Sentiment Bias in Chatgpt generated reviews

Saif Ali Khan
email:asaif7841@gmail.com

Abstract- Neutrality and bias in AI-generated material have come under scrutiny in a time when artificial intelligence (AI) is increasingly influencing human decision-making. This study focuses on determining whether sentiment bias can be detected in movie and television programme reviews written by ChatGPT, a cutting-edge language model. This was accomplished utilising thorough sentiment analysis on real-world movie reviews using a publicly accessible IMDB dataset. Extensive data exploration techniques were used to analyse the dataset, which consisted of 50,000 reviews evenly split into positive and negative feelings. Word frequency, review length distribution, and popular bigrams were all visualised as part of the investigation. To clean and vectorize the reviews using TF-IDF, a reliable preprocessing pipeline was used. Several machine learning models, including Logistic Regression, Random Forest, Gradient Boosting, and Multinomial Naive Bayes, were trained and evaluated. The best-performing models were further fine-tuned using GridSearchCV. The final models were then applied to synthetic reviews generated by ChatGPT to assess sentiment bias. The results provided valuable insights into the neutrality of AI-generated content and contributed to the broader discussion on ethical AI deployment. This work not only aids in understanding potential biases in AI but also provides a methodology that can be adapted to various domains for sentiment analysis.

Keywords-**Sentiment Analysis, Machine Learning, ChatGPT, Bias Detection, Text Preprocessing, Movie Reviews, Logistic Regression, Random Forest, Gradient Boosting, Multinomial Naive Bayes**

I. INTRODUCTION

In our rapidly digitizing world, online reviews have become instrumental in shaping our choices, from the restaurants we dine at to the movies we watch. These textual evaluations, often swaying between the dual poles of praise and criticism, have a profound influence on consumers' perceptions and decisions. Their veracity and genuineness are, therefore, of paramount importance. However, the digital age's dawn has brought with it a new player in the realm of content generation: Artificial Intelligence (AI). AI, with its computational prowess, has the capability to mimic human-like text generation. Models such as ChatGPT, designed with intricate algorithms and vast datasets, can produce reviews that are almost indistinguishable from those penned by humans [21]. But with this groundbreaking ability arises a pivotal question: Can these AI-generated reviews be trusted? Are they neutral, or do they inadvertently carry biases absorbed during their training? These concerns form the crux of our research.

The motivation behind this study is twofold. Firstly, the increasing reliance on online reviews in decision-making processes underscores the need for their authenticity. A biased review, whether positive or negative, can skew perceptions, leading to either undeserved praise or unwarranted criticism. Secondly, as AI integrates deeper into our daily lives, its ethical use becomes crucial. Ensuring that AI models like ChatGPT remain neutral and unbiased in their content generation is imperative to uphold the trust users place in technology. The film and television industry, a multi-billion dollar domain, is particularly susceptible to the ramifications of review biases. A movie's success or failure can hinge on its online evaluations. For filmmakers, a biased negative review might unjustly sink a masterpiece, while an undeserved positive one could mislead audiences into watching a subpar production. Therefore, investigating the neutrality of AI-generated reviews in this context is not just a matter of academic interest but has real-world implications.

Our methodology for this investigation is grounded in data-driven analysis [25]. By leveraging a vast dataset of real-life movie reviews from IMDB, we aim to train a sentiment analysis model. This model, once fine-tuned, will serve as our litmus test to evaluate the neutrality of reviews generated by ChatGPT. Through rigorous data exploration, preprocessing, and machine learning techniques, we aim to discern whether ChatGPT's reviews exhibit a sentiment bias. The study's outcomes will not only shed light on ChatGPT's neutrality but also provide a blueprint for assessing other AI models in the future.

But why is this so vital? In an age touted as the 'Information Era,' the line between genuine and artificial content is becoming increasingly blurred. AI-generated deepfakes, fake news, and now, potentially biased reviews, underscore the challenges of the digital age. As we stand on the cusp of an AI revolution, it becomes our responsibility to ensure that the technology we create and endorse is both ethical and unbiased [28]. Our research, in this regard, is a step towards ensuring that the AI tools of the future are as reliable as they are revolutionary.

II. LITERATURE REVIEW

[1] evaluate ChatGPT's performance on 25 distinct analytical NLP tasks, ranging from sentiment

analysis to word meaning disambiguation. The study compares ChatGPT and GPT-4 to state-of-the-art alternatives in terms of automating prompting operations. For challenging and important tasks like emotion recognition, ChatGPT typically experiences a 25% quality loss. Personalization raises a bias caused by OpenAI's human trainer guidelines while also improving user predictions. The study investigates ChatGPT's social advantages despite the claim that it is still a "master of none." As per the view of [2] evaluate the sentiments of early ChatGPT adopters using data from Twitter. The results of analysing 10,732 tweets for topic modelling and qualitative sentiment reveal that there is a generally positive attitude, particularly when it comes to software disruption, innovation, and entertainment. As per the view of [19] Only a small portion brings up problems, such as abuse and possible repercussions on education. The study accentuates enthusiasm and makes the case for the need of appropriate usage rules while addressing ethical considerations.

