2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
October 1-4, 2023, Oahu, Hawaii, USA

Unmasking Deception: A Comparative Study of Tree-Based and Transformer-Based Models for Fake Review Detection on Yelp

Pengqi WANG
College of Future Technology
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, Guangdong, China
eric.wangpq@outlook.com

Yue LIN
HKU Business School
The University of Hong Kong
Hong Kong SAR
yuelin_judy@outlook.com

Junyi CHAI *(Senior Member, IEEE)*
Faculty of Business and Management
BNU-HKBU United International College
Zhuhai, Guangdong, China
donjychai@uic.edu.cn

2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC) | 979-8-3503-3702-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/SMC53992.2023.10394573

***Abstract*—The increasing prevalence of fake online reviews jeopardizes firms' profits, consumers' well-being, and the trustworthiness of e-commerce ecosystems. We face the significant challenge of accurately detecting fake reviews. In this paper, we undertake a comprehensive investigation of traditional and state-of-the-art machine learning models in classification, based on textual features, to detect fake online reviews. We attempt to examine existing and noteworthy models for fake online review detection, in terms of the effectiveness of textual features, the efficiency of sampling methods, and their performance of detection. Adopting a quantitative and data-driven approach, we scrutinize both tree-based and transformer-based detection models. Our comparative studies evidence that transformer-based models (specifically BERT and GPT-3) outperform tree-based models (i.e., Random Forest and XGBoost), in terms of accuracy, precision, and recall metrics. We use real data from online reviews on Yelp.com for implementation. The results demonstrate that our proposed approach can identify fraudulent reviews effectively and efficiently. Synthesizing ChatGPT-3, tree-based, and transformer-based models for fake online review detection is rather new but promising, this paper highlights their potential for better detection of fake online reviews.**

Keywords—deception detection, fake online reviews, online reviewer behavior, machine learning, Chat GPT

I. Introduction

With the popularity of online shopping, customers need ways to help them directly evaluate products online [1]. Online product reviews emerged and served as a powerful tool for consumers to make better decisions [2]. Aware of the power of online reviews, businesses developed corresponding strategies for online reviews. They managed to alter online reviews to increase their reputation online and induce consumption [3]. Fake online reviews, also known as spam reviews, emerged as an approach to mislead consumers by displaying inappropriate positive comments to promote one business' products or malicious negative comments to damage competitors' reputations [4]. The user-generated content may be generated by human being or machines, and are hard to be recognized. Jindal and Liu [5] managed to find the feature of a fake review, and they discovered that most fake reviews are a large number of duplicate or near-duplicate reviews. Through future analysis of the metadata of online reviews, it was found that fake reviews might appear in low-sale products with many positive reviews [5]. And reviewers also can be considered a criterion to identify fake reviews. If a reviewer only wrote positive reviews for one business and negative reviews for another competing business, the reviews written are more likely to be fake. For businesses, fake reviews that highlight the opposite feature of a product are harmful: a good brand might be discredited by fake negative reviews and an undeserving brand might be promoted and seen with the help of fake positive reviews. For consumers, they risk making a worse purchase decision with unwished negative consequences [6]. It also corrupts consumers' confidence and trust in online purchasing. Consumers as review readers show increasing concern about fraudulent online information. Therefore, it is essential to identify fake reviews for both consumers and businesses, for the long-term profit of both sides. Reducing the number of fake reviews in online shipping has become the current challenge [7].

