CS4248 Group 39 Final Report - Fake News Classification

A0152077H, A0219678L, A0252615J, A0253042U, A0219771X, A0218065H
Group 39
Mentored by Xian He
{e0021474,e0550336,e0958203,e0958630,e0550429,e0544101}@u.nus.edu

Abstract

Fake news has become increasingly common in the digital era, and the spread of fake news can lead to severe consequences in both social and political aspects Fleming (2021). To protect news consumers from the harmful consequences of fake news, this work explores the linguistic characteristics of reliable news against those of satire, hoax and propaganda news types, and compares the performance of different classification models in detecting and differentiating these four types of news. The best-performing model for our data is a LR (logistic regression) model with a macro F1 score of 78%. We conduct an error analysis on wrongly predicted data samples, and suggest directions of future effort.

1 Introduction

Fake news refers to false or misleading information presented as news, and can include disinformation, propaganda and hoaxes Wikipedia. Fake news can cause distrust in the news ecosystem, the spread of false or discredited science, as well as harmful conspiracy theories and hate speech Fleming (2021). Based on an Statistica survey published in 2023, more than 40% of US news consumers saw fake Covid-19 information in early 2023 Watson (2023).

Hence there is a real and urgent need for effective fake news detection models to help news consumers not be harmed by fake news consequences. The aim of this project is to build a fake news detection model, utilising our learning from NLP (Natural Language Processing) classes.

2 Related Work

In Rashkin et al. (2017), the authors curated the LUN (Labeled Unreliable News) dataset, incorporating lexical features such as LIWC (Linguistic Inquiry and Word Count), which groups words into psychologically meaningful categories Tausczik and Pennebaker (2010). These features encompass the use of strongly and weakly subjective language through sentiment lexicons, hedging forms indicating vagueness and lack of commitment, and lexicons gauging the degree of dramatization in articles.

Their analysis revealed that deceptive news types tend to employ first-person and second-person pronouns more frequently, which suggested that reliable news sources tend to use less personalisation in their articles. Furthermore, fake news tended to exhibit significant hedging and exaggeration, in contrast to reliable new sources exhibiting more assertive language and providing concrete terms like numerical figures.

Additionally, the authors also observed that *Satire* news tended to use adverbs more, *Hoax* news tended to use less superlatives and comparatives, and *Propaganda* news tended to use more assertive verbs and superlatives.

Apart from the lexical feature analysis, Rashkin et al. (2017) also compared the performances of fake news classification with Naive Bayes (NB), Maximum Entropy (MaxEnt), and Long Short-Term Memory (LSTM) models. They found that LSTM performed the best when only considering textual inputs, but by incorporating lexical features, the performances of both NB and MaxEnt were similar to that of LSTM.

Similarly, Hu et al. (2021) used the same dataset to explore other fake news detection methods. Instead of relying only on content-based linguistic features, they employed a graph neural knowledge that used external knowledge bases to compare the news articles with entities. Their approach achieved micro F1 and macro F1 scores of 69.05 and 68.26 respectively.

In another study, Alarfaj and Khan (2023) conducted a comprehensive comparison between machine learning (ML) and deep learning (DL) methods such as Convolutional Neural Networks (CNN) and LSTM for fake news detection on a Kaggle dataset. Their analysis revealed that LR outperformed other models, although the DL models showed better performance in general, which agree with our results.

3 Corpus Analysis & Method

3.1 Dataset Description

We use the LUN dataset for this project, initially curated by Rashkin et al. (2017). It comprises of 48,854 training samples across four separate news types:

1. *Reliable*: Focuses on delivering accurate, impartial reporting adhering to journalistic standards.
2. *Satire*: Utilises humor and exaggeration to critique societal and political issues, aiming to entertain and provoke thought.
3. *Hoax*: Intentionally disseminates false information to deceive.
4. *Propaganda*: Pushes biased/misleading content to influence opinions towards a specific agenda.

The test data consists of 3,000 samples, with the four news types distributed equally among them. After we removed duplicate records, the train set comprises 48,652 unique samples, and the test set comprises 2,990 unique samples.

Unlike the test set, the classes are not equally distributed in the train set, with *Propaganda* and *Satire* being the two largest classes at 37% and 29% respectively, while *Hoax* was the smallest class at only 14%. To account for this class imbalance, resampling methods will be explored during training.

