CISC 451 - Project Report

Synthesizing the Future: Generating High-Quality Synthetic Data for Financial Sentiment Analysis

by

Ethan Kim (20233332), Yuchen Li (20216533) & Matteo Parry (20148521)
UG-Team 1

Queen's University
Kingston, Ontario, Canada
(April, 2024)

Problem Definition

In the field of Natural Language Processing (NLP), the task of sentiment analysis serves as a pivotal mechanism for interpreting public sentiment, particularly within the finance sector where investor sentiment profoundly influences market dynamics. The conventional approach to training sentiment analysis models heavily relies on datasets composed of human-annotated texts. Such datasets, however, are often marked by their scarcity and the high costs associated with their curation. This limitation significantly impacts the performance of sentiment analysis models, as their effectiveness is directly correlated with the diversity and caliber of the training data at their disposal. Amidst ongoing advancements in artificial intelligence, the prospect of employing synthetic data—fabricated through AI models like GPT—for the enhancement or potential substitution of human-annotated datasets presents a compelling avenue for exploration.

This innovative strategy aims to address the prevailing challenges concerning data availability, expense, and quality, thereby elevating the overall efficacy and precision of sentiment analysis models. The genesis of this problem lies in real-world scenarios that demand the analysis of user sentiment despite the absence of adequate data.

Detailed Data Description with Visualizations and Statistical Measures

We decided to select a financial dataset given it is invaluable for sentiment analysis due to its voluminous and continuously updated nature, ensuring a robust foundation for training and refining models. The diverse range of sentiments contained within, from optimistic market outlooks to concerns over economic downturns, presents a unique opportunity to enhance the accuracy of sentiment detection algorithms. Improved sentiment analysis directly impacts decision-making in finance, providing stakeholders with deeper insights for informed strategies. Moreover, the established precedent of sentiment analysis in finance offers a wealth of benchmarks for comparison, facilitating the validation of innovative approaches and demonstrating tangible advancements in the field.

Detailed Statistical Analysis and Linguistic Exploration

The provided text offers an overview of a dataset designed for sentiment analysis within the financial news domain, specifically from a retail investor's perspective. This dataset is structured around two primary columns: "Sentiment" and "News Headline", capturing financial news headlines' essence and emotional tone as individual investors might perceive them. Each entry in the dataset is categorized into one of three sentiment classes: negative, neutral, or positive, reflecting the sentiment conveyed by the news headline regarding its potential impact on financial markets or investments. The dataset comprises 4,846 textual entries, indicating a substantial volume of data for analysis. The quality and composition of this dataset are critical, as the accuracy and effectiveness of sentiment analysis models trained on this data directly depend on the representation and balance of sentiments within the dataset. This dataset thus

serves as a foundational resource for developing and evaluating algorithms aimed at understanding and predicting investor sentiment based on financial news headlines.

Our detailed investigation into the dataset, uncovers a nuanced sentiment distribution (Figure 1). Neutral sentiments overwhelmingly dominate the dataset, accounting for 59.41% of the total entries. This significant majority suggests that the dataset's content primarily focuses on delivering information rather than expressing opinions. Positive sentiments are observed in 28.13% of the entries, highlighting a considerable presence but substantially less than neutral tones. Negative sentiments are the least represented, making up only 12.46% of the data. This distribution pattern indicates a potential challenge for sentiment analysis models due to a skewed representation toward neutral and positive sentiments.

