Master of Science for Business Analytics

NUS Business School & School of Computing

BT5153: Group Project Report for Group 7

Members:
Alex Chen Chen (A0280580E)
Laura Ngoc Ha Do (A0280548X)
Lim Ciwen Brendan (A0216513N)
Toh Yu Qi Chermaine (A0172396W)
Wong Cheuk Wah (A0280543H)

Fake Review Detection

Abstract

In recent years, online consumer reviews have gained increasing importance and have become a fundamental aspect of the shopping experience and decision making for customers across e-commerce and traditional retail sectors. The rise in fake reviews driven by their profitability has led to a growing presence of deceptive feedback, posing risks to both consumers and businesses. Consequently, identifying these fake reviews is essential to protect consumers and honest businesses. In this project, we explore various deep learning models using an Amazon e-commerce dataset containing both GPT-2 generated fake reviews and genuine product reviews to develop a fake review detector. We train, validate, and refine the model using this dataset to improve its performance. Our experiments show that the deep learning model we propose effectively detects fake reviews, achieving impressive performance metrics including precision, recall, and F1-score, thereby demonstrating its state-of-the-art efficacy. We additionally show how Part-of-Speech (PoS) features are used for the interpretation and generalisability of our model to examine any weaknesses of our methodology.

1 The importance of fake review detection and project objective

The growing reliance of consumers on online reviews, coupled with the surge in fake reviews, emphasize the need for a fake review detector capable of distinguishing between authentic and AI-generated reviews.

1.1 The impact of fake review detection in business integrity

In today's competitive market landscape, fake reviews have become widespread. A Bloomberg report from 2020 revealed that 42% of 720 million Amazon reviews posted were classified as fake (Lee, 2020). One major catalyst behind this trend is the profitability linked to soliciting fake reviews. According to findings from the Federal Trade Commission (FTC), the investment in fake reviews can yield a sales revenue twenty times greater than the initial expenditure.

However, the evolution of online consumer reviews (OCRs) into a fundamental component of the shopping journey and purchasing decision poses detrimental effects on both consumers and businesses. Fake reviews can deceive numerous consumers into purchasing and spending more on low-quality products, ultimately diminishing overall trust in OCRs and impacting the sellers' reputations when these customers are dissatisfied with their purchases. For businesses, fake reviews also present a significant threat. Those resorting to generating fake reviews to bolster their sales may jeopardize their credibility and reputation when they fail to meet expectations. Additionally, honest businesses may find themselves engaged in advertising and campaigning efforts to counteract competition resorting to unethical practices.

By leveraging deep learning techniques, the identification of fake reviews can achieve higher levels of accuracy and scalability compared to manual detection, especially considering the exponential growth of online reviews (Salminen et al., 2022).

1.2 Objective of the project

With the advancement in natural language processing techniques, computer-generated fake reviews can be produced quickly, at scale and lower cost compared to human-created ones where paid content creators craft authentic-seeming yet fictitious reviews without product interaction. Differentiating between genuine and fake OCRs is increasingly challenging as they can mimic the language and structure of real ones.

There is an urgent need for greater transparency regarding review authenticity to safeguard consumers and honest businesses. The main objective of this project is to use deep learning to discern these AI-Generated fake reviews. It allows for a more nuanced analysis of language patterns and structures, enabling our model to better discern fake reviews and achieve higher accuracy compared to traditional machine learning models. Ultimately, we aim to develop a transparency tool to empower consumers to make well-informed purchasing decisions by identifying potentially AI-generated OCRs and protect businesses from these unethical marketing strategies.

1.3 Overview of the models

To classify if a review is fake, we explore and benchmark the following models. A more detailed section on these models will be discussed in Section 2 of the report.

Table 1. Models implemented

	Model	
1	Logistic regression	Baseline
2	DistilBERT	Pre-trained
3	RoBERTa	Pre-trained
4	XLNet	Pre-trained
5	Multi-head Self-attention	Self-trained

2 Data Description, Preprocessing and Data Labeling

In this section, we will delve into the dataset utilized for our exploration and discuss the preprocessing steps taken before training the model.