Regarding the data sources, while *Reliable* news samples in the train set have the same origin as the test set, the sources for the remaining three news types differ from train and test sets. This divergence could present challenges for classification models because different sources have different editorial styles and hence different language characteristics (e.g. writing styles).

Table 1: Dataset Summary

Label	#Train Samples	Train Samples Source	#Test Samples	Test Samples Source
Reliable	9932	Gigaword News	748	Gigaword News
Satire	13911	The Onion	745	The Borowitz report Clickhole
Hoax	6939	American News	750	DC Gazette
Propaganda	17870	Activist Report	747	The Natural News

Table 2: Text Properties Summary

Label	Mean Sentence Count	Mean Tokens Count	Mean Tokens Count per Sentence	Ratio of Capitalized Tokens
Reliable	20.3	456	25.2	17.4%
Satire	12.5	305	27.6	13.4%
Hoax	11.1	196	18.2	22.9%
Propaganda	45.8	927	20.3	27.6%

3.2 Exploratory Data Analysis (EDA)

The following sections below show our analysis and findings on different properties of the dataset.

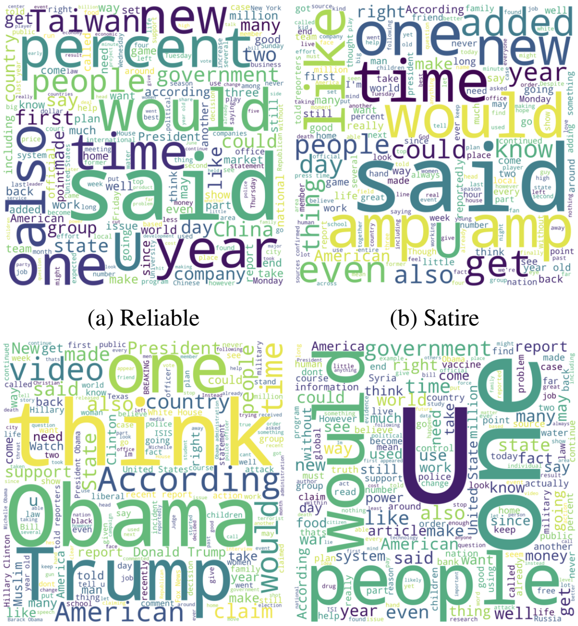
3.2.1 Text Properties

Various text properties are summarized in Table 2.

- *Propaganda* samples were the longest among all classes.
- Conversely, *Hoax* samples were the shortest.
- Despite not having the most sentences or tokens, *Satire* and *Reliable* samples featured longer sentences, as shown by their higher tokens per sentence counts.
- *Propaganda* samples also had the highest proportion of capitalized tokens, followed by *Hoax*.

3.2.2 Word Cloud

Word cloud charts are drawn in Figure 1 to analysis the high-frequency words for each class.



(a) Reliable

(b) Satire

(c) Hoax

(d) Propaganda

Figure 1: Word Cloud for 4 classes

Table 3: Sentiment Analysis Summary

Label	%Positive	%Neutral	%Negative	Subjectivity Score
Reliable	60.8	3.4	35.8	0.37
Satire	63.4	1.8	34.8	0.47
Hoax	39.7	0.8	59.4	0.45
Propaganda	41.9	7.7	50.4	0.40

- The presence of political figures like Obama and Trump is particularly prominent in *Hoax* samples, prompting us to explore NER (Named Entity Recognition) for feature engineering.
- *Reliable* samples frequently feature country and region names such as China, U.S., and Taiwan, again encouraging NER exploration.
- The word *said* is frequently used in both *Reliable* and *Hoax* samples, highlighting the paraphrasing nature of text in these categories

3.2.3 Sentiment Analysis

For sentiment analysis, we employed **Vader** to extract sentiment scores, and **TextBlob** to extract subjectivity scores from samples. Subjectivity scores range from 0 to 1, with 0 signifying high objectivity and 1 signifying high subjectivity. From Table 3 we can observe that:

- *Hoax* contains the highest proportion of samples with negative sentiment scores, followed closely by *Propaganda*. This is consistent with the findings in Salsabila and Suhardijanto (2020) and Danylyk and Vysotska (2024). Intuitively, *Hoax* and *Propaganda* articles tend to misinform or mislead readers by provoking negative sentiments.
- Conversely, *Satire* contains the lowest proportion of samples with negative sentiment scores, likely due to its comedic nature.
- *Reliable* news had the lowest subjective score, while *Satire* had the highest, aligning with our expectations of their nature of reporting and entertainment.