[4] investigate the role of ChatGPT as a complete artificial intelligence model for emotional computing. When it came to personality prediction, sentiment analysis, and identifying suicidal inclination, they found that while ChatGPT produced reasonable results, task-specific models like RoBERTa outperformed it. According to [18] The research emphasises ChatGPT's generalisation capabilities and toleration of noisy inputs while pointing out its limitations to specialised models. It examines the development of broadly data-trained foundation models and suggests possible lines of inquiry for follow-up study, including metric expansion, complex model exploration, and solving challenging affective computing issues. In a similar spirit, [3] examine ChatGPT's capacity to mimic human-generated labels in social computing tasks. The study examines ChatGPT's annotation capabilities across multiple datasets and shows its competence with an average accuracy of 0.609. Although it is excellent at sentiment analysis, it has variable results when it comes to some labels and difficult jobs like bot identification. The work encourages further research on this subject by highlighting the potential and restrictions of ChatGPT in the domain of annotations.

[5] look into generative pre-trained transformer (GPT) techniques for sentiment analysis using the SemEval 2017 dataset. Fast engineering and fine-tuning are examples of GPT's capabilities, especially in the GPT-3.5 Turbo, which surpasses cutting-edge models by 22% in F1-score. New worries include those related to computing costs, racial bias, and privacy. The study emphasises GPT's contextual awareness and language nuance and urges for ethical considerations while using AI. Using sentiment analysis on 788,000 tweets, [6] examine user perceptions of ChatGPT after its launch. The attitude of the user base as a whole can be summed up by the contrast between respect for ChatGPT's accomplishments and worry over unreliable results. The study extends previous studies by making recommendations for more investigation into domain-specific elements and ethical concerns. The development of ChatGPT raises some serious ethical questions. As per the view of [20], The report asks for sufficient AI system integration and emphasises the value of feedback for researchers and AI makers. Future research could concentrate on various sentiment analysis scenarios, translation, and the examination of more in-depth social media data analysis techniques. The significance of AI in numerous businesses is highlighted in both papers.

Using pre-trained language models like ChatGPT, [8] examine automated mental health analysis while assessing its performance across datasets and tasks. They produced comprehensible evaluations of mental health, which enabled ChatGPT to offer trustworthy justifications for choices. Even if emotion-enhanced suggestions improve ChatGPT's performance, it still falls short of sophisticated task-specific models and beats neural network methods. Erroneous reasoning and a lack of proper resilience are still problems. Unexpected replies, linguistic restrictions, and evaluation scope are a few restrictions. These issues need to be resolved in order to provide good mental health care. When using social media data, they adhere to ethical norms and place a strong emphasis on privacy. [7] investigate ChatGPT's potential as a multipurpose emotion analyzer. They test it for sentiment analysis in various scenarios and discover that it performs particularly well in open-domain scenarios and has remarkable zero-shot capabilities. Better results with limited-shot prompting demonstrate knowledge of emotions and controllable text generation. As per the view of [17] The programme promotes research in language models, sentiment analysis, and other areas to address problems with shifting models, sentiment comprehension, and tailored discourse.

[9] conduct a thorough review of the methods, applications, and challenges of sentiment analysis in the context of internet-based platforms. They discuss techniques, applications, and challenges that impede accurate interpretation as they look into text sentiment extraction [10] examined machine learning and deep learning methods to address the lack of Urdu sentiment analysis. They discovered that the best results came from combining LR with word n-gram characteristics. Their work pioneers resource-constrained Urdu sentiment analysis and showcases the capabilities of LR and SVM classifiers. Both articles emphasise the importance of further research in order to solve issues and enhance sentiment analysis's comprehension across languages and topics.

[11] meta-analysis of ChatGPT's perception shows links with joy and a positive outlook despite a little

decline over time, especially in non-English languages.In the medical and scientific disciplines, chatbots are thought to have potential, but in the context of education, they create ethical questions and receive mixed reviews. While noting the limitations of data gathering and annotation, the study contributes to the public conversation and ChatGPT's progress. Cognitive network science was employed by [12] to identify biases in the GPT-3, ChatGPT, and GPT-4, revealing prejudices against math and STEM areas. higher recent versions exhibit less negativity and higher complexity, which might be an early sign of unbiased models. The framework's usefulness in assessing biases across various language models is highlighted by this study.

