Investigating Misinformation: Analyzing Social Media to Identify Fake News

Joshua Ndala

INTRODUCTION

In the era of digital communication, the quick spread of misinformation on social media sites like X (formerly Twitter) presents serious problems for the validity and reliability of online discourse. The platform introduced "Community Notes" in response to the criticism that X has turned into a place for the spread of fake news. This tool lets users report inaccurate posts with fact-checked information, which is later verified by platform administrators. Although this community-based strategy helps fight misinformation, there is still an opportunity to improve this process using automation.

This research project intends to improve the detection of fake news on X by using advanced machine learning (ML) and natural language processing (NLP) methods. The main goal of the research is to assess the performance of deep learning models, particularly Long Short-Term Memory (LSTM) networks and Recurrent Neural Networks (RNNs). Although LSTMs are made to retain knowledge over extended periods of time, they may be better at detecting subtle and context-heavy fake news content. RNNs, on the other hand, are skilled at processing sequential data, which makes them suitable for the dynamic nature of social media content.

Additionally, this research will investigate the possibilities of the Google-developed transformer model, BERT (Bidirectional Encoder Representations from Transformers). By efficiently using modern GPU architectures, BERT's ability to process data in parallel as compared to sequentially (like in RNNs and LSTMs) may provide better performance. In the case of detecting fake news on X, this study predicts that BERT, with its advanced processing capabilities, will perform better than traditional deep learning models.

Research Questions:

- How have predictive models, particularly those that include stance detection, been used to identify fake news, and what is the comparative analysis of their techniques and effectiveness?

- What performance metrics (accuracy, precision, recall) do these models show when it comes to identifying fake news on social media platforms, particularly X?
- How can machine learning and natural language processing be combined in a way that works well together to identify signs of misinformation on social media, as currently addressed by "Community Notes" on X?
- In the complex and changing world of social media-based fake news, what new techniques or methods may be developed to improve the automation and accuracy of stance detection-based models?

This research aims to contribute significantly to the body of knowledge in fake news detection by exploring and comparing the performance of advanced machine learning models, thereby offering insights and potential tools to enhance the effectiveness of misinformation combat strategies on platforms like X.

LITERATURE REVIEW

Detecting Fake News Using Machine Learning A Systematic Literature Review

Alim Al Ayub Ahmed, Ayman Aljarbouh, Praveen Kumar Donepudi, Myung Suh Choi

This paper evaluates the state of machine learning for fake news detection through an in-depth analysis. It emphasizes the range of approaches used and the importance of diverse and high-quality datasets in building reliable predictive models, recommending multiple approaches to increase detection accuracy.

Research questions guiding the study:

- How have predictive models, particularly those that include stance detection, been used to identify fake news, and what is the comparative analysis of their techniques and effectiveness?
- What performance metrics (accuracy, precision, recall) do these models show when it comes to identifying fake news on social media platforms, particularly X?

A Literature Review of NLP Approaches to Fake News Detection and Their Applicability to Romanian-Language News Analysis

Costin Busioc, Stefan Ruseti, Mihai Dascalu

This study emphasizes the importance of building large and accurate datasets while offering an in-depth review of NLP techniques in the fight against fake news. It focuses attention on the challenges and opportunities that come with using these methods for low-resource languages, emphasizing the demand for creative solutions that take into account individual linguistic and cultural settings.

Research questions guiding the study:

- How can machine learning and natural language processing be combined in a way that works well together to identify signs of misinformation on social media, as currently addressed by "Community Notes" on X?
- In the complex and changing world of social media-based fake news, what new techniques or methods may be developed to improve the automation and accuracy of stance detection-based models?

