Comparing Large Language Model Fine-Tuning and Context-Based Learning for Fake Review Detection

Mariah Nicole Roberts
Department of Computing Sciences
College of Engineering and Computer Science
Texas A & M University-Corpus Christi
mroberts4@islander.tamucc.edu

Shruthi Balakrishnan Nair
Department of Computing Sciences
College of Engineering and Computer Science
Texas A & M University- Corpus Christi
snair1@islander.tamucc.edu

Chaitanya Goud Bandi
Department of Computing Sciences
College of Engineering and Computer Science
Texas A & M University- Corpus Christi
cbandi@islander.tamucc.edu

Yasaswani Narina
Department of Computing Sciences
College of Engineering and Computer Science
Texas A & M University- Corpus Christi
ynarina@islander.tamucc.edu

***Abstract*—Shopping online is more popular than ever. Online reviews significantly influence the success of a business, but fake reviews distort consumer opinions. Traditional methods for detecting a fake review depend on a large amount of data for training but do not yield the best results. However, fine-tuning pre-trained models such as GPT-3, which require minimal training data, has surpassed traditional models in accuracy. Alternatively, context-based learning supplements reviews with contextual cues to enable deceptive detection without relying on large models. Our experiments benchmarked these approaches on artificial and genuine reviews, evaluating their relative accuracy, efficiency, and generalization. We created a user interface (UI) to analyze real-time performance. Through rigorous comparisons and analysis, we provided data-driven results on optimal AI techniques for a robust fake review detection system across various domains. This project aims to address the growing problem of manipulated reviews undermining e-commerce integrity.**

I. Introduction

Fake reviews have become a huge problem in e-commerce platforms. Consumers who buy online products heavily depend on online reviews when making purchases. However, in some cases, the reviews are mostly anonymous, which makes businesses post a positive review for their products or a negative review for their competitor's products. This manipulates the consumer's perception of buying a product. Detecting fake reviews is an important challenge that needs to be addressed to maintain the integrity of online opinions of a specific product.

As consumers depend on online reviews and ratings to inform their buying decisions, there are unethical businesses to unfairly boost their products using computer-generated reviews to disparage competitors through fraudulent reviews. As a result of using computer-generated reviews, businesses are at risk for financial loss and honest businesses with genuinely good products will be overlooked. The impact of this problem is massive and the categories involved are mostly sites that utilize e-commerce and their services such as electronics, apparel, hotels, restaurants, and many more.

The typical consumer is unable to differentiate the differences between a real review and a computer-generated one. The challenge is unique given the volume of reviews and the sophistication of computer-generated reviews mimicking genuine ones. Nonetheless, the rise of fake reviews is a vital problem that needs solutions to restore faith in online opinions.

The specific problem we aim to tackle is identifying automated techniques that can accurately flag fake and deceptive reviews. Various methods have been proposed to identify computer-generated reviews, including deep learning, XGBoost, machine learning models, and many more. In this research paper, we propose a refined approach to detecting computer-generated reviews using OpenAI's Ada, Davinci, and GPT-3.5-Turbo models, which are based on GPT-3 architecture.

Machine learning has made remarkable improvements in various areas such as text completion, detection, sentiment analysis, etc. Supervised and unsupervised learning are two of the most used machine learning techniques. In supervised learning, the model will be given a data set that is labeled to train and will be tested on the test data set. However, in unsupervised learning, the model will be trained on unlabelled data and will be tested on the test data to predict the outcomes. One other area in which machine learning has grown significantly is Natural Language Processing (NLP).

NLP in recent times has become a crucial component in applications such as virtual assistants, automatic translations,

and chatbots. Figure 1 represents the transformer model architecture which has an encoder and decoder where the encoder iterative processes the input tokens one layer after another through its encoding layers and the decoder iterative processes the encoder's output and the decoder output tokens so far through its decoding layers. The emergence of transformer models has been a path to many AI assistants in recent years, which led to the development of Generative Pre-trained Transformers (GPTs). These models are trained by large corpora of data. There are various models such as ADA, Davinci, Curie, and GPT-3.5-Turbo in GPT-3 category. Among the models, the model that costs the least for fine-tuning is Ada which is $0.0004 per 1000 tokens, and for testing the fine-tuned $0.0016 per 1000 tokens [12].

Our goal was to explore modern AI techniques like large language model fine-tuning and context-based learning that can generalize across domains, and detect fake reviews with minimal data. The massive amount of reviews makes manual identification of deception impractical. Rule-based systems that involve human-crafted rules have been tried but they fail to scale and miss semantic nuances. We need automated solutions that can flag fake opinions accurately without needing prohibitive human effort. Traditional machine learning methods like SVMs, logistic regression, etc. rely on hand-engineered input features. Feature engineering is labor intensive and the learned patterns do not transfer across domains.