The investigation of review spamming has been initially used to study the task of fake review detection, where duplicate detection and spam classification are both employed to perform spam detection [8]. The authors classified fake reviews into two categories: mendacious opinions (e.g., unworthy positive reviews or unfair negative comments) and non-reviews (e.g., advertisements). By the analysis of Amazon data, the authors have concluded that it may be difficult to manually discern fake reviews, and alternatively, it is suggested to utilize duplicates or nearly identical responses as spam to create a model that can identify fake reviews. The research conducted by Li et al. [8] and Feng et al. [9] applied a similar method, indicating a strong association between duplication and untruthful review. Generally, there are two types of opinion spam detection approaches supervised learning [5], [10], [11] and unsupervised learning [10], [11], where supervised learning offers comparatively decent performance and flexibility, given the correct feature and labeled training data [14]. Mukherjee et al. [15] attempted to uncover Yelp's detection algorithm by analyzing the filtered reviews. The research methodology was based on the combination of LIWC,

The third author thanks the grants from BNU-HKBU United International College Research Grants with No. of UICR0200009-23 and UICR0300012, Guangdong Higher Education Upgrading Plan (2021-2025) with No. of UICR0400027-21 and UICR0400042-21CTL, as well as National Natural Science Foundation of China with No. 72271032.

standard word, and Part of Speech n-gram features employed by Ott et al. [16]. Mukherjee et al. [15] extended the research by including other POS-based features and wording styles. Additional textual features were further proposed by several researchers, such as writing styles and level of details [17], and emotion and semantic similarity [14].

There is a gap in former research on comparing the efficiency and accuracy of model-building methods. Research is done using different datasets to construct fake review detection models based on different algorithms and methods. However, comparisons between models using the same datasets but different algorithms and methods are not studied further. A lack of horizontal comparisons makes it difficult for online websites and businesses to choose the most ideal tool for fake review detection based on their own needs and position. In this paper, two main different types of models are built using the same dataset from Yelp, and we further investigated model performances and compare their accuracy. The most ideal methods and algorithms can be highlighted. And when others build fusion models, they can also determine the weight of different models, so it is critical to access model performances under the same dimension.

The essence of fake review detection is to select false comments from online websites' vast volume of comments. Because of the large number of comments and rich language information, two key problems exist in the efficient detection of fake reviews: one is efficient filtering, the ability to filter and operate quickly; another is the accuracy of the detection. Two types of models are constructed. For tree-based models, two classic machine learning algorithms, Random Forest and XGBoost, are used. Tree-based models detect each feature of the review based on the input features layer by layer until the final fake reviews are identified. Its advantages are that the number of data needed for model construction is small, and it can handle numerical and categorical data, with higher accuracy, and is more explanatory than in neural networks. In our research, we employed the state-of-the-art GPT-3 model, a transformative development in the field of transformer-based architectures, alongside BERT. Since the latter half of 2022, ChatGPT has garnered widespread attention due to its remarkable conversational capabilities, and the underlying GPT-3 model, which powers its exceptional performance, has been deemed worthy of exploration in the realm of fake review detection. Both the GPT-3 and BERT models excel in various natural language processing tasks, including fake review detection, and can handle large-scale training data. However, they require significant computational resources and are less interpretable compared to tree-based models, which may impact their suitability for certain applications. There is a gap in former research on comparing the efficiency and accuracy of model-building methods. Comparing the detection capabilities of the two different types of models will help Yelp choose more appropriate algorithms and model types to detect fake reviews more accurately when applied on a larger business scale and save their costs on improving computing power.

In this study, we focus on both traditional and state-of-the-art machine learning models, and detect fake reviews through the following research questions (RQs):

- **RQ1**: What textual features are effective for fake review detections? How to select these features for training detection models?
- **RQ2**: In training the model of fake review detection, how to select the sampling model in terms of efficiency?
- **RQ3**: Comparative analysis and evaluation of existing models for fake review detection.

We have highlighted 5 numeric and 4 non-numeric textual features that proved crucial to the model's performance. Regarding the sampling method, the balanced sample is more effective and efficient. The model using GPT to construct is the most accurate at detecting fake reviews, with an impressive performance in handling textual reviews of over 25,000 words for one review and higher reasoning ability on detection. A similar model can be applied by Yelp to improve their filtering efficiency and build a better review ecology for its users.