Stance Detection for the Fake News Challenge: Identifying Textual Relationships with Deep Neural Nets

Ali K. Chaudhry, Darren Baker, Philipp Thun-Hohenstein

This study explores using advanced computer models, specifically LSTM networks and Global Vectors for Word Representation (GloVe), for identifying the stance of news headlines relative to articles. LSTMs are smart algorithms that remember and use past information, making them great for understanding and predicting sequences of words. GloVe helps by turning words into numerical values based on how often they appear together, capturing their meaning. Together, they significantly improve the ability to detect whether a news headline accurately reflects the content of its article, a key step in identifying potentially misleading news.

Research questions addressed:

- How have predictive models, particularly those that include stance detection, been used to identify fake news, and what is the comparative analysis of their techniques and effectiveness?
- What performance metrics (accuracy, precision, recall) do these models show when it comes to identifying fake news on social media platforms, particularly X?

- How can machine learning and natural language processing be combined in a way that works well together to identify signs of misinformation on social media, as currently addressed by "Community Notes" on X?

Fake News Detection using Bi-directional LSTM-Recurrent Neural Network

Pritika Bahada, Preeti Saxena, Raj Kamal

This research investigates the use of Bi-directional LSTM (Bi-LSTM) networks to detect fake news, comparing them to other deep learning models such as CNN, vanilla RNN, and unidirectional LSTM. The study emphasises LSTM's better accuracy—Bi-LSTM specifically. It emphasises its ability to process textual information from both forward and backward directions to better capture the context, which is critical for accurate classification.

Research questions addressed:

- How have predictive models, particularly those that include stance detection, been used to identify fake news, and what is the comparative analysis of their techniques and effectiveness?
- What performance metrics (accuracy, precision, recall) do these models show when it comes to identifying fake news on social media platforms, particularly X?
- How can machine learning and natural language processing be combined in a way that works well together to identify signs of misinformation on social media, as currently addressed by "Community Notes" on X?
- In the complex and changing world of social media-based fake news, what new techniques or methods may be developed to improve the automation and accuracy of stance detection-based models?

Detecting Rumors from Microblogs with Recurrent Neural Networks

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, Meeyoung Cha

To detect rumours on microblogging sites, this study compares the effectiveness of RNNs against conventional machine learning techniques that mostly rely on manually generated features. The research uses the sequential structure of microblogs to recreate the time-based and social context, improving detection capabilities by gaining a deeper understanding of the text changes related to rumours. Real-world dataset results show that RNNs can perform better than current advanced methods and offer early rumour detection.

Research questions addressed:

- How have predictive models, particularly those that include stance detection, been used to identify fake news, and what is the comparative analysis of their techniques and effectiveness?
- What performance metrics (accuracy, precision, recall) do these models show when it comes to identifying fake news on social media platforms, particularly X?
- How can machine learning and natural language processing be combined in a way that works well together to identify signs of misinformation on social media, as currently addressed by "Community Notes" on X?

METHODS

Data Collection and Preprocessing

The study used two different datasets: over 40,000 labeled articles from the University of Victoria for training and around 3,000 tweets from 2015 and 2016 for testing. Additionally, models were later retrained using a dataset from a Bangladeshi source to explore performance differences with texts similar to social media.

Preprocessing involved several important steps:

- Text Normalisation: For better consistency in tokenization, all text data were changed to lowercase to standardize the input.
- Noise Removal: To clean up the data, non-helpful text elements including URLs, user mentions, and special characters were removed.
- Word cloud visualization: This tool helps to analyze unique textual features by identifying common words in both fake and real news.
- Source Removal: To stop models from associating certain publishers with credibility, identifiers such as "Reuters" were eliminated from real news stories.

Model Selection and Setup

For this investigation, two advanced neural network architectures were selected:

- RNN and LSTM: These were chosen because of their ability to manage sequence data, which is necessary for preserving the contextual connections found in text data.

- BERT (Bidirectional Encoder Representations from Transformers): Chosen for its advanced capabilities in multiple NLP applications, BERT was applied to take advantage of its deep understanding of linguistic nuances and context. Due to time restrictions, the model was fine-tuned for two epochs after being pre-trained on a large dataset; more epochs will likely be required to improve performance.