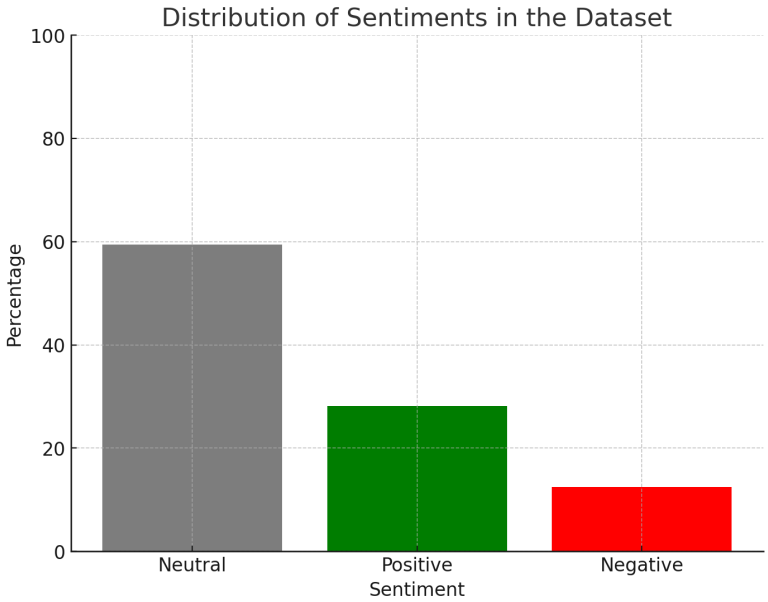

Figure 1: Frequency distribution of sentiments in the dataset.

The text length within the dataset varies significantly (Figure 2), ranging from a minimum of 9 characters to a maximum of 315 characters. The average text length is calculated to be approximately 128 characters, suggesting a wide variety in the level of detail and information conveyed in each entry. The standard deviation of text lengths is measured at 56.5, emphasizing the diverse range of communicative styles present within the dataset. This diversity is further illustrated by the quartile measures, with the 25th percentile at 84 characters and the 75th percentile at 163 characters, indicating that half of the dataset's texts fall within this length range.

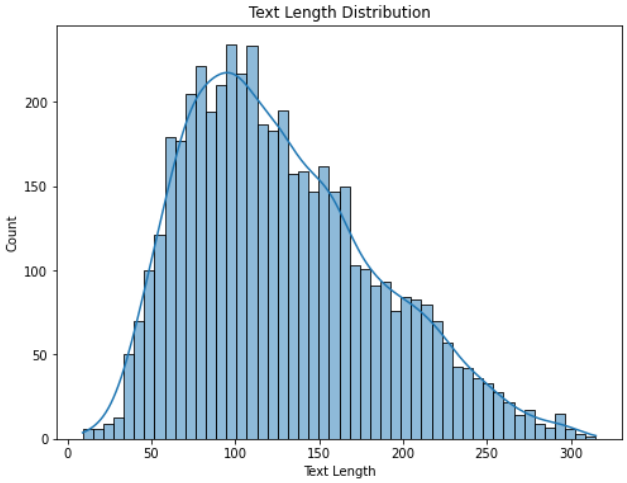

Figure 2: Frequency distribution of text length in the dataset.

Boxplot analysis comparing sentiment distribution against text length (Figure 3) reveals subtle differences in how sentiments are expressed through text length. Neutral sentiments are typically conveyed in more concise text, with negative sentiments slightly longer on average, suggesting a tendency to provide more context when expressing negativity. Positive sentiments are identified as the most verbose, potentially due to the need to use more words to convey affirmation or positivity effectively.

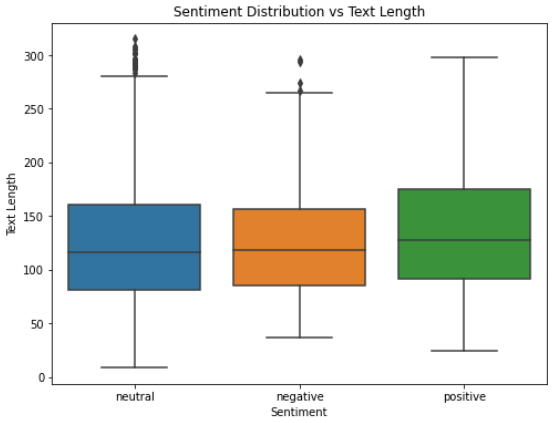

Figure 3: Box plots for sentiment distribution vs text length.