II. Data Preprocessing and Feature engineering

A. Data Collection

To explore deceptive reviews on the website, we use data from Yelp to construct the training model, since Yelp, as one of the largest review websites in the US, provided its filtered (fake) and unfiltered (non-fake) reviews for researchers to further analyze. Yelp is famous for its review filtering process and algorithm to provide users with reliable and trustworthy reviews, and its review filter is considered highly accurate and reliable [18]. The dataset we utilized in our study is YelpZip, which is authenticated and studied by Rayana & Akoglu [19]. YelpZip is a subset of the Yelp labeled dataset with the largest number of reviews and it restores the real website environment to the greatest extent, and the ratio of fake to true reviews is 1: 6.5.

B. Data Preprocessing

The data preprocessing phase is a crucial step in ensuring the quality and consistency of the data used for training and evaluating our models. Given the relatively complete nature of the provided dataset, our preprocessing steps are straightforward:

1. Confirming Data Integrity: We first checked the integrity of the review.json and metadata.json files to ensure they contain complete and consistent information. We then merged the datasets using the "*pd. merge*" function in Python to create a unified dataset containing both review text and corresponding labels.

2. Data Cleaning: We removed any duplicate rows and addressed encoding issues to ensure that the dataset is free from inconsistencies. Furthermore, we standardized the labels by converting them to numerical values, where 'True' is represented by 1 and 'False' is represented by 0. This step facilitates the smooth functioning of machine learning algorithms.

3. Text Normalization for GPT-3: For the training of GPT-3, we employed the OpenAI CLI tool to add a unique suffix, such as "\n\n###\n\n", to the review text. This modification is intended to improve the model's classification accuracy by

providing it with distinguishable patterns that differentiate between genuine and fake reviews.

C. Feature Engineering

We outline the comprehensive feature engineering process applied to extract valuable features from the raw Yelp review data. Our study not only encompasses traditional textual features, such as text length and similarity but also incorporates lesser-known textual features, including text embeddings and outlier Term Frequency - Inverse Document Frequency (TF-IDF) words, which is new to the literature to the best of our knowledge. By drawing on a diverse range of features, we aim to capture various aspects of the review text, such as length, sentiment, readability, and linguistic patterns. This holistic approach is designed to enhance the performance of both tree-based and transformer-based models in effectively differentiating between genuine and fake reviews. In Table I we provide a detailed explanation of each feature, supported by relevant literature to justify their inclusion in our analysis.

TABLE I. TEXTUAL FEATURES DESCRIPTION AND PERFORMANCE

Feature Extracted	Parameter Name	Description	Reference	Importance (*imp*)
Length of Text	text_length	This feature represents the number of characters in the review text. It is based on the idea that fake reviews may be of a different length than genuine ones.	[15], [6]	**0.1254**
Number of Words	word_count	The number of words in the review text. Similar to text_length, fake reviews might have a different number of words compared to genuine reviews.	[15], [6]	0.0942
Top Words Summary	top_words	A measure of the frequency of the most common words in the review text. This feature is used to determine if there are any patterns in the usage of specific words in fake reviews.	[15]	-
Sentiment Polarity	sentiment_polarity	The sentiment polarity of the review text was calculated using sentiment analysis techniques. Fake reviews might exhibit different sentiment patterns compared to genuine ones.	[16]	**0.1418**
Readability	flesch_kincaid	The Flesch-Kincaid readability score, which measures the ease of reading a text. Fake reviews might be more or less readable compared to genuine reviews.	[6]	**0.1295**
The ratio of Unique Words	type_token_ratio	The ratio of unique words to the total number of words in the review. Fake reviews might use a more limited vocabulary, leading to a lower type-token ratio.	[25]	0.0954
The ratio of Special Characters	special_char_ratio	The ratio of special characters (e.g., punctuation) to the total number of characters in the review. Fake reviews might have a different usage pattern of special characters.	[25]	**0.1280**
Frequency-Based Shifts	frequent_word_shift	A measure of the difference between the frequency of the most common words in the review and the frequency of those words in genuine reviews. This feature can help identify if fake reviews consistently use specific words more or less often.	[11], [25]	0.0307
Text Embedding	text_embedding	A numerical representation of the review text generated using the pre-trained word embedding model GloVe. This feature captures semantic and syntactic information from the text, which can be useful for detecting fake reviews.	[26]	-
The ratio of POS Tags	pos_tag_ratios	The ratios of different parts-of-speech (POS) tags in the review text. Fake reviews might have different patterns in the usage of POS tags compared to genuine reviews.	[11]	-
The ratio of Repeated Words	repeated_words_ratio	The ratio of repeated words to the total number of words in the review. Fake reviews might have more or fewer repeated words compared to genuine reviews.	[11]	0.0961
Outlier TF-IDF Words	outlier_tfidf_words	Words with exceptionally high or low term frequency-inverse document frequency (TF-IDF) scores. These words can indicate that a review is different from the majority of reviews and might be fake.	[27]	-
The ratio of Consecutive Capital Letters	consecutive_caps_ratio	The ratio of consecutive uppercase letters to the total number of characters in the review. Fake reviews might use more or fewer uppercase letters compared to genuine reviews	[11]	0.0458
Ratio of Stopwords	stopwords_ratio	The ratio of stopwords (commonly used words like "the" and "is") to the total number of words in the review. Fake reviews might have a different usage pattern of stopwords	[11]	**0.1136**