These models also need large training datasets which are expensive to obtain, as real fake reviews are hard to find. We aim to experiment with advanced AI techniques based on neural networks and deep learning that can overcome these limitations. Specifically, we explore two promising approaches - large language model fine-tuning and context-based learning. Large pre-trained Language Models like GPT-3 have shown impressive capability for natural language understanding. Fine-tuning them on modest domain data could work very well for fake review detection. On the other hand, explicitly providing useful context to lightweight models can also potentially boost accuracy. Our goal is to thoroughly evaluate both techniques on diverse review data and determine optimal solutions that combine accuracy, scalability, and generalization ability for spotting fake reviews.

In our research, we address computer-generated review detection through the following research questions (RQs):

> RQ1: How effectively can large pre-trained language models be fine-tuned to discern fake reviews from genuine ones?
>
> RQ2: Can context-driven learning models enhance the accuracy of fake review classification compared to established pre-trained models?

These research questions outline the core problems we want to explore through rigorous experimentation. RQ1 examines whether leveraging the knowledge already encoded in large pre-trained LMs can produce highly accurate fake review detectors with minimal training data requirements. Large recent LMs like BERT, GPT-3, etc. have shown remarkable performance on many language tasks. Their representational capacities allow them to capture nuanced semantics, context, and meaning in text. RQ1 seeks to determine whether fine-tuning such models on modestly labeled fake review data can result in excellent classifiers that surpass traditional ML approaches reliant on extensive feature engineering.

We hypothesize the innate linguistic sophistication of large LMs will enable accurate fake review detection with orders of magnitude less training data than needed for conventional models. RQ2 explores an alternate angle - whether we can improve the performance of pre-trained LMs by explicitly providing useful contextual cues to simple feed-forward networks. Reviews don't occur in isolation. Metadata provides useful context to the model which helps in the classification. Incorporating such contextual information along with the review text could potentially boost performance for models. RQ2 asks whether such context-based techniques can improve the learning capacities of huge LMs, instead of using fine-tuned methods.

For fake review detection, we can take an existing pre-trained open-source LM and fine-tune it on a labeled fake review data set. We hypothesize that fine-tuned LMs will outperform training shallow neural networks from scratch given their advanced language capacities. We will also experiment with different model sizes and architectures. They comprise hundreds of millions of parameters, allowing them to learn extremely sophisticated representations of language structure and meaning. The pre-trained models can then be fine-tuned to adapt to specialized downstream NLP tasks by training them on much smaller supervised data sets. Critically, fine-tuning large pre-trained LMs requires far less data than training deep neural models from scratch. This makes them highly valuable for low-resource domains where labeled data is scarce.

On the other hand, explicitly providing useful context along with the primary text can potentially improve model performance without needing very large models. For fake review detection, context can include product, user, and rating metadata. Prior research has shown the value of incorporating contextual information for text classification tasks. However, textual data does not exist in isolation. Reviews are complemented by useful contextual cues like product information, user profiles, ratings, etc. Explicitly incorporating such contextual metadata along with the input review text could help improve model performance even without requiring training large models. For instance, an overly positive review for a product that mostly has negative ratings is likely insincere.

Through a comprehensive comparative analysis of synthetic and real-world data sets, we aim to determine the relative merits of each approach. Our experiments will evaluate accuracy, training costs, robustness against adversarial examples, and real-time user testing. This will provide data-backed insights into optimal techniques for building scalable and generalized fake review detection systems. The methods can potentially be extended to other deceptive content detection domains beyond just online reviews. Additionally, we developed an interactive UI to assess real-time performance based on user feedback on flagged reviews. This multi-dimensional evaluation provides

comprehensive insights into the strengths and weaknesses of each approach. Our goal is to determine the technique that combines accuracy, efficiency, and generalization ability for detecting fake reviews. This can guide the adoption of optimized solutions.

Our method for fake review detection marks a significant shift from traditional machine learning techniques such as SVMs and Naive Bayes. Existing models heavily rely on manual feature engineering and domain knowledge, whereas our neural network approach automatically learns feature representations directly from the raw data without human effort.

Additionally, most current methods analyze just the review text, but our context-aware technique also encodes metadata from the review to capture contextual clues missed by text alone. Furthermore, our approach with pre-trained language models like GPT-3 achieves high accuracy with only hundreds or thousands of labeled examples, unlike conventional models needing millions.

This adaptability to limited data makes our method invaluable for low-resource settings. The inclusion of contextual cues from reviews also improves performance without requiring more complex networks. We expect experiments will demonstrate clear accuracy gains over models without pre-training or context, highlighting the advantages of our approach for fake review detection. Overall, our technique reduces the need for extensive feature engineering and training data while leveraging both text and context for enhanced deception detection.

