentiment and post length in identifying misinformation.

2) Pros and Cons: Compared to this study's text-focused approach, multimodal analysis offers richer context by capturing visual and textual manipulations that contribute to fake news. The inclusion of visual sentiment and image complexity could align well with human-interpretable features like textual sentiment and post length, providing a more comprehensive understanding of misinformation. However, the approach introduces challenges, such as the need for larger datasets containing paired multimedia elements and higher computational requirements for processing multiple modalities.

3) Possible Outcome: A multimodal approach incorporating textual and visual sentiment, as well as visual complexity, may outperform text-only models, especially for posts with rich multimedia content. It could reveal how visual and textual cues interact to mislead audiences and extend the interpretability

framework of this study to visual features, enhancing user-facing warnings and explanations.

B. Quiz 2: Incorporating Graph-Based Neural Networks

1) Idea: Another promising direction is using graph-based neural networks (GNNs) to analyze the spread of fake news based on social network interactions. This approach models the relationships between users, tweets, and their retweeting or sharing behavior as a graph. GNNs, such as GraphSAGE or GCNs, could learn patterns of propagation and user influence, identifying content likely to be fake based on spread dynamics. Additionally, this method could integrate human-interpretable features like average sentiment or post length of tweets within a network cluster, providing an extra layer of analysis.

2) Pros and Cons: The advantage of this approach lies in its ability to account for the social context of fake news and how it propagates through networks. By incorporating aggregate features like cluster-level sentiment or length averages, it could align closely with human-readable insights while leveraging social behavior patterns. However, GNNs require detailed social media interaction data, which can be challenging to obtain due to privacy concerns. This approach may also struggle with standalone content lacking sufficient propagation data.

3) Possible Outcome: By analyzing how human-interpretable features, such as sentiment and post length, vary across propagation networks, this method could enhance detection accuracy while providing additional insights into how misinformation spreads. It may also offer a framework for user-level interventions based on behavioral clusters prone to sharing fake news.

C. Quiz 3: Leveraging Generative Adversarial Networks (GANs) for Data Augmentation

1) Idea: Generative Adversarial Networks (GANs) could be used to generate synthetic fake news content for augmenting datasets and improving model training. GANs can create realistic but artificial text by training a generator to produce fake news and a discriminator to distinguish it from real content. Incorporating human-interpretable features like sentiment, post length, and even readability into the GAN's objective function could ensure that the synthetic content reflects realistic variations of these features, enhancing the quality of the augmented data.

2) Pros and Cons: The advantage of GAN-based data augmentation is its ability to create diverse, high-quality examples of fake news, potentially making detection models more generalizable. Unlike the current study, which relies on a static dataset, this method could dynamically expand training data, including underrepresented types of fake news. However, GANs may produce unrealistic text that lacks coherence, particularly if the generated sentiment or readability deviates too far from natural patterns, reducing the value of augmented data.

3) Possible Outcome: Augmented datasets enriched with human-recognizable features like sentiment and post length could improve model robustness and interpretability. This approach might also uncover how these features evolve in fake news over time, providing additional research opportunities into the relationship between linguistic patterns and misinformation. Incorporating readability as a refinement parameter in GANs could also test its relevance in distinguishing fake from real news more effectively.

Ⅶ. CONCLUSION

This study offers valuable insights into the detection of fake news during the COVID-19 pandemic by examining the interplay between advanced transformer-based models and human-interpretable linguistic features. Using a pretrained BERT model, the research demonstrated that while features like sentiment and post length are statistically significant in distinguishing fake news from real news, their integration into the model did not enhance predictive performance. These findings underscore the inherent strength of transformer architectures in capturing complex linguistic patterns, while also highlighting the potential value of human-recognizable features for interpretability and user engagement.

The broader implications of this work extend beyond the COVID-19 pandemic. The insights gained here can inform strategies for addressing misinformation in other critical domains, such as political disinformation, climate change denial, and public health crises. By leveraging both machine learning advancements and human-interpretable features, social media companies, policymakers, and researchers can develop tools and policies that strike a balance between accuracy, transparency, and user trust.

Ultimately, this paper aims to contribute to the growing body of research on combating misinformation by emphasizing the importance of understanding both machine-driven and human-centric approaches. Future work should explore innovative methods to integrate interpretability into detection systems, ensure adaptability to evolving misinformation trends, and enhance collaboration between technical and social domains to mitigate the pervasive impact of fake news.

REFERENCES

[1] J. Alghamdi, Y. Lin, and S. Luo, “Towards COVID-19 fake news detection using transformer-based models,” *Knowledge-Based Systems*, vol. 274, p. 110642, Aug. 2023, doi: https://doi.org/10.1016/j.knosys.2023.110642.

[2] S. Gundapu and R. Mamidi, “Transformer based Automatic COVID-19 Fake News Detection System,” *arXiv:2101.00180 [cs]*, Jan. 2021, Available: https://arxiv.org/abs/2101.00180

[3] A. Bridgman *et al.*, “The causes and consequences of COVID-19 misperceptions: understanding the role of news and social media,” *Harvard Kennedy School Misinformation Review*, vol. 1, no. 3, Jun. 2020, doi: https://doi.org/10.37016/mr-2020-028.

[4] B. Farhoudinia, S. Ozturkcan, and N. Kasap, “Emotions unveiled: detecting COVID-19 fake news on social media,” *Humanities and Social Sciences Communications*, vol. 11, no. 1, pp. 1–11, May 2024, doi: https://doi.org/10.1057/s41599-024-03083-5.

[5] Y. M. Rocha, G. A. de Moura, G. A. Desidério, C. H. de Oliveira, F. D. Lourenço, and L. D. de Figueiredo Nicolete, “The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review,” *Journal of Public Health*, vol. 1, no. 10, Oct. 2021, doi: https://doi.org/10.1007/s10389-021-01658-z.