Model Training and Validation

- Vectorization and Tokenization: To guarantee that every input sequence was the same length, texts were tokenized and padded to a uniform length based on the word count distribution for RNN/LSTM. Tokenization for BERT included using its tokenizer to turn words into token IDs and attention masks.
- Training Process: A batch size of 32 was used to train the models, and a suitable learning rate was chosen for RNN/LSTM based on initial research, and 1e-5 for BERT. To prevent overfitting, the training included monitoring accuracy and loss on a validation set.
- Validation Strategy: The data was divided into training and test sets in a 70:30 ratio for the RNN and LSTM models, and a 70:15:15 ratio for BERT as a validation set was included. The validation set was useful in fine-tuning the models and determining the most effective epoch.

Metrics for Evaluation

To thoroughly evaluate performance across classes, the models were assessed using accuracy, precision, recall, and F1-score. These indicators were used to prove the algorithms' fairness and effectiveness in identifying fake news.

Sentiment Analysis

To investigate the characteristics of real versus fake news, sentiment analysis was performed on both the news dataset and the tweets. This investigation aimed to find any emotional tone contrasts between real and fake news that might affect public opinion.

For manageability and efficient computing, 15% of the real and fake news datasets were chosen at random. Hugging Face's transformers library's distilbert-base-uncased model, recognized for its effective sentiment classification, was used for the sentiment analysis.

- Tools and Pipeline: The text was prepared and processed using an AutoTokenizer and an AutoModelForSequenceClassification. A sentiment analysis pipeline was set up, allowing the model to be applied directly to the data.
- Analysis Execution: To meet the input requirements of the model without significantly reducing the amount of information, text data was limited to the first 512 characters. The pipeline generated a sentiment score between 0 and 1 for each text entry. 1 indicates positive sentiment, 0 indicates negative, and 0.5 indicates neutral
- Data Application: Sentiment scores were calculated to compare the average emotional content of the two types of news by applying the function to both the real and fake news samples.

Comparative Study and Repeatability

The performances of RNN, LSTM, and BERT were compared, revealing differences that provide information on the practicality of each model for text-based fake news detection. To guarantee reproducibility, every step of the methodology—including specific setups and the reasoning behind each choice—was well-documented

The research recognized many limitations, including the possibility of biases in the datasets that are not obvious and the limitations caused by the computer capabilities. To improve model reliability, future research could increase the number of training epochs for BERT and investigate the effects of more parameter adjustment and cross-validation.