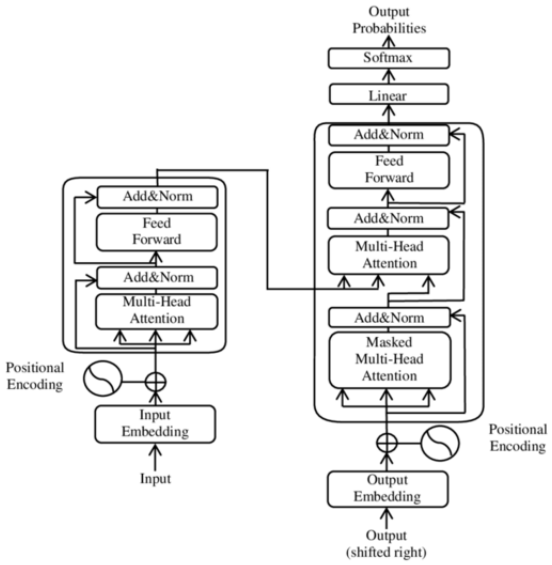

Fig. 1. Transformer Model Architecture. Adapted from [12]

II. Related works

Costa et al. [1] introduced an approach for detecting incentivized online reviews of products on Amazon using data mining techniques. The paper presents a related technique for analyzing online reviews, our approaches differ significantly in the specific objective and methods used. We want to identify fake reviews rather than just incentive-based reviews. Also, we leverage deep learning techniques rather than data mining features like review length. However, the analysis of review length and sentiment could provide useful insights that inform our feature engineering process. Overall, the paper provides a useful starting point, but our methodology will likely differ.

Dixit et al.[2] proposed an approach for predicting consumers' intention to write online reviews using an integrated model extending the Theory of Planned Behavior. While it shares similarities such as analyzing online review behaviors, our proposed approach differs significantly. Their focus was on modeling general review writing intentions based on motivations and involvement, whereas we aimed to detect fake/deceptive reviews specifically. Our techniques emphasize large language models rather than motivational modeling. However, their analysis of factors like ego involvement and vengeance motivations could provide useful insights to inform our feature engineering process. Overall the paper presents useful background research but our methodology will leverage different techniques tailored to identifying deception rather than modeling general review writing intentions.

Yogesh et al.[3] covers a wide range of topics related to digital and social media marketing research. One aspect it discusses is detecting fake or fraudulent reviews, which is also the focus of our proposed project. However, while the paper summarizes high-level findings from previous studies, our project will implement and evaluate two specific AI/ML models for identifying fake reviews - fine-tuning transformer networks and context-based learning. This paper compares these approaches by training and testing the models on review data sets, to provide more technical insights into their relative accuracy and performance for fake review detection.

For our approach, this project will have a narrower scope and will focus specifically on the comparative evaluation of these two models, rather than surveying the entire field of digital marketing research. Additionally, we intend to build a user interface for real-time testing of the models, which is a more implementation-oriented contribution not covered in the literature review. Overall, it is relevant to fake review detection, our project will provide unique insights into the performance comparison of two ML techniques, offering a more hands-on and technical perspective.

Gobi and Rathinavelu [4] introduced an approach for analyzing cloud-based reviews for product ranking using a feature-based clustering algorithm. While it shares the goal of analyzing online reviews, our proposed approach differs significantly in the techniques used. Their focus was on extracting explicit and implicit features from reviews to feed into a fuzzy clustering and ranking algorithm, whereas we want to detect fake/deceptive reviews specifically.

Our techniques emphasize large language models and contextual learning rather than traditional feature extraction and clustering. However, their analysis of mining both explicit and implicit features provides useful insights that could inform our feature engineering process when extracting signals from

review texts. Overall the paper presented a related domain area of online review analysis but our methodology will leverage very different techniques tailored to identifying deception rather than product ranking.

The target stated in our project is also shared by the study paper by Christopher G. Harris [5], which utilized machine learning approaches to identify false reviews of hotels. There are a few significant variations. In contrast to the service reviews we want to use, the study concentrates on product reviews and makes use of more conventional machine learning models like Naive Bayes and SVM. In contrast to our goal to analyze in-context learning, the research does not appear to examine giving context to models while they are being trained. Additionally, we also want to test interactive UI components while the paper evaluates categorization performance.

Qasim et al. [6] share the high-level goal of using machine-learning techniques for text classification. The paper utilizes traditional machine learning models like logistic regression and Naive Bayes, whereas we intend to leverage large language models. The paper also focuses on types of tweets rather than product reviews and does not examine providing context to models during training, unlike our plan to test in-context learning. Additionally, the paper evaluates models solely on classification performance metrics, while we also aim to assess interactive UI components.

Li et al.[7] introduced an approach for sentiment analysis of Cantonese political posts on Hong Kong local forums using fine-tuned mBERT. The paper presents a related technique for sentiment analysis utilizing mBERT fine-tuning. However, our approaches will differ in the domain application and specific techniques used. While mBERT fine-tuning is promising, our methodology will likely test different large language models beyond just mBERT. The paper provides useful analysis on sentiment analysis that could improve our proposed approach.