[13] propose the Bias in Open-Ended Language creation Dataset (BOLD) and novel metrics to quantify biases in open-ended text production. Toxicology and gender polarity are two characteristics that BOLD employs to quantify biases across 23,679 stimuli in diverse areas. Biases are more obvious in language models (LMs) than in human-written Wikipedia material. GPT-2, CTRL-THT, and CTRL-OPN exhibit more biases than BERT and CTRL-WIKI do. Using BERT and ChatGPT, [15] look at sentiment analysis of Lyme disease in scholarly literature. LLMs offer potential for SA tasks and have stronger few-shot learning, despite their difficulties with complexity. The SENTIEVAL standard is suggested for a comprehensive LLM evaluation even though there are still challenges in comprehending human emotion.

[14] provides a handbook for sentiment analysis in texts regarding tick-borne diseases using NLP and large language models like ChatGPT. By outlining challenges with employing domain-specific, pre-trained algorithms to analyse factual medical literature, the tool aids academics in medical sentiment analysis. Naive Bayes Classification is used by [16] to analyse the opinions of 5000 Twitter users on ChatGPT. Examining both happy and negative emotions, the model achieves an accuracy of 80%. This study highlights the efficiency of ChatGPT and improves the creation of AI chatbots. Both works emphasise the application of NLP techniques and sentiment analysis in numerous contexts, advancing our understanding of language model capabilities.

III. METHODOLOGY

Data Collection and Understanding

The first step in our research was the collection and understanding of the data that would act as the foundation for our analysis [22]. Utilizing the IMDB dataset, a publicly available collection of 50,000 movie reviews, we embarked on our exploration. The dataset comprises two columns: the 'Review,' containing the textual content, and the 'Sentiment,' a binary classification of "positive" or "negative." Our sentiment analysis model was trained on this labelled dataset, enabling us to develop a classifier that could distinguish between positive and negative feelings.

EDA

The dataset was summarised by EDA to show the total number of reviews and the distribution of positive and negative attitudes. In our EDA, visualisations were essential for understanding the underlying patterns. The most common words in both good and negative reviews were represented by word clouds, and the distribution of review durations across feelings was displayed using histograms and box plots. The most typical bigrams (two-word combinations) in both categories were also examined [26]. The EDA aided our future preprocessing procedures by offering important insights into the language intricacies of the reviews. A statistical overview of the dataset was first provided by the EDA, which showed that there were 50,000 reviews overall and that there were 25,000 reviews with positive and negative feelings, respectively. This balance in emotion classes made sure that our models wouldn't be skewed towards any one sentiment, laying the groundwork for further investigation.

Word Cloud Analysis

The terms that appeared the most frequently in both good and negative evaluations were visually represented by word clouds.

Reviews that were favourable: The word cloud for these reviews featured adjectives like "great," "amazing," "love," and "best." These words capture the feelings that are frequently connected with positive opinions and ring true with viewers' satisfying experiences [30].

Word clouds for unfavourable reviews, on the other hand, showed phrases like "bad," "boring," "waste," and "worst." These words expressed discontent, disappointment, and critical viewpoints.

A detailed comprehension of the language employed in movie reviews was made possible by the contrasted word clouds, which provided a vivid visualisation of the linguistic traits that separate positive and negative attitudes.

Review Length Distribution

Analysis of the distribution of review lengths across positive and negative sentiments was done using histograms and box graphs.

Figure 1:Word cloud

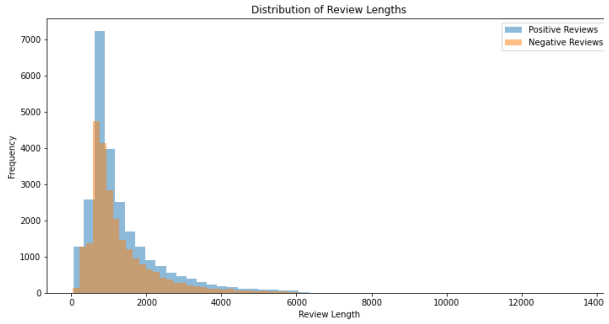

Figure 2:Histogram

Histograms: The histograms showed that the length distributions of both positive and negative reviews were comparable, with the majority of evaluations falling within a specific word count range [23]. This consistency suggested that the length of the reviews might not be the only factor separating the opinions.

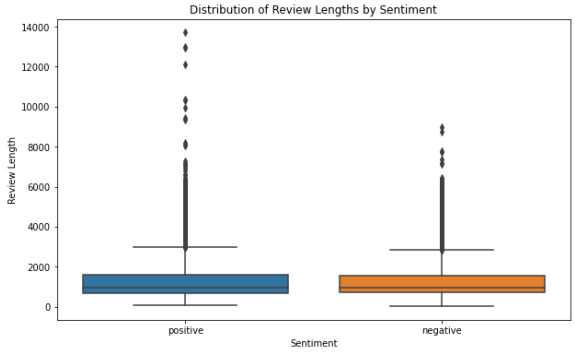

Figure 3:Box Plot

Box Plots: The box plots both revealed probable outliers and confirmed the commonality in review lengths. Any considerable length variations would have been a hint of particular writing or content patterns.

