Forecasting Bitcoin Price Trends: Leveraging Natural Language Processing and Bing GPT Data Augmentation for Enhanced Predictive Insights

1st Stepan Tytarenko
Graduate Student Researcher
Fordham Graduate School of Arts and Sciences
New York, USA
stytarenko@fordham.edu

1st Kibru Temesgen Lefebo
Graduate Student Researcher
Fordham Graduate School of Arts and Sciences
New York, USA
klefebo@fordham.edu

***Abstract*—The volatility of cryptocurrency markets poses a significant challenge for investors and traders. This study explores the application of Natural Language Processing (NLP) techniques, specifically leveraging transformer models, to predict the impact of news articles on Bitcoin prices. We have a small dataset that was manually labeled for us by a domain expert (a person doing business and trading in crypto). Using this small, manually labeled dataset of 1800 news pieces, categorized as positively impactful, negatively impactful, or not important, this research aims to expand the dataset by automatically labeling an additional 1000 news articles using a custom Bing API and Bing GPT. The study focuses on the fine-tuning of BERT, a pre-trained transformer model, to predict the impact of news on Bitcoin prices. By addressing the limitations of manual labeling through automated methods, this research attempts to establish a proof of concept for predicting cryptocurrency price movements based on news sentiment analysis. We also provide an open-source GitHub[1] repository with all of the experiments that we have run, for reproduction and improvement.**

***Index Terms*—NLP, BERT, GPT, Cryptocurrency**

I. Introduction

Introduction The cryptocurrency market has emerged as a transformative force in the global financial landscape, offering a decentralized and democratized alternative to traditional financial systems. However, the inherent volatility of cryptocurrency prices poses a significant challenge for investors and traders. Accurately forecasting price movements remains a complex and elusive endeavor, with traditional technical analysis and fundamental analysis often proving inadequate.

In this context, Natural Language Processing (NLP) techniques have emerged as a promising avenue for enhancing cryptocurrency price prediction. NLP enables the extraction of meaningful insights from unstructured text data, such as news articles, social media posts, and online forums. By analyzing the sentiment and tone of these textual sources, NLP models can potentially uncover valuable information that can inform investment decisions.

This study delves into the application of NLP techniques, specifically leveraging transformer models, to predict the impact of news articles on Bitcoin prices. Transformer models, such as BERT [1], have revolutionized NLP by demonstrating remarkable capabilities in tasks such as language translation, text summarization, and question answering. Their ability to capture long-range dependencies and contextual relationships in the text makes them well-suited for sentiment analysis and predicting the impact of news on cryptocurrency prices. To address the limitations of manual labeling, which is a time-consuming and resource-intensive process, this research proposes an automated approach to labeling news articles. Utilizing the Bing API and Bing GPT, a large language model from Bing AI, we aim to expand our dataset by automatically labeling an additional 1000 news articles. We would like to see if this automated labeling approach will significantly enhance the scope and effectiveness of our NLP models.

The study commences by establishing a benchmark result using classical machine learning algorithms like KNN, decision trees, random forests, and Adaboost. These algorithms, while well-established in their own right, often struggle to capture the nuances and complexities of natural language. Our primary focus lies in fine-tuning BERT pre-trained transformer models to predict the impact of news on Bitcoin prices. The fine-tuning process involves adjusting the models' parameters to optimize their performance on our expanded dataset of manually and automatically labeled news articles. We also compare our results with a much bigger and more powerful transformer model from OpenAI called Ada, to see how it behaves in the identical scenario. By leveraging NLP techniques and transformer models, this research endeavors to establish a proof of concept for predicting cryptocurrency price movements based on news sentiment analysis. The successful application of NLP could revolutionize cryptocurrency trading, providing investors with valuable insights to make informed investment decisions and navigate the volatile cryptocurrency market with greater confidence.