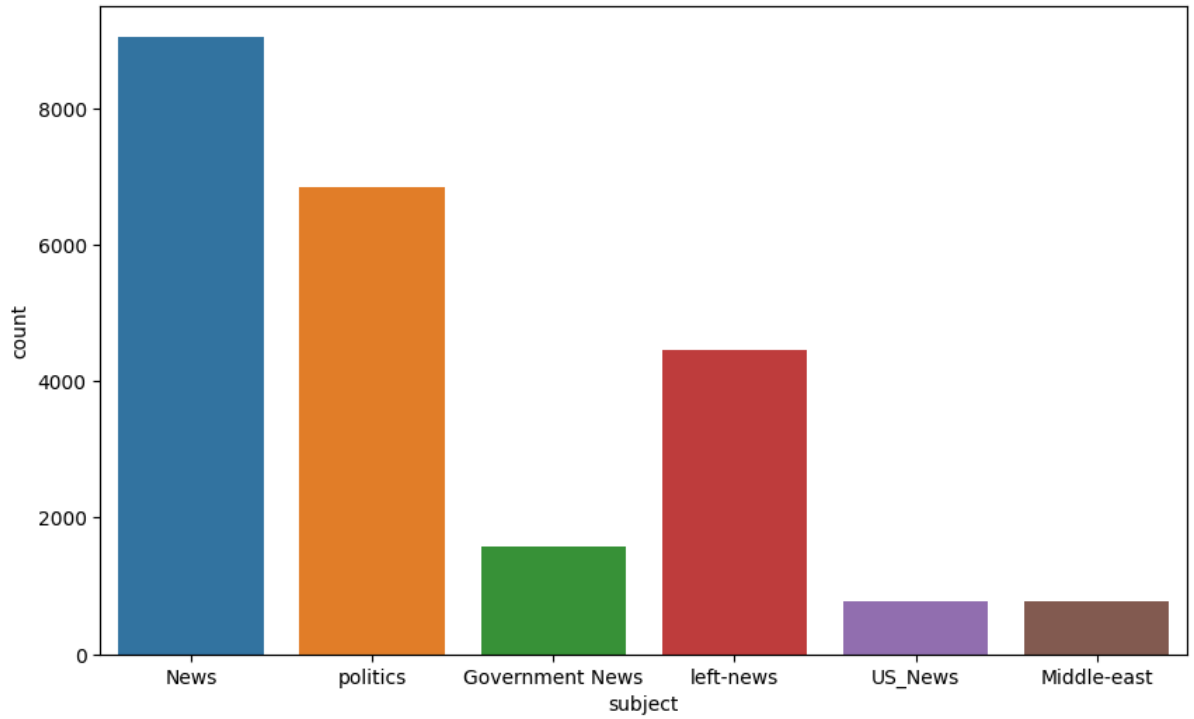
RESULTS

Distribution of News Articles by Category

An analysis of how the articles from the news dataset are distributed among different categories suggests general news and politics are the main topics of interest. These two categories make up the bulk of the content in the dataset, as Figure 1 shows. This distribution emphasises the media's major focus to these topics, giving context for the types of content most commonly related with real and fake news stories.

Figure 1

Bar Graph Representing the Number of Articles for Each News Category



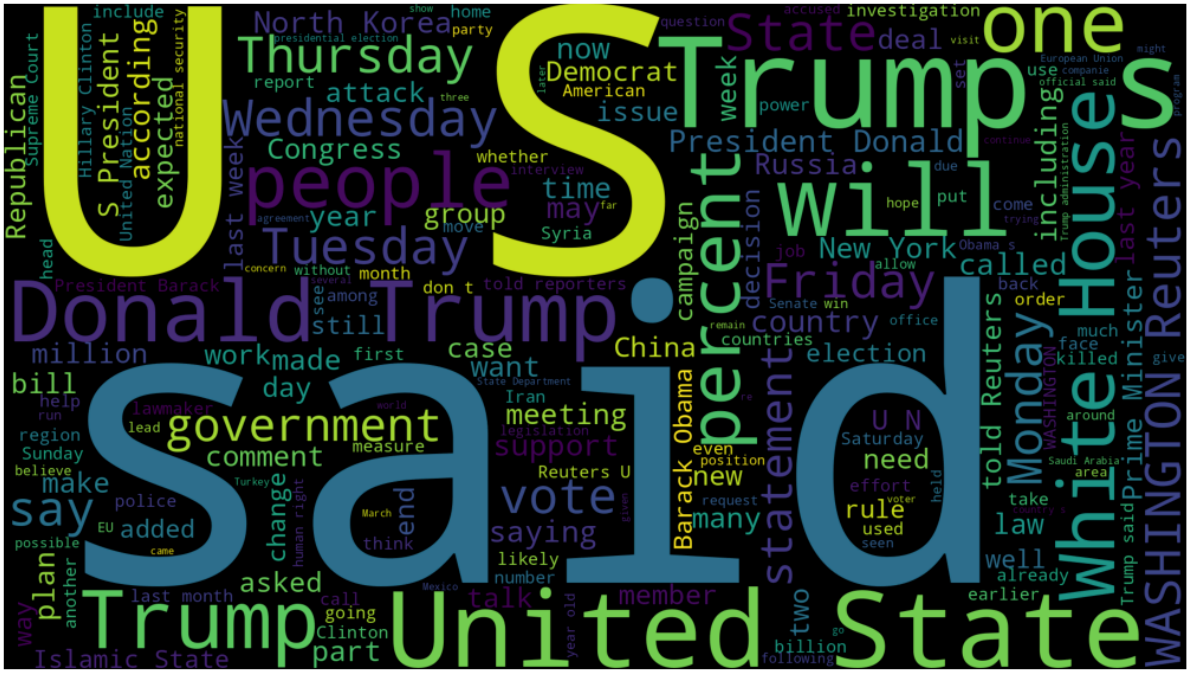
NOTE: This graph visualizes the volume of articles categorized under News, Politics, Government News, Left News, US News, and Middle East. The 'News' and 'Politics' categories are highly common in the dataset, as shown by their notable distribution.