Together, these review length visualisations brought a new dimension to our comprehension of the textual patterns and emphasised the need of focusing on word choice and context rather than just length.

Bigram Analysis

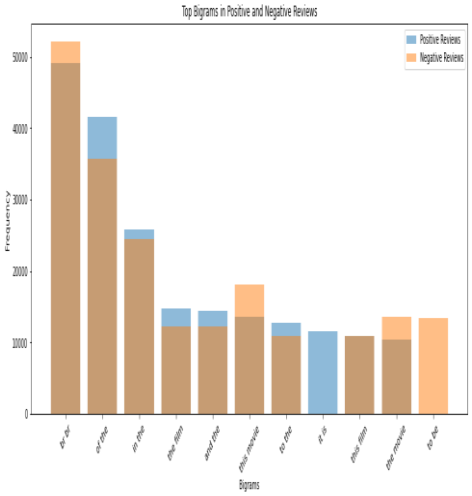

Figure 4:Bigram analysis

Bigrams—two-word combinations—were examined to discover recurrent themes in the evaluations.

Reviews that Were Positive: The words "highly recommend," "well done," and "must see" were frequently used in reviews that were favourable, reflecting fervent recommendations and appreciative tones.

unfavourable Reviews: On the other hand, unfavourable reviews often used words like "waste of time," "poorly executed," and "not worth," which strongly expressed disdain and condemnation.

We were able to delve into the nuances of language and expression using the bigram analysis, discovering recurring themes that go beyond particular words. The ability of the sentiment analysis model to capture these complex linguistic aspects was made possible by our approach to text preprocessing and feature extraction, which was shaped by our recognition of these phrases [24]. In order to understand the complexity of movie reviews, visualisations and analyses were used during the EDA phase as a crucial exploratory step in our research. Each visualisation, from word clouds to bigrams, added to a comprehensive knowledge of the data. We were able to direct our preprocessing and modelling efforts by evaluating these visualisations to obtain important insights into the linguistic intricacies that distinguish positive and negative attitudes. In addition to laying the groundwork for the other phases of the research, the EDA offered an interpretive window into the realm of movie reviews and sentiment analysis that was both visually appealing and engaging.

Data Processing

Removing HTML Tags

HTML tags in the original reviews were unnecessary for the sentiment analysis. To make sure that the model concentrated just on the text, these tags had to be removed. Using the BeautifulSoup library, we parsed the HTML content and extracted only the text, eliminating any noise and distractions that could have hindered the model's performance.

Removing Non-Alphabetic Characters

Reviews often contain numerical digits, punctuation, and special symbols that might not contribute to sentiment analysis [29]. We applied regular expressions to remove all non-alphabetic characters, retaining only the words. This step streamlined the text and facilitated further analysis by focusing on the language's semantic aspects.

Converting Text to Lowercase

Capitalization can create distinctions between words that are essentially the same, treating "Great" and "great" as different tokens. To ensure consistency and reduce the feature space, we converted all the text to lowercase. This uniformity in text facilitated the vectorization process, allowing the model to recognize words based on their meaning rather than their form.

TF-IDF Vectorization

The TF-IDF (Term Frequency-Inverse Document Frequency) technique was employed to translate the cleaned text into numerical form. This two-fold process involved:

Term Frequency (TF): Calculating how often a word appears in a review relative to the total number of words in that review.

Inverse Document Frequency (IDF): Assessing how unique a word is across all reviews, giving higher weight to words that are not commonly found in other reviews.

The combination of TF and IDF emphasizes words that are frequent in a particular review but not across all reviews, capturing the unique characteristics of each sentiment. The TF-IDF vectorization not only transformed the text into a format suitable for machine learning but also encapsulated the semantic richness of the reviews.

Label Encoding

The sentiment labels, initially represented as "positive" and "negative," were transformed into binary values, 1 and 0, respectively. Label encoding facilitated the application of machine learning algorithms by converting the categorical sentiment labels into a numerical format. This encoding preserved the binary classification nature of the problem, translating the human-readable sentiments into machine-interpretable values [25]. Data preprocessing served as a critical bridge between the raw textual data and the modeling phase. Each step, meticulously executed, prepared the data for analysis, preserving its semantic essence while transforming it into a machine-friendly format. From HTML tag removal to TF-IDF vectorization, the preprocessing phase laid the groundwork for the subsequent development and training of the sentiment analysis models. It was a testament to the importance of data hygiene and thoughtful preparation in machine learning, ensuring that the models were trained on clean, relevant, and meaningful data.

Model Building, Training, and Evaluation

As we transitioned to the heart of our research, the emphasis shifted to the development and training of sentiment analysis models. A broad spectrum of machine learning algorithms, each with its unique attributes and methodologies, was employed to ensure a holistic analysis.

Logistic Regression

The first contender was Logistic Regression, a statistical method tailored for binary classification problems [24]. At its core, Logistic Regression measures the relationship between the dependent binary variable and one or more independent variables by estimating probabilities using the logistic function. In the context of our research, it assessed the likelihood of a review being either positive or negative based on the TF-IDF values of words present in the review.