[1] https://github.com/StepanTita/crypto-text-classification

II. Background

Cryptocurrency, epitomized by its flagship currency, Bitcoin, has revolutionized the financial landscape, ushering in a decentralized paradigm that transcends traditional banking systems. The allure of digital currencies lies not only in their potential for rapid value appreciation but also in their susceptibility to substantial price fluctuations, making them a fertile ground for investment and speculation. However, this volatility is a double-edged sword, posing formidable challenges for investors and traders seeking to comprehend and capitalize on market dynamics.

Amidst this ever-evolving landscape, the role of information dissemination and its impact on cryptocurrency prices have garnered significant attention. News, often disseminated through various media channels, holds the potential to sway market sentiment, triggering notable price movements in the realm of cryptocurrencies. The inherent interconnectedness between news sentiment and price fluctuations has spurred researchers and market participants alike to explore methodologies that could effectively predict and leverage these relationships.

Natural Language Processing (NLP) has emerged as a formidable tool for unraveling the intricate relationship between news sentiment and cryptocurrency price movements. Leveraging NLP techniques, particularly transformer models such as BERT and Curie, presents a promising avenue for discerning the impact of news articles on cryptocurrency markets. These models, adept at processing and understanding textual data, offer a nuanced comprehension of language semantics and sentiment, enabling the extraction of valuable insights from vast repositories of textual information.

Traditionally, efforts to predict cryptocurrency price movements have encountered challenges stemming from limited and often subjective datasets. Manual labeling, despite its efficacy in certain scenarios, presents constraints in scale and objectivity. This limitation necessitates the exploration of automated labeling methods that not only expand dataset sizes but also mitigate the inherent biases associated with human annotations. By surmounting these challenges through automated labeling mechanisms, this research aims to pave the way for more robust and scalable methodologies for predicting cryptocurrency price fluctuations based on news sentiment analysis.

In light of the complexities surrounding cryptocurrency markets and their susceptibility to external influences, this study endeavors to explore the fusion of NLP techniques with machine learning models to navigate and forecast market behavior. This fusion aims to empower stakeholders with more informed decision-making capabilities, potentially mitigating risks and maximizing opportunities in the volatile terrain of cryptocurrency trading. Through a comprehensive exploration of NLP-driven sentiment analysis and its implications on cryptocurrency markets, this research seeks to contribute to the evolving discourse on predictive analytics in the financial technology domain.

III. Related Work

Related Works Recent advances in pre-trained language models (PLMs) have opened up new possibilities for automating the labeling of extensive datasets, a critical step in developing machine learning models. Label functions (LFs) are heuristic data sources that provide weak supervision signals for training machine learning models. Traditionally, creating accurate LFs has required domain expertise and substantial effort. However, PLMs have demonstrated the potential to learn from large amounts of text data and extract meaningful patterns, making them well-suited for generating LFs. Several studies have explored the use of PLMs for LF generation. For instance, [2] proposed an interactive framework called DataSculpt that utilizes PLMs to generate LFs from natural language descriptions of data samples. Their study demonstrated that DataSculpt can generate accurate LFs for a variety of tasks, but further research is needed to refine the capabilities of PLMs in this domain and address the challenges associated with their application to complex and diverse tasks.

Paper [3] explores the challenge of obtaining large-scale relevance labels for search systems and proposes an innovative approach to improving label quality using large language models (LLMs). Traditionally, acquiring relevant labels from real users through feedback is deemed the highest-quality data but is limited in scalability. Current practices rely on third-party labelers, risking low-quality data due to potential misunderstandings of user needs. The paper introduces an alternative methodology leveraging LLM prompts based on high-quality first-party user feedback. By deploying language models for relevance labeling at Bing and drawing insights from TREC data, the study demonstrates that LLMs can match human accuracy levels and exhibit a similar capability in selecting complex queries, top-performing results, and optimal groups. Notably, systematic changes and even simple paraphrases in LLM prompts significantly impact accuracy. Results show that with high-quality ”gold” labels from real users, LLMs outperform third-party workers in producing superior labels at a reduced cost, leading to enhanced training for better search result rankings.