2.1 Real Amazon E-commerce Reviews and GPT-2 Generated Reviews for Fake Review Detection

The dataset used to develop our solution is "fake reviews dataset[1]" which is openly available on Kaggle. It contains approximately 40,000 product reviews for the top 10 product categories with the highest number of product reviews evenly divided between authentic and fake entries. This dataset, created by Salminen, et al. for a research study, includes authentic reviews sourced from Amazon product reviews in 2018 which are assumed to be real and fake reviews were generated using GPT-2. These fake reviews were generated through a stratified sampling method, randomly selecting an equal proportion of reviews per category. These categories were chosen as they represent 88.4% of the reviews in the Amazon product reviews, thereby providing a representative sample.

2.2 Dataset Description

The dataset consists of 4 columns: *category* for product categories, *rating* for customer ratings ranging from 1-5, *label* indicating Original (OR) or Computer-Generated (CG), and text_ containing the review text. Each row corresponds to a specific product, including its category, customer rating, label, and corresponding review. Fig. 1 illustrates the distribution of labels across various attributes. Given the absence of class imbalance, over- or under-sampling is not required. Additionally, Table 2 indicates that the length of OR reviews is marginally longer than those of computer-generated (CG) reviews.

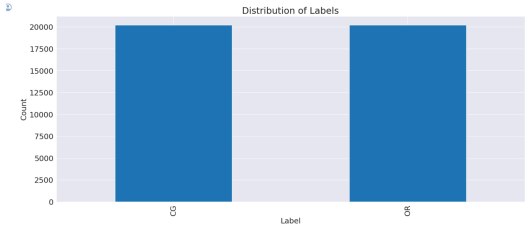

Fig. 1 (a) Overall Class Distribution

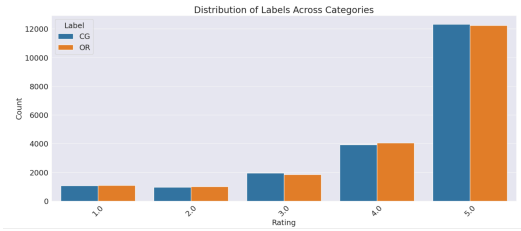

Fig. 1 (b) Class Distribution across Ratings

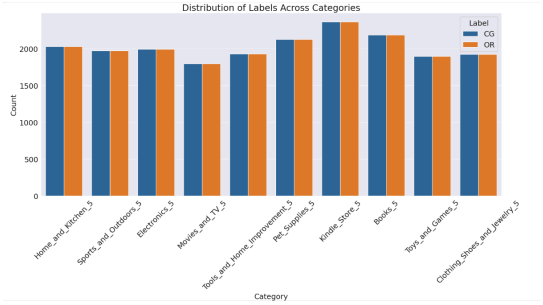

Fig. 1 (c) Class Distribution across Categories

Table 2. Length of review text for each class

Label	Min	Max	Mean	Std
CG	24	1717	305.57	308.04
OR	28	2827	396.97	418.43

[1] https://www.kaggle.com/datasets/mexwell/fake-reviews-dataset

Further analysis was done on the text in order to better understand sentence structure and other text features interpretable from a human's point of view. This was done in order to be able to qualify some of the differences based on the authenticity of the reviews before inserting it into a word embedding. Firstly, the polarity and subjectivity of reviews was assessed using Textblob. The analysis showed that in general, computer generated reviews had polarity scores 11% higher and subjectivity scores 5% higher than that of original reviews. This could indicate that the generation method used was unable to account for the strength of the words used in expressing opinions in a review, and in general less specific, leading to a larger subjectivity score.