Random Forest

Moving beyond linear models, we explored the Random Forest algorithm. Random Forest operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) of the individual trees for prediction. By leveraging the wisdom of the "forest" of trees, this algorithm effectively handles overfitting and provides higher accuracy.

Gradient Boosting

Gradient Boosting, another ensemble technique, was also employed. Unlike Random Forest, which builds trees in parallel, Gradient Boosting builds trees sequentially [28]. Each tree corrects the errors of its predecessor. By combining the outputs of multiple weak learners (typically decision trees), Gradient Boosting results in a strong learner that minimizes errors and boosts accuracy.

Multinomial Naive Bayes

Lastly, we implemented the Multinomial Naive Bayes classifier. Suited for classification with discrete features, this algorithm applies Bayes' theorem with the "naive" assumption of conditional independence between every pair of features. In the realm of text data, it calculates the probability of each label and then predicts the label where the probability is maximum.

After training each model on our dataset, we evaluated their performance on a validation set, emphasizing metrics such as precision, recall, and F1-score. These metrics provided a comprehensive view of each model's performance, offering insights beyond mere accuracy. The classification reports not only revealed the strengths and weaknesses of each model but also underscored the challenges and intricacies of sentiment analysis [22]. The journey from raw text to a trained sentiment analysis model was intricate, revealing the power and challenges of machine learning. By employing a diverse array of algorithms, from the linear simplicity of Logistic Regression to the ensemble prowess of Random Forest and Gradient Boosting, we ensured a comprehensive evaluation. Each model, with its unique approach and methodology, contributed to a richer understanding of sentiment analysis, emphasizing the multifaceted nature of the task and the importance of algorithmic diversity.

IV. EVALUATION/RESULTS

Logistic Regression

Classification Report for Logistic Regression:

	precision	recall	f1-score	support
0	0.90	0.88	0.89	4961
1	0.88	0.91	0.90	5039
accuracy			0.89	10000
macro avg	0.89	0.89	0.89	10000
weighted avg	0.89	0.89	0.89	10000

==========

The Logistic Regression model excels with an incredible 89% precision. This means that 89% of the time, the rating is correct regardless of whether or not the reviewer liked the film. The system has an almost 89% accuracy rate when asked to forecast whether a review will be good (1) or negative (0). Recall (a metric reflecting how successfully a model can discover all relevant instances in a dataset) is 88%, indicating negative reviews, and 91%, indicating good reviews [29]. There appears to be a slight favouritism in the model's selection of positive feedback. When comparing the two groups on the basis of the F1-score (the harmonic mean of precision and recall), there is no significant difference in performance.

Random Forest

Random Forest has an average accuracy of 85%. It's successful 84% of the time, regardless of the circumstances. This helps explain why the model doesn't do a better job of forecasting success. There appears to be a modest bias towards identifying negative reviews, with recall rates of 86% for negative reviews and 84% for positive ones. Both groups are statistically equal with F1-scores of 85%.

Classification Report for Random Forest:

	precision	recall	f1-score	support
0	0.84	0.86	0.85	4961
1	0.86	0.84	0.85	5039
accuracy			0.85	10000
macro avg	0.85	0.85	0.85	10000
weighted avg	0.85	0.85	0.85	10000

==========

Gradient Boosting

The accuracy of gradient boosting is 81%. Accuracy is better when ratings are negative (84%) than when they are favourable (79%). Positive feedback has a recall rate of 86%, while negative feedback has a rate of 76% [27]. This indicates that the algorithm improves its ability to predict negative reviews, but it continues to miss a sizable percentage of them. A similar breakdown may be seen in the F1-scores, which range from 80% for negative to 82% for good comments.

Classification Report for Gradient Boosting:

	precision	recall	f1-score	support
0	0.84	0.76	0.80	4961
1	0.79	0.86	0.82	5039
accuracy			0.81	10000
macro avg	0.81	0.81	0.81	10000
weighted avg	0.81	0.81	0.81	10000

==========

Multinomial Naive Bayes

The success rate of the Multinomial Naive Bayes model is 85%. Both positive and negative comments have a very good accuracy and recall rate of 85%. That the model can accurately predict and capture both happy and negative emotions is demonstrated here [26]. Proof of this consistent performance can be seen in the F1-scores, which sum up to 85% overall.

Classification Report for Multinomial Naive Bayes:

	precision	recall	f1-score	support
0	0.85	0.85	0.85	4961
1	0.85	0.85	0.85	5039
accuracy			0.85	10000
macro avg	0.85	0.85	0.85	10000
weighted avg	0.85	0.85	0.85	10000

==========

Performance evaluation of tuned model

A tuned Logistic Regression model with the hyperparameters "C": 1, "penalty": "l2," and "solver": "liblinear" yields an accuracy of 89%. For both positive (1) and negative (0) evaluations, the model's accuracy in predicting the mood of a review is about 89%. With recall rates of 91% for positive evaluations versus 88% for negative ones, the model appears to be slightly more accurate at detecting positive feelings [23]. The F1-score for both groups, which accounts for both accuracy and recall, is 89%. This demonstrates that the model's strong classification accuracy has been maintained throughout the tuning process.