Recent advancements in pre-trained language models (PLMs) have revolutionized the automation of dataset labeling, a pivotal step in machine learning model development. Label functions (LFs) and heuristic data sources offering weak supervision signals traditionally demanded expertise and effort for accurate creation. However, PLMs, adept at learning from vast text data and extracting meaningful patterns, show promise in LF generation. [4] introduced DataSculpt, an interactive framework utilizing PLMs to generate accurate LFs from natural language descriptions, demonstrating efficacy across diverse tasks. While showing potential, further research is needed to refine PLMs' capabilities for complex tasks. [5] address the challenge of obtaining large-scale relevance labels for search systems, proposing an innovative approach leveraging large language models (LLMs) for label quality enhancement. Traditional reliance on real user feedback for high-quality data faces scalability limitations, leading to third-party labeler usage and potential data quality issues. By deploying LLMs for relevance labeling and drawing from TREC data, the study reveals LLMs' capability to match human accuracy in selecting complex queries and top-performing results. Systematic changes in LLM prompts notably impact accuracy, showcasing their superiority over third-party workers in producing high-quality labels at reduced costs and enhancing training for improved search rankings.

Recent advancements in natural language processing have underscored the significance of language model pre-training as a foundational approach to acquiring comprehensive and universal language representations. Among the forefront of such pre-training models stands BERT (Bidirectional Encoder Representations from Transformers), renowned for its breakthroughs in various language understanding tasks [6]. The authors embark on an exhaustive exploration through a series of meticulous experiments aimed at dissecting and evaluating diverse fine-tuning methodologies specifically tailored for BERT's application in text classification tasks. By delving into the intricacies of fine-tuning strategies, the research endeavors to unravel optimal approaches and synthesize these findings into a comprehensive, adaptable solution for fine-tuning BERT. The culmination of these endeavors leads to a novel methodology that not only refines BERT's performance but also yields pioneering results, establishing new benchmarks on eight widely studied text classification datasets. This robust and versatile solution marks a significant advancement in the realm of language model fine-tuning, showcasing its efficacy in pushing the boundaries of text classification performance.

IV. Methodology

A. Dataset

We employ two datasets in this study. The first dataset comprises a collection of news articles related to Bitcoin and its price movements. The data was gathered from multiple international news outlets and contains articles published between December 1, 2022, and December 31, 2022. The dataset consists of approximately 1,800 news articles, each characterized by the following features: Source, Title, Publication date, currencies, news URL, description, and Label To ensure the quality and relevance of the data, the following criteria were applied during the data collection process: News Source: Only reputable and established news sources were considered to minimize the inclusion of biased or unreliable information.

The second dataset is an unlabeled dataset. It consists of 18545 crypto-related news. The dataset contains data from 3 sources: cryptonews.com, cryptopotato.com, and cointelegraph.com, in the period between December 2021 and May 2023. We are using only a small subset of this data from the timeframe that is aligned with the one we use in the first dataset. We use BING GPT with Google search to label a subset of 900 data entries. Due to technical limitations, it is very difficult to label more data than that.

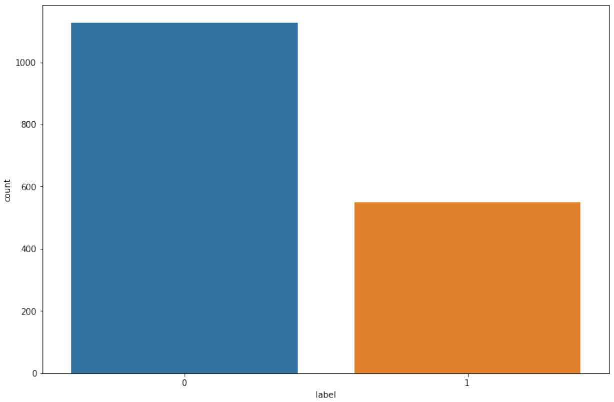

Fig. 1. Labeled Data Balance

News Relevance: Articles were selected if they specifically focused on Bitcoin and its price movements, excluding articles with only tangential mentions of Bitcoin.