Next, part-of-speech (PoS) tagging was performed to examine if any stark differences in the proportion of each word type used. To examine this, the count of each PoS tag was summed across the full review and divided by the total word count of the review. Across the full range of PoS tags, the tags with the largest differences in percentage of occurrence are adverbs, determiners and pre-determiners. Pre-determiners (e.g. *all, once*) observed the greatest difference, appearing 64% less in computer generated reviews than original reviews. This is followed by determiners (e.g. *the, that*) occurring 38% more and adverbs appearing (e.g. *above, happily*) 15% less than original reviews. The data indicates that fake reviews from this dataset do exhibit a slightly different profile in PoS tag occurrence, which however can only be observed on an aggregate level and not on individual reviews.

PoS analysis between the categories do not exhibit similar profiles. For example, reviews from the Books category do not exhibit similar features when compared to other categories with the same label. This indicates that the review generation process for each category was probably not done in silo but with reviews from multiple categories being fed into the generation simultaneously. This also tells us that based on the context, the profile of PoS word types does differ within authentic reviews.

Analysis on textual features enables us to better understand the generation process of the fake reviews and while this may not be consistent across multiple methods of fake review generation, it does allow us to gain a better understanding into this specific generation method used by Salminen, et al.

3 Models Development and Evaluation

As a binary classification task, the model considers the label column (OR or CG) as the prediction target and the review text as the input feature. CG is assigned as the positive class (1) and OR as negative (0).

We will evaluate model performance using standard metrics such as accuracy, precision, recall, and F1-score to gain a comprehensive understanding of how our models detect fake reviews accurately. Given the balanced class distribution in our dataset, we will prioritize accuracy. Therefore, we will select the classifier model with the highest accuracy score to minimize misclassification of computer-generated reviews as real ones, or vice versa.

3.1 Baseline model - Logistic Regression

To develop an effective fake review detector, we tested models of varying complexity to compare their advantages and benchmark their performances. We began with a simple and efficient baseline model, Logistic Regression. In the modelling process, only two columns were utilised, *label* column, serving as the target and *text_* column, acting as the feature. Preprocessing of the text was performed by fitting TfidfVectorizer to the training set and subsequently transforming the feature in both the validation and test sets.

3.1.1 Benefits and advantages

This model offers speed and explainability, providing a foundational performance benchmark and insights into how more complex models might perform. Moreover, their computational efficiency enhances its suitability for rapid prototyping. This makes it particularly advantageous for businesses seeking to deploy a minimum viable product for proof of concept purposes.

3.1.2 Result and performance metrics (Precision, Recall, F1 Score, Accuracy)

Table 3. Performance of Logistic Regression

	Validation	Test
Accuracy	0.907	0.911
Precision	0.915	0.922
Recall	0.898	0.899
F1 Score	0.907	0.910

With comparable validation and test results, the logistic regression model yielded a relatively low recall score, suggesting that it may struggle to accurately identify reviews that are truly computer-generated.

3.1.3 Limitations of the Baseline Models and Selection of more advanced Models

The main limitations of the baseline model are the ability to handle the complexity of large texts and the overall context of the reviews. This model is able to identify the

probability based on certain words, resulting in an assigned weight in the classification of the text sentiment. Logistic regression models are not suitable for this type of data due to the violation of the independence of residuals assumption. Reviews written by the same reviewer may exhibit correlation, as each reviewer expresses their unique knowledge and ideas through their distinct style. Moreover, variability among reviewers and reviews nested within reviewer clusters cannot be adequately addressed by single-level logistic models. (Le et. al, 2022)

3.2 BERT models

Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained deep learning model developed for natural language processing (NLP) tasks. It has revolutionized the field by significantly enhancing performance across various NLP tasks. Consequently, we aim to evaluate the performance of optimized versions of this transformer language model, DistilBERT and DistilRoBERTa from the Hugging Face transformers library. Initially, our proposal suggested the use of Recurrent Neural Networks (RNN). However, leveraging transformer models can address the limitations of RNNs, such as difficulty in capturing long-term dependencies.