3.2.4 Named Entity Recognition

Motivated by the high-frequency words in the word clouds, we use **Spacy** for NER across all classes. We observed that:

- *Reliable* and *Propaganda* classes frequently feature organisational entities, which aligns with their news reporting context.
- *Hoax* and *Satire* classes most commonly feature personal entities, consistent with their narrative-driven content.
- The prevelance of non-geopolitical locations is higher in *Propaganda* than in other classes.
- Percentages are featured more prominently in *Reliable* and *Propaganda*, likely due to their higher usage of statistical data.

3.3 Data Cleaning

We cleaned the text by removing punctuation, hyperlinks, and splitting words which were joined together, especially in in *Propaganda* (e.g. "ToldMichael" instead of "Told Michael").

3.4 Data Preparation

We processed the raw text into a tokenized format suitable for model training using several steps.

3.4.1 Preprocessing

Text was standardized through case-folding and tokenization to eliminate case sensitivity, remove non-English words, emoticons, and punctuation, while retaining hashtags for their topical relevance. Stopwords removal was done using **NLTK**. Lemmatization was preferred over stemming as it slightly enhanced the model's performance.

3.4.2 Vectorization

Vector representations were generated using Count Vectorizer, TF-IDF Vectorizer for unigrams and bigrams, and GloVe, which is a word-embedding model trained on co-occurrence statistics of the corpus, and captures semantic relationships between words (StanfordNLP (2023)). Pre-trained word vectors from Wikipedia and CommonCrawl were used for GloVe. In our LR model we find that Count Vectorizer is more effective without stopwords removal, while TF-IDF performs better with stopwords removal.

3.4.3 Class Balancing

To address class imbalance, we explored **Oversampling** which generates new samples by sampling with replacement from underrepresented classes, **SMOTE** (Synthetic Minority Oversampling Technique) which generates new synthetic samples through interpolation, and **Undersampling** which removes samples from majority class.

Oversampling proved most effective in our LR model. Our hypothesis is that Undersampling leads to loss of information, while SMOTE, through the use interpolation, introduces noise and changes the meaning of original text.

4 Experiments

This section outlines our various experimental procedures and findings. Initially, we explored multiple vanilla models such as LR, and Deep Learning (DL) models such as Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT) to assess their baseline performance. This was to identify models with decent scores that showed potential, before fine-tuning each of them to optimise further. Our metric for performance will mainly be on the model's macro F1 score.

Table 4: Performance Metrics Across Different Models

Model	F1 Score (S, H, P, R)	Macro F1
Baseline	0.77, 0.55, 0.56, 0.79	0.665
TF-IDF (vanilla)	0.82, 0.59, 0.70, 0.84	0.738
TF-IDF (preprocessed)	0.77, 0.68, 0.75, 0.84	0.759
TF-IDF* (overall best)	0.77, 0.72, 0.79, 0.86	0.784
LSTM (vanilla)	0.67, 0.59, 0.49, 0.72	0.617
LSTM*	0.77, 0.64, 0.70, 0.85	0.745
BERT (vanilla)	0.88, 0.76, 0.04, 0.66	0.585
BERT*	0.88, 0.89, 0.01, 0.66	0.610

Note: S,H,P,R stands for Satire, Hoax, Propaganda, and Reliable, respectively in current and future tables. * indicates best performing for that model type.

4.1 Baseline Model

For our standard baseline model, we employed a LR model with bag-of-words features, using Count Vectorizer without any preprocessing. Despite its simplicity, this model already yielded a macro F1 score of 0.665. However, we note that it struggled with categorising *Hoax* and *Propaganda* classes.

4.2 Logistic Regression

Using TF-IDF instead of bag-of-words gave us an 11.0% increase in the macro F1 score. It also performed much better at categorising *Propaganda* classes correctly. By employing pre-processing methods like stopword removal and case-folding, we gained a further 2.85% increase in performance. This refinement reduced noise and produced a more meaningful feature set, enhancing our overall model performance.

4.2.1 Feature Engineering

We explored additional features, namely, length-related features, sentiment analysis, readability scores, named entity recognition (NER), custom vocabulary, capitalisation count, emotions, and hyperlink counts. We tested each feature individually, as well as in different combinations, on an LR model to see if it performs better than random guess (i.e. get a macro F1 score of >0.25). Note: The features are extracted prior to any preprocessing.