III. MODEL IMPLEMENTATION AND EVALUATION

A. Tree-Based Models (RF and XGB)

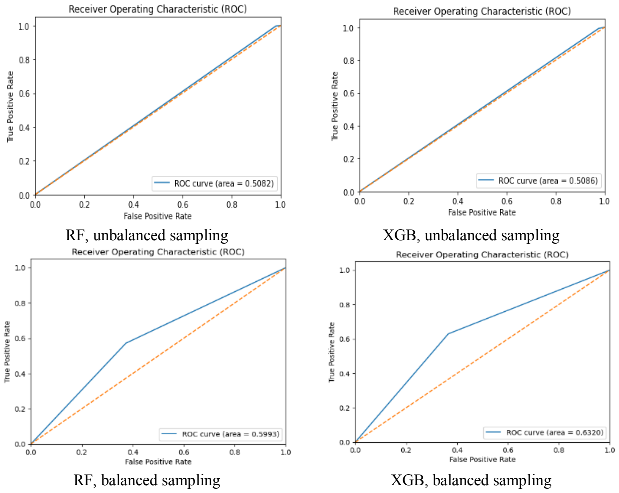
Random Forest and XGBoost, which are traditional ensemble machine learning algorithms that have gained widespread acceptance, are first trained and evaluated in this research. Due to their ability to combine multiple weak learners and create a robust model, generalization and prediction performance are enhanced.

Random Forest is an ensemble of decision trees, constructed by bootstrapping samples and using random feature subsets for each tree [20]. This approach helps to reduce overfitting and improves accuracy by aggregating multiple tree predictions through a majority vote or averaging. XGBoost stands for eXtreme Gradient Boosting, a highly optimized implementation of the gradient boosting algorithm. It builds trees sequentially, aiming to minimize an objective function that combines a loss function and a regularization term [21].

While both algorithms are tree-based ensemble models and can be used for similar tasks, there are some differences between them. RF builds multiple trees independently and combines their outputs, while XGBoost builds trees sequentially, where each tree tries to correct the mistakes made by the previous tree. This makes XGBoost more prone to overfitting compared to RF, but it also enables XGBoost to achieve higher accuracy in some cases.