We applied an 80-10-10 split for the training, validation and test splits respectively. The random state is maintained across all models to ensure all the data is trained and tested on the same partition of the dataset, ensuring compatibility of metric performances across all the models. The review text column was then tokenized using the tokenizers from the respective models in the Hugging Face transformer Library. We then fine-tuned the parameters in the last layer by training it on our training set. The model that gives the best performance on the validation set was then used in the classification of the test set.

3.2.1 Benefits and advantages

The primary advantage of BERT transformer models lies in their capability to analyze word sequences through the attention mechanism. It considers inputs before and after each word, enabling better context comprehension and observation of relationships between words. particularly important for detecting fake reviews. Leveraging the contextual understanding can accurately determine whether a review is genuine or fake. (Refali & Hajek, 2021). Additionally, being pre-trained on extensive corpus allows for higher accuracy in sentiment prediction.

DistilBERT was first evaluated as it was optimized from BERT using a compression technique crafted to train a smaller model with the aim of emulating BERT's functionality. Subsequently, DistilRoBERTa, distilled version of RoBERTa was assessed for comparison. RoBERTa was optimized and trained along larger datasets than the BERT model, with 124 million parameters (DistilBERT has 67 million). Therefore, it should have a deeper knowledge of nuances that could also appear on Amazon reviews, such as slang and abbreviations, and will likely outperform DistilBERT.

3.2.2 Result and performance metrics (Precision, Recall, F1 Score, Accuracy)

Table 4(a). Performance of DistilBERT

	Validation	Test
Accuracy	0.883	0.884
Precision	0.889	0.889
Recall	0.878	0.877
F1 Score	0.884	0.883

Table 4(b). Performance of DistilRoBERTa

	Validation	Test
Accuracy	0.933	0.928
Precision	0.908	0.903
Recall	0.966	0.960
F1 Score	0.936	0.930

The validation and test results were relatively similar, with the DistilRoBERTa model performing slightly better. Its higher recall rate indicates that it is effective at capturing a larger proportion of actual fake reviews, minimizing the number of undetected fraudulent reviews.

3.3 XLNet Model

XLNet, is a bidirectional transformer model that leverages the best of both autoregressive modeling (AR) and autoencoding (AE) while avoiding their limitations. It incorporates permutational language modeling which can capture bidirectional context (Yang et al., 2019).

Similarly, we applied a 80-10-10 split for the training, validation and test splits respectively and set the random state. We utilized the Hugging Face transformer Library for both tokenization and model implementation. The review text was tokenized as the preprocessing step. The parameters in the last layer were also fine-tuned by training it on our training set. Subsequently, the model that gives the best performance on the validation set was then used in the classification of the test set.

3.3.1 Benefits and advantages

The XLNet model offers several advantages over BERT. Firstly, as a generalized auto-regressive (AR) language model, XLNet avoids the pretrain-finetune discrepancy and eliminates the independence assumption inherent in BERT. Unlike BERT, which predicts only the masked 15% tokens, XLNet predicts all tokens but in random order. XLNet was trained on over 130 GB of textual data, incorporating three additional corpora compared to the datasets used in BERT (Cortiz, 2021). This extensive training enables XLNet to outperform BERT.

3.3.2 Result and performance metrics (Precision, Recall, F1 Score, Accuracy)

Table 5. Performance of XLNet Model

	Validation	Test
Accuracy	0.862	0.859
Precision	0.924	0.915
Recall	0.768	0.779
F1 Score	0.839	0.841

We observed that the XLNet model performed more poorly compared to BERT. Due to the large number of parameters, it is more prone to overfitting, especially when working with small datasets. Regularization techniques and careful hyperparameter tuning are necessary to mitigate overfitting and ensure generalization.

3.4 Multi-head Attention Model

Moving on from BERT models, we also assessed a Multi-head Self-Attention model. This was implemented by constructing custom layers using the Keras Package. For the classification task, we utilized only the encoder block of the original transformers model, designed specifically for sequence problems. Multiheaded attention enhances the model's capacity to focus on multiple positions within the text simultaneously, depending on the number of heads set in the model. This approach provides the attention layer with multiple "representation subspaces". The model then groups all the information learnt from each head and forms a classification of the text just like previous models.