Table 5: Performance of Logistic Regression Model with Various Features

Feature	F1 Score (S,H,P,R)	Macro F1
Length Related	0.11, 0.22, 0.58, 0.26	0.29
Sentiment Analysis	0.36, 0.40, 0.01, 0.08	0.21
Readability	0.00, 0.43, 0.58, 0.20	0.30
Named Entity Recognition	0.41, 0.39, 0.45, 0.54	0.45
Custom Vocabulary	0.53, 0.29, 0.52, 0.61	0.49
Capitalization Count	0.12, 0.27, 0.34, 0.22	0.24
Emotions	0.20, 0.32, 0.25, 0.18	0.24
Hyperlink Counts	0.00, 0.40, 0.59, 0.00	0.25

4.2.2 Feature Analysis

We delve into the performance of key features. The values enclosed in brackets denote the macro F1 scores with its associated class.

TF-IDF with Custom Vocabulary: Effective for *Satire* (0.52) and *Reliable* News (0.61).
Analysis: The custom vocabulary comprises a curated list of the top ∼20 tokens of each class, along with a vocabulary consisting of dramatic words extracted from Wikitionary. (Refer to A). One notable observation is adverbs (such as "undoubtedly", "obviously") were found to appear most often in satire, due to their use of sarcasm, and comparative forms appeared most often in reliable news, due to its analytic nature.

Readability Score: Moderately effective for *Hoax* (0.43) and *Propaganda* (0.58)
Analysis: Hoaxes often adopt a straightforward writing style to appeal to a broader audience, resulting in higher readability scores. In contrast, propaganda can incorporate specialized jargon and complex constructions to mimic credible news sources, leading to lower readability scores.

Length Related Features: Most effective for *Propaganda* (0.58)
Analysis: Metrics such as word count, sentence length, and number of characters were particularly useful. Propaganda articles tend to be longer and more detailed, possibly aiming to persuade or mislead readers through extensive information and detailed arguments.

Sentiment Analysis: Most effective for *Hoax* (0.36) and *Satire* (0.40)
Analysis: The emotional tones and sentiment biases are captured well in hoaxes and satire. Hoaxes often aim to spread fear, leading to a negative sentiment, while satire employs humor, resulting in a more positive score. However, sentiment analysis proves less effective for *Propaganda* (0.01) and *Reliable* (0.08) classes, where the tone may be more neutral or mixed. This suggests that sentiment analysis is more useful in detecting emotionally charged false narratives.

Named Entity Recognition (NER): Most balanced feature across all classes, most effective for *Reliable* news (0.54)
Analysis: NER identifies real people, places, and institutions referenced in news articles, which provides a much richer context that can be verified against known facts. However, the presence of specific entities may also be used to assert biased viewpoints, commonly found in *Propaganda* (0.45). This suggests that NER captures both informative and manipulative elements present in news articles.

4.2.3 Feature Combination and Best Overall Results

We experimented with different feature combinations - and concluded combining length related features, sentiment analysis, readability score, and NER yields the best results, achieving a macro F1 Score of 0.55 (0.49, 0.44, 0.63, 0.61). The features are scaled using a min-max scaler. This combination leverages the complementary strengths of each feature, enhancing the model's overall performance.

Combining these features with TF-IDF (vectorized with preprocessed text), and using a chi-squared feature selector to select the most relevant 60,000 features, we achieve our best overall result across all models, a macro F1-Score of 0.784, marking a 17.9% increase from our baseline model, as seen on figure 2. This approach significantly reduces reliable news misclassification, from 351 to 137 false positives, a 61.0% decrease. This demonstrates the effectiveness of integrating linguistic and statistical approaches in news classification.

4.3 LSTM

Subsequently, we investigated the use of Recurrent Neural Networks (RNNs), more specifically Long-Short Term Memory (LSTM) which takes into account the sequence of the tokens, as LR may miss critical sequence-specific details that are potentially essential for fake news classification.

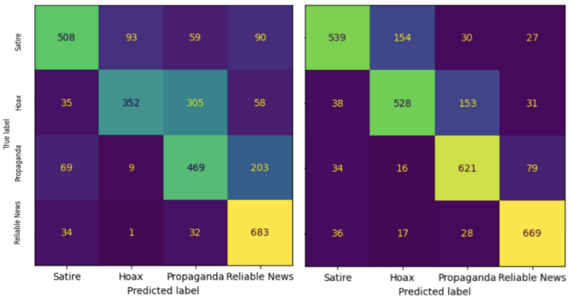


Figure 2: Baseline (left) vs Best Model (right)

Our untuned vanilla LSTM model used pre-trained word embeddings from **GloVe**. Sequence length is adjusted to follow the median to be more robust to outliers, and helps capture the typical sample in the dataset. We also do this to balance between capturing sufficient information and context of each data sample, as well as the computation efficiency of the model.