TABLE II. TREE-BASED MODELS PERFORMANCE

	Random Forest		**XG Boost**	
	Balanced Sample	***Unbalanced Sample***	***Balanced Sample***	***Unbalanced Sample***
Accuracy Score	0.5989	0.8723	0.6319	0.8692
Precision Score	0.6123	0.8745	0.6396	0.8746
Recall	0.5718	0.9968	0.6296	0.9924
F1 Score	0.5914	0.9316	0.6346	0.9298
AUC-ROC	0.5993	0.5082	0.6320	0.5086



RF, unbalanced sampling

XGB, unbalanced sampling

RF, balanced sampling

XGB, balanced sampling

Fig. 1. Resulting ROC Curves of Random Forest and XGBoost, using balanced and unbalanced sample

Both Random Forest and XGBoost have been applied in various NLP tasks, including text classification and fake review detection. In text classification, the algorithms can be used to predict the class label of a given text based on its features such as word frequency, n-grams, or sentiment scores [22]. In fake review detection, the algorithms can be used to identify fake reviews based on inconsistencies in the text and other features, such as user behavior patterns, writing style, and sentiment polarity [16].

Table II shows that when a balanced sample is used (number of fake reviews: true reviews = 1:1), the precision and accuracy of the two models (0.61 and 0.64) are lower but with higher ROC. In model training, the balanced sample should be used to gain the most accurate ROC. Fig. 1 provides further insight by displaying the resulting ROC curves of Random Forest and Boost, using both balanced and unbalanced samples. However, the tree-based models performed average in classifying true and fake reviews.

B. *Transformer-Based Models (BERT and GPT)*

BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) are both state-of-the-art natural language processing (NLP) models that have transformed the field of artificial intelligence with their impressive language understanding capabilities. They are built on the Transformer architecture introduced by Vaswani et al. [23], which uses self-attention mechanisms to process input sequences in parallel rather than sequentially. Despite sharing the same underlying architecture, BERT, and GPT have notable differences in their training objectives, model structure, and applications.

BERT, introduced by Devlin et al. [24], is designed to learn bidirectional representations by training on a masked language modeling task. It learns to predict missing words in a sentence while considering both the left and right context. This enables BERT to better capture context-dependent information and perform well on various NLP tasks like question-answering, named entity recognition, and sentiment analysis.

GPT, on the other hand, is a unidirectional, left-to-right language model. Introduced by Radford et al. [28], GPT learns to generate text by predicting the next word in a sequence given the preceding words. GPT-3, the latest iteration of GPT, has been shown to perform well on tasks like text summarization, translation, and text completion, among others. In the context of fake review detection, both BERT and GPT can be employed for different purposes. BERT's bidirectional nature enables it to understand the semantic meaning and syntactic structure of a review, which can be useful for extracting features for classification. A fine-tuned BERT model can be used to classify reviews as genuine or fake, based on their textual features [29].

GPT, due to its generative nature, can be utilized to generate reviews that resemble the style and content of fake reviews. By exposing the model to a dataset of fake reviews, GPT can be fine-tuned to generate distinct responses tailored to genuine reviews and deceptive reviews. This is achieved through fine-tuning, which involves modifying the model's parameters and training it on the fake review dataset [29]. As a result, the model can generate responses that are better suited for each category of reviews.

In summary, BERT and GPT, though based on the same Transformer architecture, have distinct training objectives, model structures, and applications. BERT's bidirectional context representation can be harnessed for fake review detection through classification, while GPT's generative capabilities can be employed for data augmentation in the same domain.

TABLE III. TRANSFORMER-BASED MODELS PERFORMANCE

	BERT	**GPT-3 (curie)**
Accuracy Score	0.6969	0.6930
Precision Score	0.7604	0.7325
Recall	0.5750	0.643545
F1 Score	0.6548	0.6851
ROC	0.6969	0.7517

Table III illustrates the performance of transformer-based models, BERT and GPT-3 (Curie), in classifying true and fake reviews. While both models display similar accuracy scores (0.6969 and 0.6930), BERT has a higher precision (0.7604), and GPT-3 (Curie) has a better recall (0.6435). GPT-3 (Curie) also exhibits a superior F1 score (0.6851) and ROC (0.7517). Compared to Table II's tree-based models, BERT and GPT-3 (Curie) show improved performance in this classification task. Table III is further supported by Fig. 2, which presents the resulting ROC curves of BERT, and Fig. 3, which showcases GPT-3 (Curie) training accuracy.