The multi-head model is trained with the same 80-10-10 train, validation, test splits to maintain the consistency across all models. The random state is also kept the same to ensure the transformer models are trained with the same dataset partitions. For tokenization, sentences are split into words to determine the vocabulary size. Subsequently, they are converted to token followed by padded sequences in encoded format, where numeric encodings are assigned to each word. The model is configured with parameters including 2 heads, 32 neurons, 50 embedding dimensions, a maximum length of 512 to align with BERT models, and a vocabulary size of 39143, matching the tokenized text size in the training set (Shaikh, 2022).

3.4.1 Benefits and advantages of Multi-head Self-Attention Model

The multihead enhanced the capability of Transformer to encode multiple relationships and nuances for each word (Doshi,2021). Splitting the embedding vectors for the input sequence across multiple heads allows for capturing richer interpretations. Each section of the embedding can learn different aspects of the meanings associated with each word, giving the ability to analyze different nuances of the same word simultaneously. This provides a deeper context to help the model identify subtle patterns and inconsistencies indicative of fraudulent reviews.

3.4.2 Result and performance metrics (Precision, Recall, F1 Score, Accuracy)

Table 6. Performance of Multi-head Self-Attention Model

	Validation	Test
Accuracy	0.935	0.94
Precision	0.9337	0.9384
Recall	0.9337	0.9384
F1 Score	0.9337	0.9384

The results show a clear improvement over all other transformer models, yielding the best performance across all the metrics. It demonstrates the complexity of the dataset, as the ability for the multi-head attention model to capture multiple relationships simultaneously gives it an advantage over its counterparts.

4. Model Generalizability

The Multi-head Self-Attention Model was identified as the top-performing model among all candidates. However, the method outlined for generating fake reviews in Section 2.1 has certain limitations. There was a target sentence length for the generated reviews and the proportions of the length was to follow the original distribution in the Amazon dataset. Specifically, there was a predetermined target sentence length for the generated reviews, and the proportions of these lengths were intended to mirror the original distribution in the Amazon dataset. For instance, if 50-word reviews accounted for 0.5% of the total reviews in the Amazon dataset, then

0.5% of the generated samples would also be 50 words in length. Thus, some of the sentences were incomplete or lacked proper punctuation, such as periods at the end. This could have contributed to the model giving very high accuracy as it was able to detect such a pattern to distinguish between the real and the generated reviews. Thus, to investigate the generalizability of our fake review detector, we applied the classifier to an independent dataset.

4.1 Dataset generation

In order to explore the generalizability of our selected fake review detection model, the Multi-head Self Attention Model, we have developed an independent dataset to test its performance. To do so, we have used another dataset containing real customer reviews about Amazon Kindles. Then, we aimed to leverage Chat GPT4 to create fake Kindle reviews. However, despite several prompting efforts, GPT 4 reported difficulties in generating fake reviews. The explanation it gave were limitations in its own environment, hindering the fake Kindle review generation. Ultimately, we asked GPT 4 to share its code it would have used to generate the reviews.

Equipped with the code, we ran it in our own environment to create fake reviews. Firstly, we gave the examples for each rating category from 1 to 5 and then ran the review generation code for 50 reviews repetitively. This is due to computation and time limitations. After dropping duplicates, we could generate 75 fake reviews with this method. Paired with 75 real reviews from the Kindle dataset, we have created a balanced, independent dataset to test our model's generalizability.

4.2 Model performance on Self-Generated Dataset

Table 7. Performance of Multi-head Self-Attention Model

	Overall Result
Accuracy	0.77
Precision	0.7749
Recall	0.7667
F1 Score	0.7649

At a threshold of 0.75, the accuracy scores are approximately 77%, showing a significant decrease from the test result presented in Table 6. This may indicate that the trained model may struggle to generalize well and has learned features that are specific to the dataset used in training.