Table 6: Word Count Summary

Label	Satire	Hoax	Propaganda	Reliable News	Everything
Median Word Count	170	694.5	355	187	239

The training set we used has a median word count of 239 in total. We set the initial sequence length to be a nice round 250, padding with zeroes or truncating data samples whenever necessary.

This initial model gave us a macro F1 score of 61.7%. We attribute this low performance due to the model being affected by words outside the word embedding set. In this version, our model assigned words not found in the word embedding set as zero vectors, and this loss of information could affect the model's ability to generalise, due to some of the context being lost.

On top of this, pure LSTM may be assigning similar importance to every word, but named entities like "Obama" should be assigned more importance than a general verb like "taking", for example.

To improve this, we combined the LSTM features with TF-IDF features in a dense neural network model. We also skipped zero-vector words and used the proceeding words in the embedding set vocabulary.

The improved LSTM model gave us a macro F1 score of 74.5%, aligning with our expectations on TF-IDF features providing the much-needed

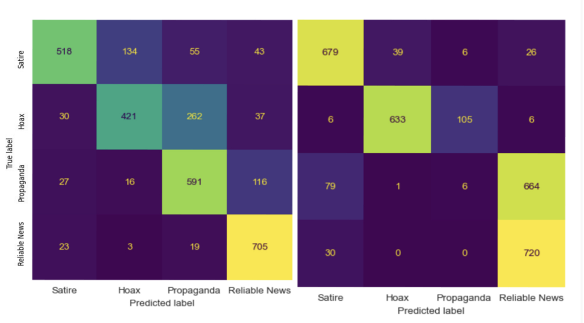


Figure 3: Final LSTM (left) and BERT (right)

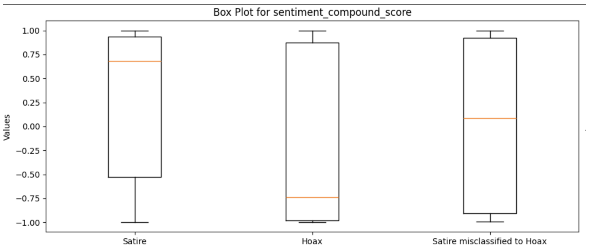


Figure 4: Satire and Hoax Sentiment Dist.

context missing from the LSTM features that the model could utilize.

4.4 BERT

We also tried BERT, specifically using the DistilBERT variant. BERT can only take in up to 512 tokens due to the model architecture, hence by default, inputs to the model are truncated to the leftmost 512 tokens, after including special tokens. Using the corresponding cased tokenizer, a baseline score of 0.58 is achieved.

BERT performs noticeably better for Satire and Reliable News, while performing slightly worse than baseline TF-IDF for Hoax, while the Propaganda was almost always being predicted as Reliable news.

Investigating Impact of Length Given that Propaganda has the highest average token count across the four classes, we tested the hypothesis that BERT's 512 token limit is the limiting factor for *propaganda*.

We considered two ways of getting around the 512-token limitation, removing all data more than 512 tokens in the training set (F1 0.88, 0.89, 0.01, 0.66), as well as chunking the text and taking the average score for predictions across each chunk (F1 0.77, 0.58, 0.16, 0.68). The improved score for *Satire*, *Hoax* and *Reliable* when removing long texts from training data suggests that the beginning and end of a document plays a large role in determining class, which is consistent with the finding in Sun et al. (2019) whose best result came from using the first 128 and last 382 tokens after tokenization. However, this finding does not apply to *Propaganda*, which F1 remained low.

Chunking increases the F1 score for *Propaganda* to 0.16, but worsened the performance of the other classes slightly. This suggests for LUN, the middle portion of *Propaganda* documents are more significant that that of other classes. Despite marginal improvements, its performance remains inadequate for detecting propaganda. It is possible that long-distance (>512 tokens) context matters for propaganda that is prevented due to the 512 tokens size limitation (due to the self-attention in BERT).

Even dropping other classes to let for a binary classification between Propaganda and Reliable news could not result in a F1 score above 0.16 for Propaganda, suggesting that with the size of the dataset, BERT is unable to dfferentiate them.