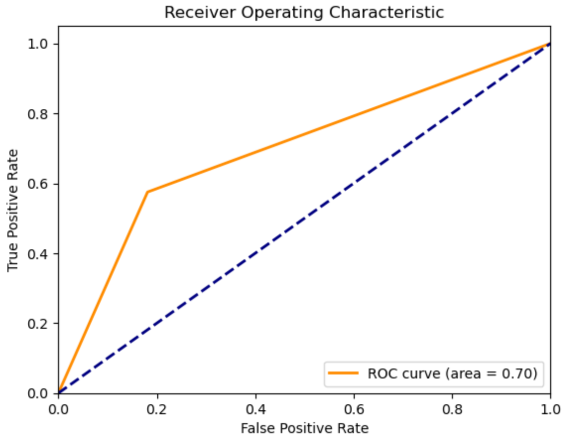


Fig. 2. Resulting ROC Curves of BERT

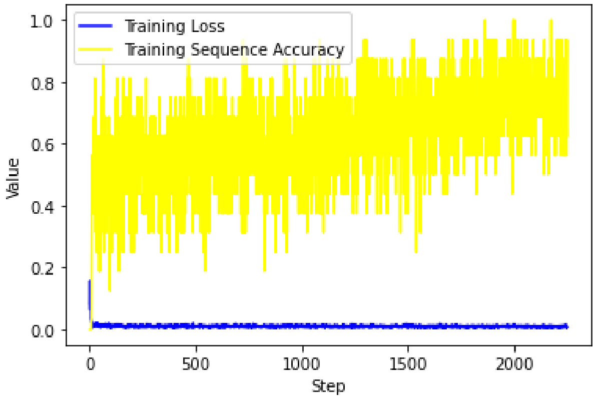


Fig. 3. GPT Training Accuracy (due to training limit, no ROC curve)

IV. Results and Analysis

In this study, we investigated a total of 14 textual features to determine their effectiveness for fake review detection and to inform the selection of features for training detection models. Among these features, 10 are numeric and can be directly measured for their importance and contribution to the model, while 4 are non-numeric and cannot be directly assessed for importance but can be converted for input into the model. The analysis identified the top 5 most important numeric features as Sentiment Polarity (imp=0.1413), Readability (imp=0.1295), Ratio of Special Characters (imp=0.1280), Length of Text (imp=0.1254), and Ratio of Stopwords (imp=0.1136). These features contribute significantly to the model's performance by capturing emotional tones, text comprehensibility, structural aspects, and linguistic patterns that may indicate deceptive content.

The non-numeric features, Top Words Summary, Text Embedding, Ratio of POS Tags, and Outlier TF-IDF Words, also proved crucial to the model's performance. They were transformed into numeric representations suitable for model input, which allowed us to assess their importance. These features help identify patterns specific to genuine or fake reviews, capture the semantic meaning of the review text, reveal linguistic patterns associated with deception, and detect unusual term frequencies that can serve as indicators of deceptive content. Incorporating these significant textual features, both numeric and transformed non-numeric, enhances the performance and accuracy of fake review detection models.

In addressing the research question concerning the selection of an efficient sampling model for fake review detection, we investigated the performance of different sampling strategies, particularly balanced and unbalanced sampling approaches. To determine the optimal sampling model for training our fake review detection model, we compared the efficiency and effectiveness of various sampling strategies. These strategies were evaluated based on their impact on the model's performance, including measures such as precision, recall, F1-score, and accuracy.

Our analysis demonstrated that balanced sampling, where the ratio of genuine (TRUE) to fake (FALSE) samples is maintained at 1:1, outperforms unbalanced sampling (using a ratio similar to the real-world distribution). The balanced sampling approach resulted in a more efficient and effective model for fake review detection. This is because balanced sampling mitigates the issue of class imbalance, which could otherwise bias the model towards the majority class and negatively impact its performance in detecting the minority class (i.e., fake reviews).