Zhang and Hu [8] introduced a two-stage fine-tuning approach for pre-trained models like BERT using hidden states. While it shares similarities such as utilizing pre-trained models and hidden states as a form of context, our proposed approach differs significantly. Their technique involves modifying the base BERT architecture to incorporate hidden states which may be complex to implement. Our focus is instead on techniques like large language model fine-tuning and contextual learning that do not alter the fundamental model structure. Overall, the paper provides related work on fine-tuning pre-trained models, but our techniques will diverge due to not modifying base model internals and focusing more on language-specific contextual learning.

Sumathi et al.[9] introduced an approach for detecting fake reviews of e-commerce electronic products using machine learning techniques. This paper shares some commonalities with our planned approach, such as collecting reviews from multiple sources and extracting linguistic features to train machine learning models for fake review detection. However, there are also several key differences. The paper focuses on product reviews rather than service reviews and evaluates more traditional machine learning algorithms like Random Forest rather than large language models such as Claude which we intend to test.

The paper also does not seem to examine providing context to models during training, which we plan to explore through in-context learning. Additionally, while the paper only evaluates classification performance, we aim to test interactive UI components. In summary, while the high-level goals of detecting fake reviews with ML are aligned, the specific models, features, datasets, and evaluation metrics differ between this paper and our intended approach described in the abstract. Testing large language models and UI interactions could provide additional insights beyond the techniques explored in this paper.

Sihombing and Fong [10] introduced an approach for detecting fake reviews on the Yelp website by comparing different machine learning classification techniques whereas the best result was obtained using XG Boost technique scoring 99% prediction. The research paper shares some high-level similarities to our planned approach, such as using machine learning models to detect fake reviews and extracting features from review text and ratings. However, there are also several key differences. The paper focuses on analyzing product reviews rather than service reviews and implements more traditional machine learning algorithms instead of large language models that we intend to leverage.

Additionally, the paper does not seem to examine providing context to models during training, which we plan to test through in-context learning. While the paper only evaluates the classification performance of models, our goal is to also assess interactive UI components. In summary, while both approaches utilize machine learning for fake review detection, the specific models, features, data sets, and evaluation metrics differ between this paper and our intended methodology outlined in the abstract. Our plan to test large language models and UI interactions could yield additional insights compared to the techniques explored in this paper.

III. Proposed Work

A. Fine-Tuning

Our proposed method for detecting fake reviews was to utilize the use of large language models (LLMS) such as GPT-3 which have shown impressive capabilities for understanding natural language. Our proposed approach is divided into two methods: fine-tuning the large language models and context-based learning to effectively detect fake reviews.

The Large Language Model (LLM) has a large parameter space which allows them to learn sophisticated representations of linguistic structures. For the first method, the pre-trained models Ada, Davinci, and GPT-3.5-turbo aka (ChatGPT) were used. We fine-tuned it using a labeled data set that consists of original and computer-generated reviews.

In our experimental setup, we conducted all implementations on Google Colab, taking advantage of its diverse run time resources. We employed a range of libraries, including OpenAI, pandas, matplotlib, OS, Gradio, sklearn, and seaborn. We used the pandas library for data preprocessing, converting

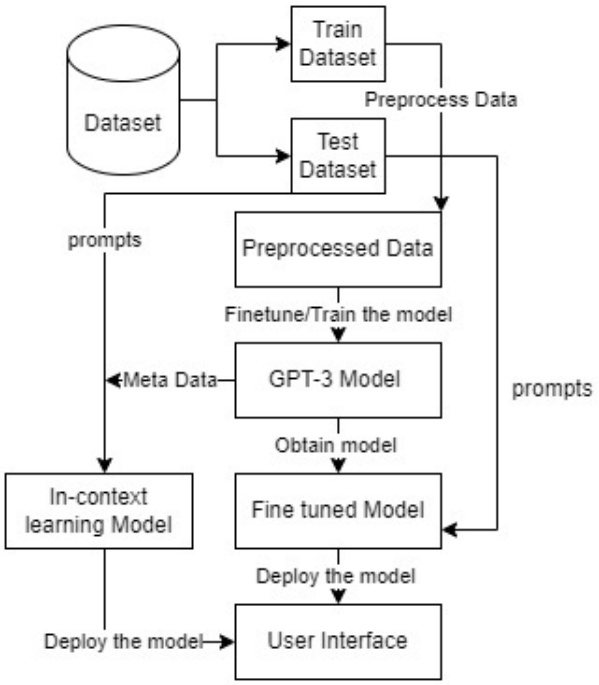

Fig. 2. Our Proposed Architecture

CSV files into JSONL format utilizing its data frame feature. Additionally, pandas facilitated the process of fetching each review for model testing and recording the results in a data frame.