5 Error Analysis

Our error analysis begins by examining patterns in misclassification depicted in the confusion matrix of our best performing model in Figure 2. Notably, *Satire* tends to be misclassified as *Hoax*, *Hoax* tends to be misclassified as *Propaganda*, and *Progaganda* tends to be misclassified as *Reliable* news.

For each pair of misclassified classes, we took a closer analysis in the statistical distribution of the features to pinpoint the possible factors of the misclassification.

Satire misclassified as Hoax due to low sentiment compound score Analysis of the box plot in Figure 4 reveals that misclassified satire texts exhibit a lower median sentiment compound score, as compared to the correctly classified satire texts. When we filtered satire texts with sentiment compound scores below -0.9 from our test set, our macro F1 score improved to 0.792, correctly predicting 474 satire articles and reduced the number of misclassified to 114.

Further investigation into low sentiment scores reveals the use of sarcastic language in both *Satire* and *Hoax*, which may have caused confusion in classification. For instance, a Satire sample discusses Trump running for Prime Minister of Canada, but this could also be interpreted as a hoax.

What stands out in both texts (See appendix

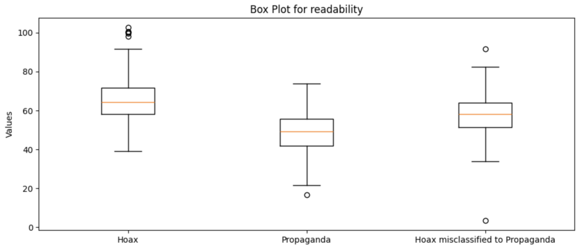


Figure 5: Hoax and Propaganda Readability Dist.

Table 7: Word Frequency Analysis

Frequency Rank	Hoax	Hoax predicted as Propaganda	Propaganda
1	obama	people	people
2	think	like	government
3	trump	just	just

B) is the use of sarcastic language. Elements of satire, such as Trump running for Prime Minister of Canada, could be rightly classified as a Hoax.

The similar low sentiments caused by negative sarcasms in both texts makes it hard to differentiate, even for the human eye.

Hoax misclassified as Propaganda due to low readability Figure 5 illustrates that misclassified *Hoax* articles exhibited lower readability, compared to correctly predicted *Hoax* articles. *Propaganda* texts, also tended to have lower readability, explaining the misclassification. Filtering *Hoax* articles with readability scores below 50 improved the F1 score to 0.793.

Hoax misclassified as Propaganda due to the word "people" We conducted a word frequency analysis on the *Hoax* and *Propaganda* training set, as well as *Hoax* articles that were misclassified as *Propaganda*.

The analysis revealed that while the word "*people*" was not among the top 3 words in the hoax training data, it emerged as the top word in both *Propaganda* as well as *Hoax* articles misclassified as *Propaganda*. To test this hypothesis, we treated the word "*people*" as a stopword and removed it from our training set. This increased our F1 score to 0.785.

5.1 Summary of Error Analysis

From our rigorous analysis, there was no singular feature that contributed to significant misclassification. The misclassifications often stemmed from inherent textual similarities, and some of them could arguably even be misclassified by humans, as some of the samples might necessitate real-world context and facts to be correctly classified.

6 Discussion

6.1 Why does simple LR perform better than DL models?

The comparative performance of LR models over more computationally intensive DL models such as LSTM and BERT in our study raises some points of discussion about the dynamics of model selection in the context of dataset size and complexity.

Dataset Size The dataset used is too small for the complex deep learning models such as BERT and LSTM as mentioned in Sutradhar et al. (2023). Unlike LR models, which are less prone to overfitting with smaller datasets, deep learning models have many parameters and are at a higher risk of overfitting if the dataset is not sufficiently large. This is evidenced in the higher test errors observed for the DL models compared to their lower training errors, indicating that these models were likely overfitting.

Feature Engineering The manual and careful approach to feature engineering can sometimes allow simpler models to excel, particularly in small datasets which we have, where clear trends and features can be observed manually, which corroborates the findings by Rashkin et al. (2017). Conversely, DL models would likely perform better if our dataset were more varied and the underlying features less discernible.

Models Limitations LSTM and BERT, while powerful in capturing contextual nuances in data, are inherently limited by their architecture to handle only predefined sequence lengths. This limitation becomes a critical bottleneck, particularly in applications like ours where the ability to process and learn from longer data sequences could be essential for achieving high accuracy and comprehensive understanding of the content.