We further assessed the performance of four popular models: Random Forest, XG Boost, BERT, and GPT-3 (Curie). The models were evaluated based on several key metrics, including accuracy, precision, recall, F1 score, and ROC AUC score.

The results indicate that the transformer-based models, BERT and GPT-3 (Curie), outperformed the tree-based models, Random Forest and XG Boost, across most performance metrics. However, it is important to note that tree-based models still offer certain advantages, such as relatively faster training times and easier interpretability, which might be beneficial in certain applications or scenarios.

BERT achieved the highest accuracy (0.6969) and precision (0.7604), while GPT-3 (Curie) attained the highest recall (0.6435), F1 score (0.6851), and ROC AUC (0.7517) scores. The superior performance of transformer-based models can be attributed to their advanced architecture, which allows them to capture complex patterns, semantics, and long-range dependencies within the text. This ability to comprehend the nuanced relationships between words and phrases in reviews is essential for accurately identifying deceptive content.

On the other hand, Random Forest and XG Boost demonstrated relatively lower performance, with accuracy scores of 0.5989 and 0.6319, respectively, and F1 scores of 0.5914 and 0.6346, respectively. Despite their lower performance in this study, tree-based models remain valuable tools for various applications, particularly when computational resources are limited or when model interpretability is a priority. In conclusion, our comparative analysis demonstrates that utilizing transformer-based models, such as BERT and GPT-3 (Curie), can lead to more effective fake review detection. However, tree-based models like Random Forest and XG Boost still have their merits, and the choice of the appropriate model should be based on the specific context and requirements of the task at hand.

V. Conclusion

This paper examines tree-based models and transformer-based models in detecting fake online reviews, with their applications on Yelp.com. Tree-based models tend to be overfitting with higher accuracy and performed average in review detection. Transformer-based models, however, with the impressive capability of natural language processing, performed better than the tree-based model. With higher accuracy and ROC, transformer-based models, especially GPT-3, can help the platform better classify fake reviews. After developing the tree-based models, we found various textual features (e.g. sentiment polarity, readability, the ratio of special characters, etc.) with high effectiveness in classification. We would also point out that to our best knowledge, this study first employed GPT-3 curie, which is a state-of-art model, to detect the fake review. In future works, we plan to explore the potential of ensemble learning methods, investigating the potential of combining the predictions coming from multiple models. We intend to delve into the development of models that incorporate both behavioral and non-behavioral features [30], as well as examine the applicability of other advanced NLP models and techniques, such as GPT-4 or BERT variations in fake review detection.

References

[1] Y. Zhao, S. Yang, V. Narayan, and Y. Zhao, “Modeling Consumer Learning from Online Product Reviews,” *Marketing Science*, vol. 32, no. 1, pp. 153–169, Jan. 2013,

[2] J. Berger and R. Iyengar, “Communication channels and word of mouth: How the medium shapes the message,” *Journal of consumer research*, vol. 40, no. 3, pp. 567–579, 2013,

[3] T. Lappas, G. Sabnis, and G. Valkanas, “The Impact of Fake Reviews on Online Visibility: A Vulnerability Assessment of the Hotel Industry,” *Information Systems Research*, vol. 27, no. 4, pp. 940–961, Dec. 2016,

[4] D. Mayzlin, “Promotional Chat on the Internet,” *Marketing Science*, vol. 25, no. 2, pp. 155–163, Mar. 2006,

[5] N. Jindal and B. Liu, “Review spam detection,” in *Proceedings of the 16th international conference on World Wide Web - WWW ’07*, Banff, Alberta, Canada: ACM Press, 2007, p. 1189.