News Timeframe: The data collection timeframe was restricted to the month of December 2022 to capture recent news events and their impact on Bitcoin's price during that period.

News Labeling: The news value label (1.0, -1.0, or 0.0) was assigned by a domain expert with extensive experience in cryptocurrency trading and news analysis to ensure the accuracy and consistency of the labels. However, for the purpose of this research, we are going to focus on two possible labels: 0 - no impact on the price and 1 - some impact on the price (either negative or positive). In Figure 1, we provide a balance of the target label before adding augmented data.

B. Data preprocessing

Since we are dealing with textual data, we deal with a lot of noise. Due to the nature of this data retrieval process, it contains a lot of artifacts and errors. Specifically, duplicated records, HTML tags due to parsing issues, typos, and some more. We not only address these problems but also try various preprocessing techniques to see which one works the best. The first key step is removing HTML tags since these carry no value to the classification problem. Next, on the clean data, we employ 3 strategies to see the one giving the best outcome.

- remove all the punctuation and stopwords
- remove all the punctuation, stopwords, and lemmatize
- only employ named entity extraction

Besides we also analyze the most frequently used words and bigrams, as shown in Figure 2 and Figure 3.

Eventually, we conclude that for the OpenAI Ada and BERT, it is best to just keep the data as it is, without steps like stopwords and punctuation removal. This is due to the

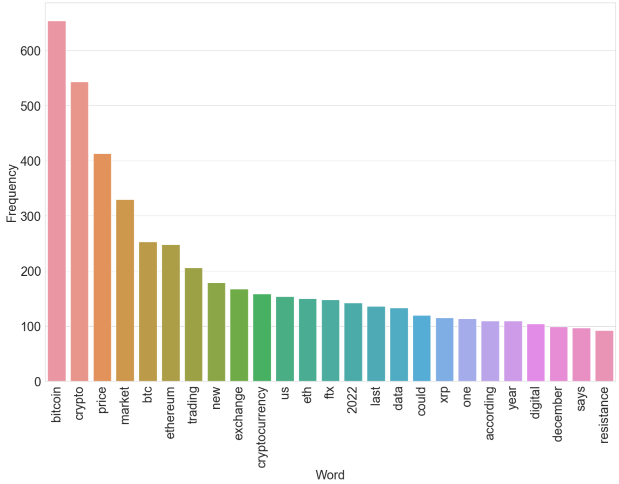

Fig. 2. Top 25 words frequencies

nature of the models, which are transformers. Transformers are context-aware models. Thus, they need as much context as possible to be most accurate. That is why we concatenate the source, title, and description of the news in the following manner:

> *Source: [source-name]; Title: [title-text]; Description: [news-description-text]*

Alternatively for the traditional ML we observe that using only description with stop words and punctuation removal works best. This is also expected since we are doing Term Frequency-Inverse Document Frequency (TF-IDF) (more on this in the next chapter), and this technique works best when we can extract some meaningful words unique to the document. Adding a source or title would add more cliche words (e.g., Bitcoin, crypto), which are, obviously, in almost every document.

C. Benchmarking

The initial phase entails the transformation of textual descriptions within the dataset into TF-IDF vectors. This technique captures the importance of terms within individual descriptions while mitigating the influence of commonly occurring words across the corpus.

Following the TF-IDF transformation, Truncated Singular Value Decomposition (SVD), configured with 75 components, is employed. Truncated SVD reduces the dimensionality of the TF-IDF vectors, distilling them into 75 essential features while retaining crucial information vital for predictive analysis. This reduction aids in handling high-dimensional data and potential noise, facilitating more efficient model training.

Subsequently, a set of diverse machine learning algorithms is selected to serve as benchmark models for prediction. The models used for benchmark predictions included Nearest Neighbors, Linear SVM, RBF SVM, Gaussian Process, Decision Tree, Random Forest, Neural Net, AdaBoost, and QDA. Each of these models represents a distinct approach to classification tasks, spanning from traditional methods like decision trees and SVMs to more complex ensemble techniques like random forests and boosting algorithms.