Parameter Tuning For DL models, achieving the optimal parameter configuration can be challenging, due to the large number of parameters and its requirement for more computational power, memory, and time. It's possible that these models have not reached their optimal parameter settings.

Conclusion This discussion underscores the importance of understanding the specific attributes of the dataset to select appropriate models, rather than blindly applying the latest technologies. A thoughtful approach that combines domain knowledge with tailored feature engineering can yield more effective and efficient outcomes than relying solely on cutting-edge models unsuited to the data's complexity and size.

6.2 How will the models perform when applied to future news events?

We hypothesise that our LR model are prone to knowledge cut-off when encountering events not in their training datasets. These models tend to overfit specific terms associated with past events (See appendix C), limiting their ability to generalize to new, unseen topics.

Conversely, DL models excel at generalization across diverse news topics and events. Their capacity to learn complex patterns from large datasets enables them to abstract underlying features that are not tied to specific training data.

To test this, the CoAID dataset by Cui and Lee (2020) was used to test the models. The CoAID dataset covers only news on the the COVID-19 crisis in 2020, as compared to the dataset from Rashkin et al. (2017), which includes news dating back to the late 2000s. Only reliable news was selected, as the fake news dataset did not have a further breakdown of the different classes of fake news. The text used for classification is the concatenation of the title and content (and the title itself if there is no content).

The baseline LR model performed horribly when predicting the reliable news dataset, only managing to classify 14% of the dataset correctly as reliable news, and even classifying 73% of the dataset as propaganda.

Because TF-IDF gives higher weightage to terms frequent in one document but not others, and terms such as "vaccine" had a high TF-IDF rank in the propaganda dataset collated by Rashkin et al. (2017), it stands to reason that LR will favour predicting Propaganda when vaccines are included. The model misclassified Berman (2020) as propaganda, with model giving a -0.2155 ("vaccines") penalty for reliable news and +0.6383 ("vaccines") and +0.6079 ("vaccine") bonuses for propaganda. The BERT and LSTM models we had, while performing relatively poorly on the train and test dataset, had a 27% and 29% accuracy respectively unlike the naive TF-IDF + LR which only had 14% accuracy (See appendix D), which confirms our hypothesis.

Regardless, the challenge remains that what may be considered propaganda or a hoax today could mistakenly be trusted in the future, underscoring the need for continuous refinement and vigilance in model training and application.

7 Conclusion

This study underscores the complexities in developing models for effectively detecting and classifying fake news.

Our research faced limitations such as:

- **Dataset Constraints:** The LUN dataset presents challenges primarily due to the homogeneity of sources within each news class and the variation between the sources used for training and testing data. Most classes are predominantly composed of articles from a single source, which limits the models' ability to generalize across different editorial styles. It is also evident from 6 that our models are prone to knowledge cut-off time.
- **Model Constraints:** Logistic Regression, although robust with enhanced features, cannot grasp the deeper contextual nuances accessible to models like BERT. Meanwhile, BERT and LSTM's dependency on sequence length highlight the practical challenges in handling variable-length text data.

For future enhancements, we suggest:

- **Ensemble Techniques:** Integrating different models strategically, combining their individual strengths. For example, utilising the contextual depth of BERT with the feature sensitivity of LR could improve overall accuracy.
- **Data Diversification:** Expanding the dataset to include a wider variety of sources and a balanced class distribution might help in building more generalized models.
- **Advanced Processing:** Employing deeper semantic analysis and context-aware embeddings could refine model performance further. Advanced architectures like Longformers for BERT, and using more sophisticated feature engineering techniques could be considered.

In conclusion, the continuous evolution of news and misinformation necessitates ongoing advancements in data processing and modeling techniques to keep pace with the changing landscape of information dissemination.

References

Fawaz Khaled Alarfaj and Jawad Abbas Khan. 2023. Deep dive into fake news detection: Feature-centric classification with ensemble and deep learning methods. *Algorithms*, 16(11):507.

Robby Berman. 2020. Vaccine misinformation can be dispelled by conversation.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset.

Vitalii Danylyk and Victoria Vysotska. 2024. Information technology for detecting fakes and propaganda based on machine learning and sentiment analysis. *Qeios*.

J. Fleming. Fake news: Consequences of fake news [online]. 2021. Accessed: 19 Nov 2023.

Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 754–763.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Aghnia Salsabila and Totok Suhardijanto. 2020. Sentiment analysis on indonesian political hoaxes. In *International University Symposium on Humanities and Arts (INUSHARTS 2019)*, pages 15–21. Atlantis Press.