[6] C. G. Harris, “Detecting fraudulent online Yelp reviews using K-L divergence and linguistic features,” *Procedia Computer Science*, vol. 204, pp. 618–626, 2022,

[7] M. Luca and G. Zervas, “Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud,” *Management Science*, vol. 62, no. 12, pp. 3412–3427, Dec. 2016,

[8] N. Jindal and B. Liu, “Analyzing and Detecting Review Spam,” in *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, Omaha, NE, USA: IEEE, Oct. 2007, pp. 547–552.

[9] S. Feng, L. Xing, A. Gogar, and Y. Choi, “Distributional Footprints of Deceptive Product Reviews,” *ICWSM*, vol. 6, no. 1, pp. 98–105, Aug. 2021,

[10] S. Feng, R. Banerjee, and Y. Choi, “Syntactic Stylometry for Deception Detection,” p. 5, Jul. 2012.

[11] A. Mukherjee *et al.*, “Spotting opinion spammers using behavioral footprints,” in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, Chicago Illinois USA: ACM, Aug. 2013, pp. 632–640.

[12] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, “Exploiting Burstiness in Reviews for Review Spammer Detection,” *ICWSM*, vol. 7, no. 1, pp. 175–184, Aug. 2021,

[13] A. Mukherjee, B. Liu, and N. Glance, “Spotting fake reviewer groups in consumer reviews,” in *Proceedings of the 21st international conference on WWW*, Lyon France: ACM, Apr. 2012, pp. 191–200.

[14] Y. Li, X. Feng, and S. Zhang, “Detecting Fake Reviews Utilizing Semantic and Emotion Model,” in *2016 3rd International Conference on Information Science and Control Engineering (ICISCE)*, Beijing, China: IEEE, Jul. 2016, pp. 317–320.

[15] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, “What Yelp Fake Review Filter Might Be Doing?,” *ICWSM*, vol. 7, no. 1, pp. 409–418, Aug. 2021,

[16] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, “Finding Deceptive Opinion Spam by Any Stretch of the Imagination,” p. 11, 2011.

[17] S. Banerjee, A. Y. K. Chua, and J.-J. Kim, “Using supervised learning to classify authentic and fake online reviews,” in *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*, Bali Indonesia: ACM, Jan. 2015, pp. 1–7.

[18] K. Weise, “A Lie Detector Test for Online Reviewers - Bloomberg,” Sep. 30, 2011.

[19] S. Rayana and L. Akoglu, “Collective Opinion Spam Detection: Bridging Review Networks and Metadata,” in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney Australia: ACM, Aug. 2015, pp. 985–994.

[20] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001,

[21] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794.

[22] B. Liu and L. Zhang, “A Survey of Opinion Mining and Sentiment Analysis,” in *Mining Text Data*, C. C. Aggarwal and C. Zhai, Eds., Boston, MA: Springer US, 2012, pp. 415–463.

[23] A. Vaswani *et al.*, “Attention is All you Need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Long Beach, CA, USA: Curran Associates, Inc., 2017.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv, May 24, 2019.

[25] A. Mukherjee, B. Liu, and N. Glance, “Spotting fake reviewer groups in consumer reviews,” in *Proceedings of the 21st international conference on World Wide Web*, France, Apr. 2012, pp. 191–200.

[26] S. Arora, Y. Liang, and T. Ma, “A SIMPLE BUT TOUGH-TO-BEAT BASELINE FOR SEN- TENCE EMBEDDINGS,” in *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, Toulon, France, Apr. 2017.

[27] J. Ratkiewicz, M. Conover, M. Meiss, B. Goncalves, A. Flammini, and F. Menczer, “Detecting and Tracking Political Abuse in Social Media,” *ICWSM*, vol. 5, no. 1, pp. 297–304, Aug. 2021,

[28] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving Language Understanding by Generative Pre-Training”,

[29] OpenAI, “Fine-tuning guide,” *OpenAI Platform Documentation*, 2021.

[30] J. Chai, and E. W. T. Ngai, "The Variable Precision Method for Elicitation of Probability Weighting Functions," Decision Support Systems, 128, 113166.