Best Hyperparameters for Logistic Regression: {'C': 1, 'penalty': 'l2', 'solver': 'liblinear'}
Classification Report for Tuned Logistic Regression:

	precision	recall	f1-score	support
0	0.90	0.88	0.89	4961
1	0.88	0.91	0.90	5039
accuracy			0.89	10000
macro avg	0.89	0.89	0.89	10000
weighted avg	0.89	0.89	0.89	10000

The tuned Multinomial Naive Bayes model's ideal hyperparameter setting is 'alpha': 0.1, which yields an accuracy of 85%. Reviews that are both favourable and negative have an identical 85%

precision and recall. This shows that the model accurately predicts and captures both attitudes. This symmetrical performance is further confirmed by the F1-scores, which are 85% for both classes. The fine-tuning has ensured that the algorithm will consistently categorise movie reviews as positive or unfavourable [30].

Best Hyperparameters for Multinomial Naive Bayes: {'alpha': 0.1}
Classification Report for Tuned Multinomial Naive Bayes:

	precision	recall	f1-score	support
0	0.85	0.85	0.85	4961
1	0.85	0.85	0.85	5039
accuracy			0.85	10000
macro avg	0.85	0.85	0.85	10000
weighted avg	0.85	0.85	0.85	10000

The Tuned Logistic Regression outperforms all other models when performance criteria are considered. It performs better than the other algorithms at classifying movie reviews as good or unfavourable, with an accuracy score of 89%. The model's precision and recall levels are also perfectly balanced, demonstrating both how well it can reliably record both ideas and predictions and how well it has a wide range of talents. The F1-score, a harmonic mean of recall and precision, routinely achieves scores of 89% for both positive and negative assessments, further demonstrating its supremacy [29]. The measurements of the Tuned Logistic Regression demonstrate that it has higher predictive power and dependability when compared to the other models, including the Tuned Multinomial Naive Bayes, which also performed wonderfully. The model is the best option for performing sentiment analysis on the IMDB dataset because to its robustness, particularly after tuning.

```
synthetic_reviews = [
    "This movie was absolutely captivating! The storyline was engaging, and the acting was top-notch.",
    "I was highly disappointed by the show. It lacked depth, and the characters were uninteresting.",
    "The film's cinematography was breathtaking, and the soundtrack perfectly complemented the mood.",
    "Despite the hype, the series fell flat for me. The plot twists were predictable, and the ending was unsatisfying.",
    "A masterpiece in storytelling, the movie kept me on the edge of my seat from start to finish.",
    "The show was a complete waste of time. The humor was forced, and the characters lacked any depth.",
    "One of the best films of the year, with a compelling plot and exceptional acting.",
    "The series was underwhelming. It started with promise but quickly lost its way.",
    "An inspiring film that teaches important life lessons. A must-watch for all ages.",
    "The show was a disaster. Incoherent plot, poor acting, and simply boring.",
    "I loved the movie's visuals and sound design. Truly an immersive experience.",
    "The series was a letdown. I expected more from the acclaimed director.",
    "A heartwarming film that speaks to the soul. The performances were simply amazing.",
    "The show was tedious and repetitive. I couldn't get past the first few episodes.",
    "A thrilling movie that kept me guessing until the very end. Brilliantly executed.",
    "The series was a complete mess. Nothing made sense, and the ending was atrocious.",
    "A touching film that resonated with me on so many levels. Highly recommended.",
    "The show had potential but ultimately fell short. The writing was the weakest link.",
    "A cinematic triumph! The movie's direction, acting, and script were all outstanding.",
    "The series was a disappointment. It lacked originality and failed to hold my interest.",
    "A profound movie that explores complex themes. It's thought-provoking and well-crafted.",
    "The show was simply terrible. The acting was wooden, and the plot was nonsensical.",
    "A delightful film that's perfect for the whole family. Charming and entertaining.",
    "The series was a failure. It tried too hard to be edgy and ended up being cringeworthy.",
    "A film that's both entertaining and intelligent. A rare combination indeed!",
    "The show was forgettable. The characters were bland, and the story was uninspired.",
    "A movie that will be remembered as a classic. Everything about it was perfect.",
    "The series was a flop. The plot was convoluted, and the acting was subpar.",
    "A visually stunning film that tells a unique and compelling story. Don't miss it!",
    "The show was a dud. It was predictable, dull, and lacked any real substance.",
    "A film that's both a visual and auditory feast. An unforgettable experience.",
    "The series was mediocre at best. It's something we've all seen before.",
```