For visual representation, we utilized matplotlib, sklearn, and seaborn to plot graphs and confusion matrices, showcasing the outcomes of our experiments. We leveraged the OpenAI library extensively, using an API key for model fine-tuning, monitoring training results, and deploying the fine-tuned model for testing. The Gradio library played a crucial role in crafting an appealing user interface, enabling real-time testing of the models. Lastly, we used the OS library to configure the environment with the necessary API key.

To evaluate the performance of the model we have taken 1200 samples of data consisting of 600 original reviews and 600 computer-generated reviews from the dataset. For the testing part, we have given 100 original reviews and 100 computer-generated reviews. We also defined some of the fake reviews given by LLMs and used those reviews for the test data set. Manually verifying the authenticity will produce high-quality ground truth labels. This smaller real-world dataset will provide an out-of-sample evaluation of real-time performance on live reviews from e-commerce sites.

First, the data set was divided into training and testing data sets, we preprocessed the training data set into JSONL format which is suitable for fine-tuning large language models. The format of the data set is given as prompts and completions where the prompts are the reviews of specific categories and completions are given were either computer-generated or genuine reviews. Then, the data set is given for fine-tuning the models, and the results are noted.

The data was converted from CSV format to JSONL format in terms of prompts and completion. After conversion, each data sample looks like: ”{”prompt”: ”I needed a long 8mm Allen wrench to get to the transmission fill plug on my ’11 Mustang 5. 0._n_n###_n_n”,”completion”:” True”}”. In the data sample, we can see the prompt is ending with ”_n_n###_n_n” which helps the model to recognize the end of each prompt.

```
import openai
import os
os.environ["OPENAI_API_KEY"] = "API_KEY"
!openai api fine_tunes.create \
      -t "reviews_1200_prepared_train.jsonl" \
      -v "reviews_1200_prepared_valid.jsonl" \
      --compute_classification_metrics \
      --classification_positive_class ' True' \
      -m model \
      --n_epochs 4
```

Fig. 3. Code snippet for fine-tuning base models

To measure the performance of our fine-tuning approach, we have defined several types of metrics. The metrics consisted of training loss, training accuracy, validation accuracy, validation loss, and classification accuracy. The data was then divided into training and validation sets which are used for fine-tuning purposes. The code in Figure 3 demonstrates how we fine-tune a model.

In the first step, we imported all the necessary packages like OpenAI to call the model and OS to set the environment API key which provides the gateway to access the model. Later we give fine_tunes.create command is used to create the model, specifying the training and validation files to be used for fine-tuning with -t and -v options, respectively. To optimize the model’s performance, we include hyper-parameters such as compute_classification_metrics and classification_positive_class=’true’. The -m option is used to specify which model we want to fine-tune, In our case we have used Ada and Davinci for fine-tuning. We have set epoch value as 4. After the fine-tuning is completed, each file will be assigned each job ID to the respective model and based on that job ID we can access the results of the training procedure.

The models generated after fine-tuning of base models ADA and Davinci are ada:ft-personal-2023-11-10-22-26-43 and davinci:ft-personal-2023-12-05-01-09-35 respectively. Here we have labeled ada_model and davinci_model as model names of the fine-tuned base models. Later we used those models for testing.

For fine-tuning of GPT-3.5 turbo aka ChatGPT, we have first preprocessed the data into JSONL formats, apart from the base models. The JSONL format consists of the role which specifies the role of the model and content of the specified role whether it can be a review or result or defining the purpose of the system. The sample of the dataset for fine-tuning GPT-3.5-turbo is given as a message in which role and

```
[ ]  from openai import OpenAI
     client = OpenAI()

     client.fine_tuning.jobs.create(
       training_file="file-ZMQlYjpNx9bhAPbj03Krqctl",
       model="gpt-3.5-turbo"
     )
```

Fig. 4. Code snippet for fine-tuning GPT-3.5-turbo

content are given: ”messages”: [”role”: ”system”, ”content”: ”Tell me whether the given review is original or computer generated, only give OR for the original review and CG for computer-generated review for the given review”, ”role”: ”user”, ”content”: ”This is a great quality light duty rain suit. Comfortable to wear, packs small, doesn’t look goofy. Sized very generously.”, ”role”: ”assistant”, ”content”: ”OR”].

After the JSONL file was obtained, we gave it as input for fine-tuning of GPT-3.5-turbo model a.k.a ChatGPT. We have given the epoch value as 3 which is the default value for fine-tuning. Here, for fine-tuning GPT-3.5-turbo we have different steps compared to the base model. Figure 4 shows how we first call the OpenAI() function and later we call client.fine_tuning.jobs.create() function in which we have to specify the training dataset file and model which we will train. After training, the a job ID is given which will determine the results of the training. The obtained model after fine-tuning is ”ft:gpt-3.5-turbo-0613:personal::8RuLxAwL” which we have labeled as model_ftgpt. We use the obtained model for testing.