Examining the text reviews themselves, polarity, subjectivity and PoS tagging was conducted to analyse any differences in the dataset and the model prediction. Initially, when comparing the profile of PoS tagging between our Kaggle training dataset with the test dataset, we observe that most PoS categories have similar proportions of occurrence. For example in both the train and test dataset, the occurrence of adverbs are 16% more in computer generated reviews than original reviews. However, one of the categories which does not follow this trend is that of proper nouns, where in the train dataset, proper nouns occur 18% more in fake reviews while in the test dataset, proper nouns occur 59% less in fake reviews. For interpretation, this would mean that fake reviews containing more proper nouns would be more likely to slip through our detection model.

The same analysis was performed on the independent dataset. Firstly, the PoS profile for the 2 datasets are drastically different, and this is expected due to the difference in generation methods as well as different review context. One example of this is that the Kaggle dataset had fake reviews that were 17% longer than real reviews while the independent dataset had fake reviews that were 63% shorter than real reviews.

In order to examine model performance in terms of the PoS analysis on the different datasets, we compare the differences in percentage between the ground truth and the model predictions for both Kaggle and the independent dataset. The Kaggle dataset profile between ground truth and predicted values are similar as previously mentioned. However in the independent dataset, some PoS categories are very different between the ground truth and the predictions. The ground truth labels reveal that fake reviews had 47% less adverbs and real ones but when labels are generated by our model, we only observe a 21% less occurrence of adverbs in fake reviews compared to real ones. For a truly accurate prediction model, we would expect the predictions to reflect the same disparity in PoS occurrences. This highlights that while the PoS profile can be different for different datasets, our model performance is unable to fully articulate these differences in some word types. To interpret this, we hypothesize that the model is able to better identify fake reviews based on some word types (e.g. nouns) better than others (e.g. adverbs), and this is reflected by the magnitude of the difference in ground truth and predicted word type frequency. Hence, the model would perform better on reviews that have high proportions of nouns and low proportions of adverbs.

5 Discussion of insights and business value

Our project was able to identify the strength of transformer models, specially the multi-head attention model, in detecting fake reviews from our Amazon dataset. However, as seen from the model performance on the self-generated dataset from chatGPT 4, the effectiveness of the multi-head self-attention model was

not replicated, showing issues in generalizability of our model outside of the original Amazon dataset.

The results can still provide significant business value, as we were able to demonstrate the ability of constructing complex machine learning models that were able to capture specific patterns and nuances in fake reviews. The ability to handle large datasets such as the Amazon dataset can also provide insights to businesses on how transformer models can help reduce costs and personnel in the data analyzing of reviews. Businesses can use these insights to continuously monitor their model to be trained on new updated datasets that include the new variations of fake reviews in the market, resulting in a more generalizable model in detecting fake reviews. The higher computational budget in businesses can also take advantage of the use of high complexity models such as transformers to create even more powerful tools in detecting fake reviews, which would likely result in both higher optimization of consumer and supplier surplus in the online shopping sector of each seller.

6. Limitations and future directions

In this section, we will address the limitations and assumptions in our project and discuss how we can enhance the business insights retrieved from our solution.

6.1 Limitations and constraints

The first limitation of our proposed model is the possibility of unknown fake reviews existing within the original Amazon dataset, which could introduce bias to the applied language model. Given the challenge of definitively determining the authenticity of each review, we assume a low occurrence of undetected fake reviews (<5%) and expect that it will not substantially bias the detector.

Secondly, a general limitation of the deep learning models is the lack of awareness or understanding of their output. However, these models are subject to dataset specificity and frequent updates are necessary to maintain high performance as the nature of human-generated reviews evolves over time. Additionally, considering the varying nature of communication across different platforms (e.g. Twitter versus Amazon), it's crucial to assess the applicability of fake detection classifiers across platforms. This entails examining not only e-commerce product reviews but also other forms of reviews.