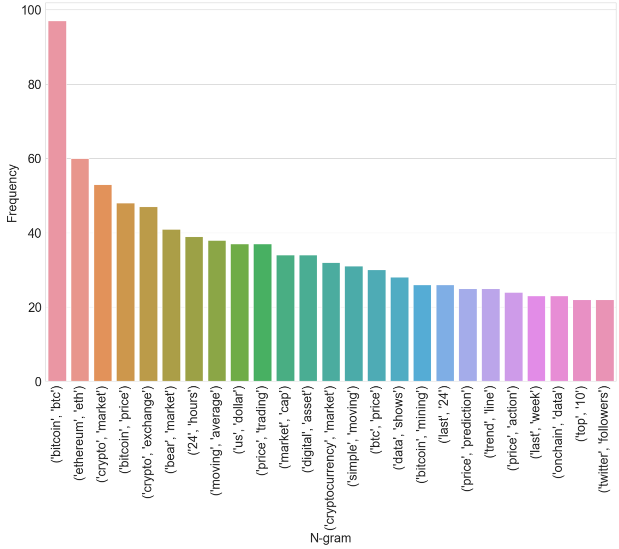

Fig. 3. Top 25 bigrams frequencies

These selected models are trained and evaluated on the reduced feature set obtained from the Truncated SVD-transformed TF-IDF descriptions. The evaluation metrics employed encompass standard measures such as accuracy, precision, recall, and F1-score to comprehensively assess each model's performance in predicting the impact of news sentiment on cryptocurrency prices.

By systematically applying these steps and leveraging a diverse array of machine learning models, this benchmarking methodology aims to establish a comparative foundation for subsequent model enhancements. The results obtained from this benchmarking process will serve as a reference point for evaluating the advancements achieved through fine-tuning transformer models and other sophisticated techniques in predicting cryptocurrency price movements based on news sentiment analysis.

D. Training BERT

BERT (Bidirectional Encoder Representations from Transformers) is a cutting-edge language representation model that revolutionized natural language processing tasks. Unlike traditional models that read text sequentially, BERT comprehends the context of words in a sentence by considering both preceding and succeeding words simultaneously [7]. This bidirectional understanding enables BERT to capture nuanced relationships within language, making it exceptionally adept at tasks like text classification, entity recognition, and understanding the meaning of words in context. Leveraging pre-trained representations of language, BERT allows for fine-tuning specific tasks, making it a powerful tool for our project in predicting cryptocurrency price movements based on news sentiment analysis.

TABLE I
PERFORMANCE METRICS FOR TRADITIONAL MACHINE LEARNING MODELS

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Nearest Neighbors	0.67	0.65	0.67	0.65
Linear SVM	0.68	0.69	0.68	0.60
RBF SVM	**0.73**	**0.72**	**0.73**	**0.70**
Gaussian Process	0.71	0.70	0.71	0.68
Decision Tree	0.61	0.59	0.61	0.59
Random Forest	0.71	0.70	0.71	0.67
Neural Net	0.69	0.68	0.69	0.67
AdaBoost	0.70	0.70	0.70	0.65
QDA	0.68	0.69	0.68	0.68

The model was trained using a stratified split, dividing the dataset into two equal-sized subsets: a training set (50%) and an evaluation set (50%). This approach ensured that both subsets maintained the same distribution of classes, preventing overfitting and ensuring a more accurate assessment of the model's generalizability. To prepare the dataset for training, the news article descriptions were preprocessed by limiting the maximum length of each sequence to 512 tokens. This step was essential to avoid memory constraints during training and to ensure that the model focused on the most relevant information in each article.