Fine-tuning involves adapting the pre-trained model to our assigned task of fake review classification by using a small domain data set to continue the training process. The models can be fine-tuned in a supervised manner by using the labeled reviews to update the internal parameters that will be more suited for the task. This allows the model to learn nuanced patterns in language used specifically to fake reviews. We hypothesize that the fine-tuning method will produce an extremely accurate classification that surpasses the traditional approaches.

B. Context-Based Learning

In addition to fine-tuning, we also proposed a second method, context-based learning by providing useful metadata alongside the input review text. According to the related works, reviews do not exist in isolation but are complemented by contextual cues. Explicitly including such metadata as additional input to the model can potentially improve performance. For instance, an extremely positive review for a product that mostly has negative ratings is likely insincere.

Providing the rating distribution as auxiliary input can help the model learn such contextual patterns. The contextual information provided useful signals that are not contained in a review text alone. LLMs can effectively leverage these cues through multi-input architecture which in turn improves the effectiveness of the model and likely returns efficient results.

In our approach, we have specified the role of the system with the following prompt: ”You are a helpful assistant. Read the following knowledge context that begins and ends with ***. This context teaches you what a fake review is with some examples, so you can develop some intuition as to how to detect it. After assimilating and understanding the knowledge context, you will be provided with a series of reviews each of which will be preceded by $$$.

Based on the context provided, and for each of the reviews, the user will respond CG if you think the review is a fake and OR if it is not a fake review. The review will end with a %%% which will indicate that there is no more work to do.” Later we have specifies the context in between *** symbols by which the learns on how to detect whether the review is fake or original based on the provided context. We have taken different contexts and performed the experiments.

This will reveal the optimal context input that will provide the greatest accuracy boost. The performance was assessed by the accuracy improvement versus just using review text alone without any context. We aimed to determine the best contextual signals that can enhance deception detection when provided the auxiliary inputs of the model. The relative increase in accuracy from the metadata’s results will be the main criteria for evaluating the context-based method. Figure 2 depicts the proposed architecture of our project.

Our proposed approach differed significantly from existing ML techniques such as SVMs and Naive Bayes for fake review detection. Traditional models depend extensively on hand-crafted input features and domain expertise. The learner requires engineers to manually identify signals such as text complexity, sentiment polarity, syntactic patterns, etc. based on rules. In contrast, our neural network-based approach automatically learns useful feature representations directly from the input data.

This eliminated the need for manual feature engineering. Additionally, most existing methods only leveraged the review text and not supplementary metadata. Our context-based method uniquely encoded contextual cues along with the main review text. This allowed for the detection of deception signals that the text alone misses. Further, conventional models need large training datasets, often in the millions, to reach optimal accuracy. Our approach with pre-trained language models achieved high accuracy with orders of magnitude fewer labeled samples, enabling effective learning from minimal data.

Our proposed approach provided significant advantages over previous techniques for fake review detection. By leveraging large pre-trained language models like GPT-3, we required substantially less training data. The vast linguistic knowledge within these models enabled accurate deception detection with only hundreds or thousands of labeled samples, compared to the millions needed by traditional ML models. This made our method invaluable for low-resource domains. Additionally, incorporating contextual metadata from reviews improved performance without requiring more extensive networks. Our experiments will demonstrate clear accuracy gains over models lacking pre-training or context.

Furthermore, our techniques eliminated the manual effort of feature engineering through automatic representation learning. Our approach also requires less computation time and power for fine-tuning and testing compared to traditional models. For example, GPT-3.5 Turbo needs only $0.008 per 1,000 tokens for training and $0.012 per 1,000 for usage, while ADA is even more efficient at $0.0004 per 1,000 tokens. The combination of efficiency, accuracy, and context modeling provided significant advantages over existing fake review detection techniques.

Table 1 displays the performance of various machine learning models in the context of fake review detection across different datasets. Specifically, we utilized the Yelp dataset by comprising reviews collected from the website to train classical models like SVM, Naive Bayes, and Logistic Regression. Among these, Logistic Regression demonstrated the highest accuracy, achieving 78% on the Yelp dataset.

The performance analysis underlines that classical machine learning models such as SVM and Naive Bayes delivered reasonable but not outstanding accuracy when applied to real-world review data from Yelp. The accuracy rates ranged from 65% to 78%. Conversely, more intricate and contemporary techniques appeared to excel in the task of identifying fake reviews, as demonstrated by OpenAI and NBSVM. OpenAI and NBSVM achieved accuracy rates in the range of 83% to 95%. This observation suggests that deep learning and ensemble models may be better suited for the precise classification of fake reviews.