Lastly, regarding the role of humans in the detection of fake reviews, there has been a debate as to whether human performance surpasses the machine. Given that fake reviews exhibit detectable yet nuanced patterns, our hypothesis is that machines would outperform humans in this detection. Due to resource constraints, we will not be unable to conduct human experiments in detecting fake reviews. Consequently, we cannot verify the accuracy of our hypothesis that machines are more effective than humans in detecting fake reviews.

6.2 Future Directions

To enhance the value of our fake review detector, the model can be trained to accommodate each language to enhance global applicability. Online businesses generally transcend borders, cultures, and languages. Therefore, it should not solely focus on one language, even if English serves as the current lingua franca. Additionally, in order to deceive the detectors, minor modifications may be made to the fake reviews. Therefore, an area for further investigation includes exploring the combination of AI-generated fake reviews subsequently edited by humans to disrupt grammar, linguistic, and spacing patterns learned by the machine. The higher the efficacy of the solution in detection, the more likely it is to deter and decrease the prevalence of fake reviews.

7. Conclusion

Our model aimed to accurately detect fake reviews from the Amazon dataset using different models ranging from the baseline models logistic regression and Näive Bayes to the more complex transformer-based models BERT and multi-head attention model. While the high complexity of the transformer models resulted in a high precision, recall and F1 score of around 93% in the validation and test dataset, the models were prone to memorizing the tendencies and patterns of the fake reviews rather than obtaining the ability to generalize from it. This was shown from the significantly weaker results obtained when tested in the self-generated dataset. Fake reviews are an increasing tendency as online shopping becomes more and more prominent, resulting in many different ways of generating fake reviews to boost the ratings and popularity of sellers' items. Although our model provides an effective way of detecting patterns in fake reviews, it would be difficult to generalize the model to detect other methods of fake review generation, such as human generated fake reviews. In conclusion, our models were able to show the effectiveness in detecting fake reviews from the Amazon dataset supported by the high accuracy scores while addressing the potential applications when tested on datasets outside of the Amazon dataset used. Addressing the issues stated and possibly training the model across diverse datasets with different methods of fake review generation could result in a refined model that is more generalizable with a higher performance on future unseen data.

The code we used to train and evaluate our models is available at:
https://github.com/janecww/bt5153_group7.git

References

Cortiz, D. (2022). Exploring transformers models for emotion recognition: A comparison of Bert, Distilbert, Roberta, XLNET and Electra. *2022 3rd International Conference on Control, Robotics and Intelligent System*. https://doi.org/10.1145/3562007.3562051

Le, T.-K.-H., Li, Y.-Z., & Li, S.-T. (2022). Do reviewers' words and behaviors help detect fake online reviews and spammers? evidence from a hierarchical model. *IEEE Access*, *10*, 42181–42197. https://doi.org/10.1109/access.2022.3167511

Lee, I. (2020). *Amazon.com (AMZN) fake reviews reach holiday season levels during pandemic*. Bloomberg.com. https://www.bloomberg.com/news/articles/2020-10-19/amazon-fake-reviews-reach-holiday-season-levels-during-pandemic

Refaeli, D., & Hajek, P. (2021). Detecting fake online reviews using fine-tuned bert. *2021 5th International Conference on E-Business and Internet*. https://doi.org/10.1145/3497701.3497714

Salminen, J., Kandpal, C., Kamel, A. M., Jung, S., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, *64*, 102771. https://doi.org/10.1016/j.jretconser.2021.102771

Shaikh, Q. (2022, July 1). Transformers for text classification. Kaggle. https://www.kaggle.com/code/quadeer15sh/transformers-for-text-classification

Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., & Le, Q.V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Neural Information Processing Systems*.

Doshi, K. (2021). *Transformers explained visually (part 3): Multi-head attention, Deep Dive*. Medium. https://towardsdatascience.com/transformers-explained-visually-part-3-multi-head-attention-deep-dive-1c1ff1024853