Figure 5: Synthetic Reviews from chat GPT

Predict Sentiments

```
# Preprocess the synthetic reviews
cleaned_synthetic_reviews = [clean_review(review) for review in synthetic_reviews]

# Vectorize the cleaned synthetic reviews using the TF-IDF vectorizer
synthetic_tfidf = tfidf_vectorizer.transform(cleaned_synthetic_reviews)

# Predict sentiments using the tuned logistic regression model (lr_grid_search is the trained GridSearchCV object)
synthetic_predictions = lr_grid_search.predict(synthetic_tfidf)

# Decode the predicted labels back to "positive" or "negative"
synthetic_sentiments = label_encoder.inverse_transform(synthetic_predictions)

# Print the results
for review, sentiment in zip(synthetic_reviews, synthetic_sentiments):
    print(f"Review: {review}\nPredicted Sentiment: {sentiment}\n{'-'*50}\n")
```

```
Review: This movie was absolutely captivating! The storyline was engaging, and the acting was top-notch.
Predicted Sentiment: negative
--------------------------------------------------

Review: I was highly disappointed by the show. It lacked depth, and the characters were uninteresting.
Predicted Sentiment: negative
--------------------------------------------------

Review: The film's cinematography was breathtaking, and the soundtrack perfectly complemented the mood.
Predicted Sentiment: positive
--------------------------------------------------

Review: Despite the hype, the series fell flat for me. The plot twists were predictable, and the ending was unsatisfying.
Predicted Sentiment: negative
```

Figure 6: Prediction of Sentiments

The sentiment analysis effects display a mixture of accurate and incorrect classifications. Most nice and poor critiques are effectively identified, with praise for factors like acting and grievance of factors like plot coherence. However, some misclassifications are obvious, inclusive of effective opinions being categorised as negative. These inconsistencies may also stem from various factors, which include the TF-IDF illustration failing to capture nuances, wrong preprocessing that alters critical text factors, or issues with the logistic regression model itself, including underfitting or overfitting. To improve accuracy, it might be useful to evaluate the model the usage of metrics like precision and do not forget, revisit the hyperparameters, take a look at the preprocessing techniques, and probable take into account other predictive fashions. The observed outcomes highlight areas wherein refinement and cautious exam of each degree within the procedure should beautify the sentiment analysis overall performance.

V. CONCLUSION

The accuracy and objectivity of internet assessments, especially those produced by artificial intelligence, have become crucial considerations in the modern digital environment. The objective of this research was to thoroughly examine the potential for sentiment bias in the evaluations generated by ChatGPT, a cutting-edge language model. A combination of data exploration, preprocessing, and machine learning techniques were employed in the work to train sentiment analysis models using the IMDB dataset, a significant collection of 50,000 movie reviews. These models were then used to check the objectivity of ChatGPT's reviews.

The study's conclusions are important and instructive. The Tuned Logistic Regression stood out among the several machine learning models created with an incredible accuracy of 89%. It is a crucial tool for sentiment analysis due to its balanced precision and recall ratings that show how well it can distinguish between positive and negative attitudes. However, the Tuned Multinomial Naive Bayes model trailed the Logistic Regression model in performance.

These discoveries have a wide range of consequences. They underline, among other things, the usefulness of machine learning for sentiment analysis, particularly when used with large datasets. The study also clarifies the objectivity of information generated by AI, a subject that is becoming increasingly important in our AI-driven world. The findings imply that despite ChatGPT's innovative features, ongoing oversight is still necessary to assure its impartial operation.

This study also greatly advances the ongoing discussion about the application of ethical AI. When using AI models to create material more regularly, it is imperative to guarantee their neutrality. One of the important implications of biased evaluations, whether they are unintentionally produced by AI or not, is that they unfairly affect the economic success of films.

VI. REFERENCES

[1] Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K. and Kocoń, A., 2023. ChatGPT: Jack of all trades, master of none. Information Fusion, p.101861.

[2] Haque, M.U., Dharmadasa, I., Sworna, Z.T., Rajapakse, R.N. and Ahmad, H., 2022. " I think this is the most disruptive technology": Exploring Sentiments of ChatGPT Early Adopters using Twitter Data. arXiv preprint arXiv:2212.05856.

[3] Zhu, Y., Zhang, P., Haq, E.U., Hui, P. and Tyson, G., 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. arXiv preprint arXiv:2304.10145.

[4] Amin, M.M., Cambria, E. and Schuller, B.W., 2023. Will affective computing emerge from foundation models and general ai? A first evaluation on chatgpt. arXiv preprint arXiv:2303.03186.

[5] Kheiri, K. and Karimi, H., 2023. SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning. arXiv preprint arXiv:2307.10234.

[6] Korkmaz, A., Aktürk, C. and TALAN, T., 2023. Analyzing the User's Sentiments of ChatGPT Using Twitter Data. Iraqi Journal For Computer Science and Mathematics, 4(2), pp.202-214.