Furthermore, it is worth noting that the nature of the dataset played a pivotal role in model performance. Models generally performed better on synthetic fake review data compared to real-world scraped reviews, which often contained more noise. In summary for the performance analysis for the classical ML models, NBSVM demonstrated the highest accuracy on fake review data, Logistic Regression also performed reasonably well on Yelp reviews.

TABLE I
DESCRIBING THE PERFORMANCE OF EXISTING METHODS

Datasets	**ML Model**	**Fake Review Accuracy**
Yelp Dataset	SVM	77%
Yelp Dataset	Naive Bayes	65%
Yelp Dataset	Logistic Regression	78%

IV. EXPERIMENTS & RESULTS

For testing the obtained models, we took 200 data samples of unseen test datasets in which there are some reviews that we have generated using language models. Here we have tested 200 data samples of model_ada, model_davinci, model_ftgpt , a zero-shot model which is the ChatGPT model, and model with context inscribed along with the input.

V. RESULTS

By interpreting Figure 5 we can say that, the accuracies of the models are 50.5 % for the zero-shot model, and next, we got a model with metadata as an in-context learning model

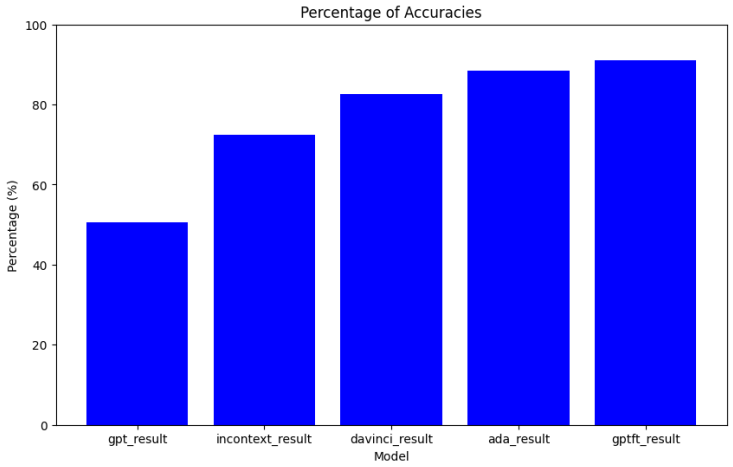

Fig. 5. Accuracies Of Models

and the accuracy of it is 72.5%, fine-tuning Davinci, Ada, and GPT-3.5-turbo resulted in the accuracies as 82.5 %, 88.5%, and 91% respectively.

For the real-time testing of the approaches we implemented a User interface to test the reviews get the result of each model and decide whether the review is fake or real. We have also given a flag feature where the testing results can be stored as log file for future reference.

We have also plotted a confusion matrix for the fine-tuned GPT-3.5-turbo model, by observing Figure 6 we can say that the model that has given the highest accuracy, has predicted one wrong as original where the correct value is computer generated and 11 wrong values as computer generated where the reviews are original.

When compared to classical machine learning models such as SVM and Naive Bayes, the fine-tuned and context-aware models demonstrated superior accuracy. Classical models like SVM, Naive Bayes, and Logistic Regression achieved accuracies between 65% to 78% on the Yelp dataset, whereas the fine-tuned OpenAI models showed significantly higher accuracies, indicating the effectiveness of deep learning and ensemble models in this domain.

VI. CONCLUSION & FUTURE WORKS

The primary focus of the study was to evaluate the efficacy of fine-tuning large pre-trained language models, such as GPT-3, and the implementation of context-based learning approaches. The research showed that these methods could significantly enhance the accuracy and efficiency of fake review detection systems.

The experiments conducted as part of this study employed a dataset comprising both original and computer-generated reviews. The fine-tuning of models like Ada, Davinci, and GPT-3.5-turbo, as well as the incorporation of context-based learning, were critical in achieving high classification accuracies. This approach not only reduced the reliance on extensive training datasets but also minimized the need for manual feature engineering. The study demonstrated that leveraging

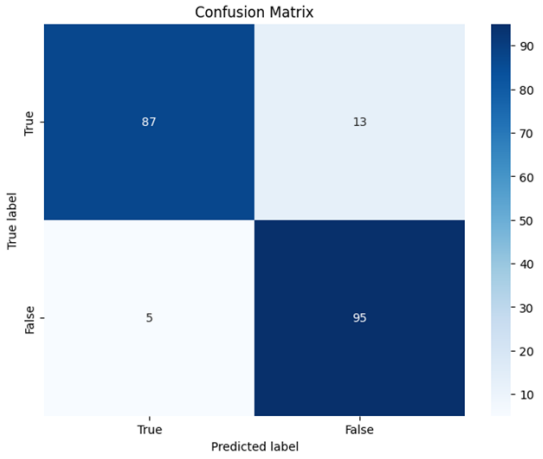

Fig. 6. Confusion Matrix for fine-tuned GPT-3.5-turbo model

the vast linguistic knowledge embedded within these models enables the accurate detection of fake reviews with substantially fewer labeled examples than traditional machine learning models.