The model was trained using a batch size of 32, meaning that it processed 32 news articles at a time during each training step. This batch size was chosen to balance training efficiency with model performance. A learning rate of $2 * 10^{-5}$ was used to control the magnitude of the updates applied to the model's parameters during training. This learning rate was selected based on preliminary experiments to optimize model convergence and avoid overshooting the optimal solution. The training process consisted of running the model through the entire training set for a single epoch, meaning that each training example was presented to the model once. This one-epoch training regime was sufficient for the model to achieve satisfactory performance and avoid overfitting.

To stabilize the training process and prevent numerical instability, a maximum gradient normalization value of 1000 was employed. This technique clipped the gradients of the model's loss function to a maximum value of 1000, preventing them from becoming excessively large and causing training instability. After completing the training process, the model was evaluated on the unseen evaluation set, consisting of the remaining 50% of the dataset. The evaluation metrics, including loss, accuracy, precision, recall, ROCAUC, and F1 scores, were recorded to assess the model's performance on unseen data. These metrics will be analyzed in detail in the results section.

E. Labeling Data

As we have discussed previously, we want to employ GPT-4 with web access to label additional data for our task and see if this boosts the performance. To force the model to produce some reasonable outcome, we provide a simple yet efficient prompt:

> *Based on the provided piece of news, search the web and find evidence if this news affected the price of the cryptocurrency or not. If it did, then just output: [effect:1], otherwise just output: [effect:0]. Only output the result, without explanations. Try to be very concise. News: [news-body]*

We ask the model to be concise because it is designed to give clear and elaborate answers, which is very time-consuming. Generating the output for a single piece of news could take between 30-90 seconds (because aside from response generation, the model needs to search and analyze multiple sources).

F. Evaluation Metrics

Since this study is a classification task, where the goal is to predict the impact of news articles on Bitcoin prices (positive, negative, or no impact), we will employ classification metrics to evaluate the performance of our NLP models. These metrics assess the models' ability to correctly classify news articles based on their assigned impact labels. The following metrics are used: Accuracy, Precision, Recall, F1-score, AUCROC, AUPRC.

V. RESULTS

A. Traditional Machine Learning Models

We initiated our study by conducting experiments with traditional Machine Learning techniques to establish a baseline for performance. The results obtained on the initial dataset are presented in Table I. Notably, the Support Vector Machine (SVM) with Radial Basis Function (RBF kernel) outperformed other models, achieving the highest scores in accuracy, precision, recall, and F1-Score.

1) Performance on Initial Dataset: Table I summarizes the performance metrics for various traditional machine learning models on the initial dataset. It is evident that SVM with RBF kernel exhibits superior performance, achieving an accuracy of 0.73 and an F1-Score of 0.70.

2) Performance on Augmented Dataset: To further evaluate the robustness of these models, we repeated the experiments using an augmented dataset generated by Bing GPT. The results are presented in Table II. Surprisingly, the performance

TABLE II
Performance metrics for traditional machine learning models after adding more data

Model	Accuracy	Precision	Recall	F1-Score
Nearest Neighbors	0.62	0.59	0.62	0.59
Linear SVM	**0.68**	**0.68**	**0.68**	0.62
RBF SVM	0.67	0.65	0.67	**0.64**
Gaussian Process	0.66	0.64	0.66	**0.64**
Decision Tree	0.65	0.63	0.65	0.63
Random Forest	0.67	0.65	0.67	0.63
Neural Net	0.64	0.62	0.64	0.62
AdaBoost	0.63	0.62	0.63	0.62
QDA	0.61	0.63	0.61	0.62

of most models dropped when trained on the augmented dataset. The SVM with linear kernel now emerges as the top-performing model, demonstrating the impact of augmented data on model generalization.

B. Open-AI Base-Models

Next, we evaluated the performance of Open-AI base models, focusing on the Ada model. The results are presented in Table III. We compared the model's performance on manually curated data against its performance on augmented data.

1) Performance on Manual Data: On the initial dataset, the Ada model achieved an accuracy of 0.70, with a precision of 0.72 and a recall of 0.89. The area under the ROC curve (AUROC) and the area under the precision-recall curve (AUPRC) was also notably high at 0.74 and 0.84, respectively. The F1-Score reached 0.80, indicating a well-balanced performance.