Further analysis using comparative histograms for each sentiment category (Figure 4) uncovers distinct patterns in text length distribution relative to sentiment. Neutral sentiments, with their wide dispersion, align closely with the dataset's average text length, suggesting a broad range of contexts where neutrality is expressed. Positive sentiments exhibit longer text lengths on average, indicating a tendency towards more detailed expressions of positivity. In contrast, negative sentiments display a similar pattern to neutral sentiments but with a slightly narrower variance, suggesting a specific range within which negative sentiments are typically expressed.

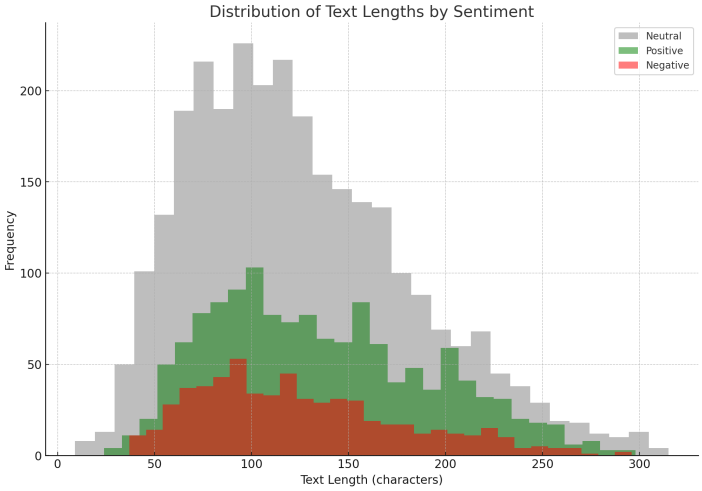

Figure 4: Frequency distribution of text length in the dataset by sentiment.

To highlight sentiment-specific vocabulary, word clouds were generated for each sentiment category (Figure 5) after a rigorous process of excluding common stopwords and terms frequently appearing across categories. This iterative refinement resulted in distinctive word clouds that effectively capture the unique lexical fields of positive, negative, and neutral sentiments, offering deep insights into each sentiment type's thematic and emotional undertones.

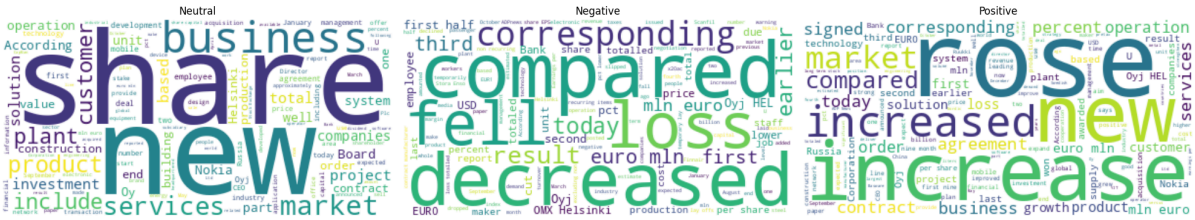

Figure 5: Wordcloud by sentiment distribution.

Refined Analysis of Linguistic Patterns & Topic Modeling

Understanding the textual content of financial news requires more than just analyzing statistical distributions; it necessitates sophisticated techniques like topic modelling to uncover underlying themes. We employed BERTopic, a cutting-edge tool leveraging transformer technology, for deep analysis. This approach is crucial for generating synthetic data that mirrors the dataset's intricacies. BERTopic's flexibility, enabled by its various components—embedding through All-mpnet-base-v2, dimensionality reduction via Umap, HDBScan for clustering, and c-TF-IDF for weighting—allows for a tailored analysis that aligns with our objectives.