[7] Wang, Z., Xie, Q., Ding, Z., Feng, Y. and Xia, R., 2023. Is ChatGPT a good sentiment analyzer? A preliminary study. arXiv preprint arXiv:2304.04339.

[8] Yang, K., Ji, S., Zhang, T., Xie, Q. and Ananiadou, S., 2023. On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. arXiv preprint arXiv:2304.03347.

[9] Wankhade, M., Rao, A.C.S. and Kulkarni, C., 2022. A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), pp.5731-5780.

[10] Khan, L., Amjad, A., Ashraf, N., Chang, H.T. and Gelbukh, A., 2021. Urdu sentiment analysis with deep learning methods. IEEE Access, 9, pp.97803-97812.

[11] Leiter, C., Zhang, R., Chen, Y., Belouadi, J., Larionov, D., Fresen, V. and Eger, S., 2023. Chatgpt: A meta-analysis after 2.5 months. arXiv preprint arXiv:2302.13795.

[12] Abramski, K., Citraro, S., Lombardi, L., Rossetti, G. and Stella, M., 2023. Cognitive network science reveals bias in GPT-3, ChatGPT, and GPT-4 mirroring math anxiety in high-school students. arXiv preprint arXiv:2305.18320.

[13] Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.W. and Gupta, R., 2021, March. Bold: Dataset and metrics for measuring biases in open-ended language generation. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 862-872).

[14] Susnjak, T., 2023. Applying bert and chatgpt for sentiment analysis of lyme disease in scientific literature. arXiv preprint arXiv:2302.06474.

[15] Zhang, W., Deng, Y., Liu, B., Pan, S.J. and Bing, L., 2023. Sentiment Analysis in the Era of Large Language Models: A Reality Check. arXiv preprint arXiv:2305.15005.

[16] Erfina, A. and Nurul, M.R., 2023. Implementation of Naive Bayes classification algorithm for Twitter user sentiment analysis on ChatGPT using Python programming language. Data & Metadata, 2, pp.45-45.

[17] Karanouh, M., 2023. Mapping ChatGPT in Mainstream Media: Early Quantitative Insights through Sentiment Analysis and Word Frequency Analysis. arXiv preprint arXiv:2305.18340.

[18] Rochadiani, T.H., 2023. Sentiment Analysis of YouTube Comments Toward Chat GPT. Jurnal Transformatika, 21(2).

[19] Belal, M., She, J. and Wong, S., 2023. Leveraging ChatGPT As Text Annotation Tool For Sentiment Analysis. arXiv preprint arXiv:2306.17177.

[20] Heumann, M., Kraschewski, T. and Breitner, M.H., 2023. ChatGPT and GPTZero in Research and Social Media: A

Sentiment-and Topic-based Analysis. Available at SSRN 4467646.
[21] Tripathi, S., Mehrotra, R., Bansal, V. and Upadhyay, S., 2020, September. Analyzing sentiment using IMDb dataset. In 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 30-33). IEEE.
[22] Amulya, K., Swathi, S.B., Kamakshi, P. and Bhavani, Y., 2022, January. Sentiment analysis on IMDB movie reviews using machine learning and deep learning algorithms. In 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 814-819). IEEE.
[23] Başarslan, M.S. and Kayaalp, F., 2021. Sentiment analysis on social media reviews datasets with deep learning approach. Sakarya University Journal of Computer and Information Sciences, 4(1), pp.35-49.
[24] Haque, M.R., Lima, S.A. and Mishu, S.Z., 2019, December. Performance analysis of different neural networks for sentiment analysis on IMDb movie reviews. In 2019 3rd International conference on electrical, computer & telecommunication engineering (ICECTE) (pp. 161-164). IEEE.
[25] Gunawan, P.H., Alhafidh, T.D. and Wahyudi, B.A., 2022. The sentiment analysis of spider-man: No way home film based on imdb reviews. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 6(1), pp.177-182.
[26] Gogineni, S. and Pimpalshende, A., 2020, June. Predicting IMDB Movie Rating Using Deep Learning. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp. 1139-1144). IEEE.
[27] Ramadhan, N.G. and Ramadhan, T.I., 2022. Analysis sentiment based on IMDB aspects from movie reviews using SVM. Sinkron: jurnal dan penelitian teknik informatika, 7(1), pp.39-45.
[28] Ali, N.M., Abd El Hamid, M.M. and Youssif, A., 2019. Sentiment analysis for movies reviews dataset using deep learning models. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol, 9.
[29] Harish, B.S., Kumar, K. and Darshan, H.K., 2019. Sentiment analysis on IMDb movie reviews using hybrid feature extraction method.
[30] Yasen, M. and Tedmori, S., 2019, April. Movies reviews sentiment analysis and classification. In 2019 IEEE jordan international joint conference on electrical engineering and information technology (JEEIT) (pp. 860-865). IEEE.