It is crucial to expand testing to encompass a wider array of datasets, covering various domains and languages, to enhance the robustness and universal relevance of these methods. This will facilitate a more nuanced comprehension of model performance in different settings, contributing to the development of more globally effective solutions.

In the future, we would want to test out more ML methods while implementing fine-tuning and context learning on the following model types: VGG-19, clustering, and classification models. We would also want to enhance the performance of these models and see how they compare to our proposed methods' performance accuracy.

VII. ACKNOWLEDGEMENTS

We thank our colleagues from Texas A&M University Corpus-Christi who provided insight and made this project possible. We also thank Dr. Carlos Rubio-Medrano for assisting and overseeing our project.

REFERENCES

[1] Ana Costa, João Guerreiro, Sérgio Moro, Roberto Henriques, Unfolding the characteristics of incentivized online reviews, Journal of Retailing and Consumer Services, Volume 47, 2019, Pages 272-281, ISSN 0969-6989, https://doi.org/10.1016/j.jretconser.2018.12.006.

[2] Saumya Dixit, Anant Jyoti Badgaiyan, Arpita Khare, "An integrated model for predicting consumer's intention to write online reviews," Journal of Retailing and Consumer Services, Volume 46, 2019, Pages 112-120, ISSN 0969-6989, https://doi.org/10.1016/j.jretconser.2017.10.001.

[3] Yogesh K. Dwivedi, Elvira Ismagilova, D. Laurie Hughes, Jamie Carlson, Raffaele Filieri, Jenna Jacobson, Varsha Jain, Heikki Karjaluoto, Hajer Kefi, Anjala S. Krishen, Vikram Kumar, Mohammad M. Rahman, Ramakrishnan Raman, Philipp A. Rauschnabel, Jennifer Rowley, Jari Salo, Gina A. Tran, Yichuan Wang, "Setting the future of digital and social media marketing research: Perspectives and research propositions, International Journal of Information Management," Volume 59,2021,102168,ISSN 0268-4012, https://doi.org/10.1016/j.ijinfomgt.2020.102168.

[4] Gobi, N., Rathinavelu, A. Analyzing cloud-based reviews for product ranking using feature based clustering algorithm. Cluster Comput 22 (Suppl 3), 6977–6984 (2019). https://doi.org/10.1007/s10586-018-1996-3

[5] Harris, C.G. (2019). Comparing Human Computation, Machine, and Hybrid Methods for Detecting Hotel Review Spam. In: Pappas, I.O., Mikalef, P., Dwivedi, Y.K., Jaccheri, L., Krogstie, J., Mäntymäki, M. (eds) Digital Transformation for a Sustainable Society in the 21st Century. I3E 2019. Lecture Notes in Computer Science(), vol 11701. Springer, Cham. https://doi.org/10.1007/978-3-030-29374-1-7

[6] Qasim R, Bangyal WH, Alqarni MA, Ali Almazroi A. A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification. J Healthc Eng. 2022 Jan 7;2022:3498123. doi: 10.1155/2022/3498123. PMID: 35013691; PMCID: PMC8742153.

[7] G. Li, Z. Wang, M. Zhao, Y. Song and L. Lan, "Sentiment Analysis of Political Posts on Hong Kong Local Forums Using Fine-Tuned mBERT," 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022, pp. 6763-6765, doi: 10.1109/BigData55660.2022.10020704.

[8] L. Zhang and Y. Hu, "A fine-tuning approach research of pre-trained model with two stage," 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA), Shenyang, China, 2021, pp. 905-908, doi: 10.1109/ICPECA51329.2021.9362566.

[9] V. P. Sumathi, S. M. Pudhiyavan, M. Saran and V. N. Kumar, "Fake Review Detection Of E-Commerce Electronic Products Using Machine Learning Techniques," 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 2021, pp. 1-5, doi: 10.1109/ICAECA52838.2021.9675684.

[10] A. Sihombing and A. C. M. Fong, "Fake Review Detection on Yelp Data set Using Classification Techniques in Machine Learning," 2019 International Conference on contemporary Computing and Informatics (IC3I), Singapore, 2019, pp. 64-68, doi: 10.1109/IC3I46837.2019.9055644.

[11] https://www.smh.com.au/technology/samsung-fined-for-hiring-bloggers-to-write-fake-reviews-attack-rival-htc-20131025-2w5nx.html.

[12] Jia, Yuening. "Attention Mechanism in Machine Translation." Journal of Physics: Conference Series 1314 (2019): n. pag. DOI 10.1088/1742-6596/1314/1/012186

[13] Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, Bernard J. Jansen, Creating and detecting fake reviews of online products,Journal of Retailing and Consumer Services, Volume 64, 2022, 102771, ISSN 0969-6989, https://doi.org/10.1016/j.jretconser.2021.102771.