Word Cloud for Real News

The terms that appear most frequently in the text of real news stories are shown in Figure 2's word cloud for real news. One of the most frequent phrases in this visual representation is "Washington Reuters," indicating that reliable news sources are often included in the dataset. Because of these indicators, source names had to be removed from the text to avoid biases throughout the model training phase. This stage was important for making sure the models did not just recall source names but instead learned to identify genuine linguistic patterns related to fake and true news.

Figure 2

Word Cloud Highlighting Frequent Terms in Real News



NOTE: The word cloud does a good job of reflecting the density and frequency of phrases. "Washington Reuters" is clearly visible on the far right, indicating how frequently it appears in the real news articles.

Initial Performance on News vs. Tweets

The performance metrics of each model to its particular original news dataset validation and tweet testing are shown in Table 1. This approach draws attention to the performance differences between more informal, social media text styles and organized news data.

Table 1

Model Performance on Original News Dataset vs. Tweets

Model	Data Type	Fake			Real			Accuracy
		Precision	Recall	F-Score	Precision	Recall	F-Score	
RNN	News Dataset	0.92	0.89	0.91	0.89	0.91	0.90	90%
RNN	Tweets	0.51	0.89	0.65	0.61	0.17	0.27	53%
LSTM	News Dataset	0.99	0.99	0.99	0.99	0.99	0.99	99%
LSTM	Tweets	0.50	0.97	0.66	0.60	0.05	0.09	51%
BERT	News Dataset	0.99	0.97	0.98	0.96	0.99	0.98	98%
BERT	Tweets	0.51	0.85	0.64	0.59	0.20	0.30	53%

Performance of Bangladesh Tweets vs. Original Tweets

After being retrained using data from the Bangladeshi source, Table 2 displays the models' performance on a test set as well as when used with the original tweets. Assessing the impact of retraining using different area data on model performance is helpful.

Table 2

Model Performance After Retraining with Bangladesh Source Data

Model	Data Type	Fake			Real			Accuracy
		Precision	Recall	F-Score	Precision	Recall	F-Score	
RNN	Bangladesh Tweets	0.63	0.32	0.42	0.62	0.85	0.72	62%
RNN	Tweets	0.51	0.75	0.61	0.55	0.29	0.38	52%
LSTM	Bangladesh Tweets	0.60	0.57	0.58	0.68	0.70	0.69	64%
LSTM	Tweets	0.49	0.90	0.64	0.48	0.09	0.15	49%
BERT	Bangladesh Tweets	0.74	0.60	0.66	0.72	0.82	0.77	73%
BERT	Tweets	0.52	0.39	0.44	0.52	0.65	0.57	52%

Sentiment Analysis

The intention of the sentiment analysis performed on tweets and news articles was to find differences in the emotional tones of fake and real news. The average sentiment scores for each category are listed in the following table, which makes it easy to compare fake and real news from different data sources:

Table 3

Sentiment Analysis

Data	Average Sentiment Score
Fake News Tweets	0.528064
Real News Tweets	0.528745
Fake News	0.506894
Real News	0.507931

DISCUSSION

Performance Variability of the Model Across Formats

Significant differences were seen in the model's performance between training on long news items and testing on shorter tweets, according to the conducted experiments. Models like RNN, LSTM, and BERT initially showed great accuracy in their training sets, indicating that they could pick up on and identify patterns in lengthy texts. When applied to the tweets dataset, however, their accuracy fell drastically, averaging only about 50%. This points to a significant flaw in the models' capacity to predict from long-form text to the short often nuanced language used in tweets.