StanfordNLP. Global vectors for word representation (glove) [online]. 2023.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583.

Biplob Kumar Sutradhar, Md. Zonaid, Nushrat Jahan Ria, and Sheak Rashed Haider Noori. 2023. Machine learning technique based fake news detection.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Amy Watson. False news worldwide - statistics and facts [online]. 2023.

Wikipedia. Fake news - wikipedia [online].

Acknowledgements

This document has been adapted from the ACL Rolling Review Template (ACL ARR) by Min-Yen Kan. You may find the original template, which NUS has also contributed to in the past, here: https://www.overleaf.com/latex/templates/acl-rolling-review-template/jxbhdzhmcpdm.

Statement of Independent Work

1A. Declaration of Original Work. By entering our Student IDs below, we certify that we completed our assignment independently of all others (except where sanctioned during in-class sessions), obeying the class policy outlined in the introductory lecture. In particular, we are allowed to discuss the problems and solutions in this assignment, but have waited at least 30 minutes by doing other activities unrelated to class before attempting to complete or modify our answers as per the class policy.

Signed,

A0152077H, A0219678L, A0252615J, A0253042U, A0219771X, A0218065H

e0021474, e0550336, e0958203, e0958630, e0550429, e0544101@u.nus.edu

A Appendix A

This Wiktionary lexicon is inspired by https://hrashkin.github.io/factcheck.html and taken from https://hrashkin.github.io/data/fact_checking_files/wiktionarylists.zip.

B Appendix B

Example satire text misclassified as Hoax and Hoax.

Satire mispredicted as Hoax - "The new bipartisan spirit sweeping the nation was captured well by House Speaker John Boehner (R., Ohio), who tearfully told reporters, This is a great day for America oh, leave me alone, goddamn it. Republican Presidential hopeful Donald Trump made no official announcement, but sources said he was considering running for Prime Minister of Canada."

Correctly predicted Hoax - "Ted Nugent Just Pissed off Every Islamic Terrorist in the World by Pointing Out One HUGE Fact... Ted Nugent just brought to light one important detail that I guarantee none of these Islamic terrorists have even considered. In the Islamic religion, pork and pork products are forbidden because the pig is considered an impure animal. In a Facebook post Ted writes Looks like these Muslim terrorists are in for a rude awakening when they reach the afterlife! Thanks for the good news Ted!!"

C Appendix C

Table 8: Top 10 words with most positive weights from TF-IDF LR for each label.

No.	Satire	Hoax	Propaganda	Reliable
1	said	think	article	said
2	Monday	reports	articles	Taiwan
3	added	Obama	YouTube	percent
4	sources	video	com	Thursday
5	reportedly	Trump	appeared	don('t)
6	confirmed	claims	war	Wednesday
7	adding	Hillary	Corbett	Friday
8	press	recent	dont	2009
9	reporters	breaking	police	cup
10	Tuesday	ISIS	NSA	Washington

Table 9: Top 10 words with most negative weights from TF-IDF LR for each label

No.	Satire	Hoax	Propaganda	Reliable
1	think	said	said	just
2	government	Monday	think	reportedly
3	Obama	Tuesday	Monday	think
4	video	don('t)	Tuesday	don('t)
5	Trump	com	reporters	Americans
6	says	added	don('t)	reports
7	Taiwan	sources	told	sources
8	post	percent	president	Trump
9	support	Thursday	Thursday	article
10	did	press	week	video

Rankings were obtained from the coefficients of the weights from a TF-IDF LR trained with xtrain.txt.

Notice that the LR learns that specific topics are covered more in some classes, such as "Trump" and "Obama" for Hoax and "war" and "police" for Propaganda.

"don('t)" is actually "don't", but was cut into "don" due to the tokenizer.

D Appendix D

Model prediction for NewsRealCOVID-19.csv from Cui and Lee (2020), showing the number of classifications for each label. The correct label is Reliable news.

Table 10: Model prediction for NewsRealCOVID-19.csv from Cui and Lee (2020), showing the number of classifications for each label. The correct label is Reliable news.

	Satire	Hoax	Propaganda	Reliable
TF-IDF LR	92	18	655	128
LSTM	115	34	478	266
BERT	115	88	444	246

E Appendix E

Our project repository for reference `https://github.com/kktai1512/cs4248_project`