2) Performance on Augmented Data: When evaluated on the augmented dataset, the Ada model's performance slightly declined. The accuracy dropped to 0.66, and there were marginal reductions in precision, recall, and F1-Score. This emphasizes the importance of carefully assessing the impact of augmented data on model performance.

TABLE III
Performance metrics for Open-AI Ada base-model

Metric	Manual Data	Augmented Data
Model	Ada	Ada
Accuracy	**0.70**	0.66
Precision	0.72	0.72
Recall	**0.89**	0.73
AUROC	**0.74**	0.69
AUPRC	**0.84**	0.76
F1-Score	**0.80**	0.72

C. BERT-base-cased

Lastly, we examined the performance of the BERT-base-cased model on both the initial and augmented datasets. The results are presented in Table IV.

1) Performance on Manual Data: On the initial dataset, BERT-base-cased achieved an accuracy of 0.66, with precision, recall, and F1-Score values indicating a balanced classification performance.

2) Performance on Augmented Data: Surprisingly, the model's performance remained consistent when evaluated on the augmented dataset. The lack of improvement or degradation suggests that the BERT-based model might not be significantly affected by the augmented data.

TABLE IV
Performance metrics for BERT-base-cased.

Data	Manual Data	Augmented Data
Model	BERT-base-cased	BERT-base-cased
Accuracy	0.66	0.66
Precision	0.44	0.44
Recall	0.66	0.66
AUROC	0.50	0.50
AUPRC	0.34	0.34
F1-Score	0.53	0.53

VI. Conclusion and Future Directions

Our comprehensive evaluation of traditional machine learning models, Open-AI base-models, and BERT-base-cased on both initial and augmented datasets has provided valuable insights into the performance and adaptability of these models.

The experiments with traditional machine learning models highlighted the impact of data augmentation on model generalization. While SVM with RBF kernel demonstrated superior performance on the initial dataset, its efficacy diminished when applied to the augmented data. Interestingly, SVM with linear kernel emerged as the top-performing model on the augmented dataset, showcasing the importance of adapting models to the characteristics of augmented data.

Our assessment of the Open-AI base models, particularly the Ada model, revealed robust performance on manually curated data. However, when evaluated on augmented data, a marginal decline in performance was observed. This underscores the need for careful consideration and validation when incorporating augmented data into models trained on manually curated datasets.

The BERT-base-cased model demonstrated consistent performance on both initial and augmented datasets. The lack of significant improvement or degradation in performance suggests that BERT-based models may have inherent robustness to variations introduced by augmented data.

While our study has shed light on the behavior of various models with augmented data, there are avenues for further exploration and improvement. Fine-tuning model hyperparameters, exploring different data augmentation techniques, considering ensemble models, and emphasizing the importance of explainability and interpretability are key areas for future research.

In conclusion, our study serves as a foundation for future research endeavors aimed at refining the integration of augmented data in machine learning models. By addressing the identified areas for improvement, we can advance the effectiveness and reliability of models in real-world applications.

We also provide an open-source GitHub repository with all of the experiments that we have run, for reproduction and improvement. It also includes all the data we have used and gathered throughout the research. The repository also provides a script for automatic data labelling with Bing GPT API, which we have created. This script is adaptable for any specific task and may serve for the development of this research or for further researches of the automatic labeling capabilities.

REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[2] Nishant Rai, Deepika Kumar, Naman Kaushik, Chandan Raj, and Ahad Ali. Fake news classification using transformer based enhanced lstm and bert. *International Journal of Cognitive Computing in Engineering*, 3:98–105, 2022.

[3] Amira Samy. Sentiment analysis classification system using hybrid bert models. *Journal of Big Data*, 10, 06 2023.

[4] Naiqing Guan, Kaiwen Chen, and Nick Koudas. Can large language models design accurate label functions?, 2023.

[5] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences, 2023.

[6] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification?, 2020.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.