Effects of Style and Content Length

The reported decreases in performance are likely caused by the difference in text length between tweets and news articles. More context is provided by news items for models to learn from, which makes pattern recognition more accurate. On the other hand, because tweets are so short, there is less data available, which makes it harder for algorithms to differentiate between real and fake news. The informal and diverse language used on social media sites like X, which can differ significantly from the more formal language used in news stories, contributes to this complexity.

Retraining Models Effectiveness Using Tweet-Like Data

The goal of retraining the models with the Bangladesh source data—texts formatted like tweets—was to make the training more closely compatible with the features of the test data. The models' performance did not significantly improve despite this, suggesting further challenges. Even with the Bangladesh test set, which replicated the format and length of the training data, the accuracy was low. This result emphasizes how challenging it is to identify fake news in tweet-sized messages because there are fewer linguistic indicators and so the content is more nuanced.

Key Findings from Sentiment Analysis

The conclusions were further complicated by the sentiment analysis results. The average sentiment scores across articles and tweets were similar for fake and real news, indicating that sentiment may not be a valid indicator of news authenticity on its own. This finding has

significance in understanding how fake news imitates the sentimental tone of real news in order to blend in and seem credible.

Future Improvements

1. **Improving Generalisation in Text Formats:**
   Further studies should concentrate on creating models that can more successfully generalize across various text forms. This could involve including multimodal data training which uses several types of media material, or using methods such as transfer learning, where a model trained on one type of text is adapted to understand another.
2. **Advanced Techniques for Text Analysis:**
   Model performance can be improved by investigating more complex text analysis methods that can manage the subtleties and shortness of social media language. Methods like deep contextual embedding could provide models with a greater understanding of the overlooked significance of short texts.
3. **Combined Detection Systems**
   Given the challenges faced, putting in place hybrid systems that blend human supervision with machine learning models could offer important benefits. These systems could make use of the automatic detection's computing power and the human reviewers' better comprehension of the context of the content.

CONCLUSION

This study revealed significant challenges in using NLP and machine learning techniques to detect misinformation across a variety of text formats. Although the models performed quite well in their training environments, they performed poorly in shorter and stylistically dissimilar texts, like tweets. The challenges of handling brief, informal content were further validated by the outcomes of retraining using tweet-like data from the Bangladeshi source. These findings highlight the importance of ongoing development in model training techniques and the possibility of applying hybrid approaches to improve the reliability of fake news detection systems.

REFERENCES

Ahmed, A. A. A., Aljarbouh, A., Donepudi, P. K., & Choi, M. S. (2021). Detecting Fake News using Machine Learning: A Systematic Literature Review. *Psychology and Education*, *58*(1), 1932-1939. https://doi.org/10.48550/arXiv.2102.04458

Bahada, P., Saxena, P., & Kamal, R. (2020, February). Fake News Detection using Bi-directional LSTM-Recurrent Neural Network. *Procedia Computer Science*, (165), 74-82. 10.1016/j.procs.2020.01.072

Busioc, C., Ruseti, S., & Dascalu, M. (2020). A Literature Review of NLP Approaches to Fake News Detection and Their Applicability to Romanian-Language News Analysis. *Transilvania*, 65-71. 10.51391/trva.2020.10.07

Chaudry, A. K., Baker, D., & Thun-Hohenstein, P. (n.d.). Stance Detection for the Fake News Challenge: Identifying Textual Relationships with Deep Neural Nets. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2760230.pdf

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., & Cha, M. (2016). Detecting Rumors from Microblogs with Recurrent Neural Networks. https://www.ijcai.org/Proceedings/16/Papers/537.pdf