Key to our analysis was fine-tuning with GPT-3.5-Turbo, which assigned interpretable labels to 90 identified topics, laying the groundwork for creating a synthetic sentiment dataset enriched with the dataset's nuanced sentiment-based textual content. This process, rooted in

BERTopic's comprehensive modelling capabilities and enhanced by GPT-3.5 Turbo's extensive context handling, aims to capture the essence of the original data while addressing its class imbalances.

With more than 80 identified topics, it becomes clear that specific topics align with distinct sentiments. However, the visualizations (Figure 6) do not explicitly reveal the correlation between particular topics and their associated sentiments, complicating the task of creating a well-balanced dataset. The breadth of topics presents a challenge in discerning overarching trends within the data. Segmenting the dataset by individual sentiments can shed light on which topics most frequently align with each sentiment category (Figure 7).

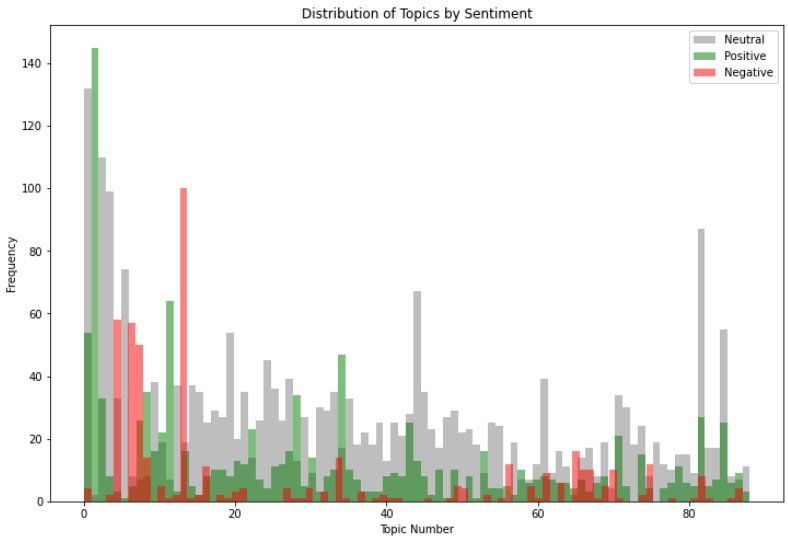

Figure 6: Distribution of topics by sentiment.

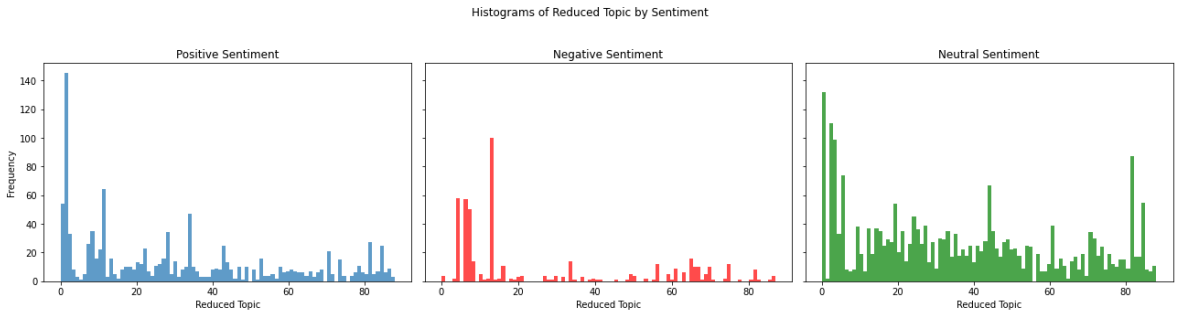

Figure 7: Histograms of reduced topic sentiment.

Narrowing down to the five most prominent topics within each sentiment category (Figure 8) enhances our grasp on the content and predominant sentiment each carries. This refined focus provides deeper insights into the thematic essence and sentiment inclination of the dataset's content.

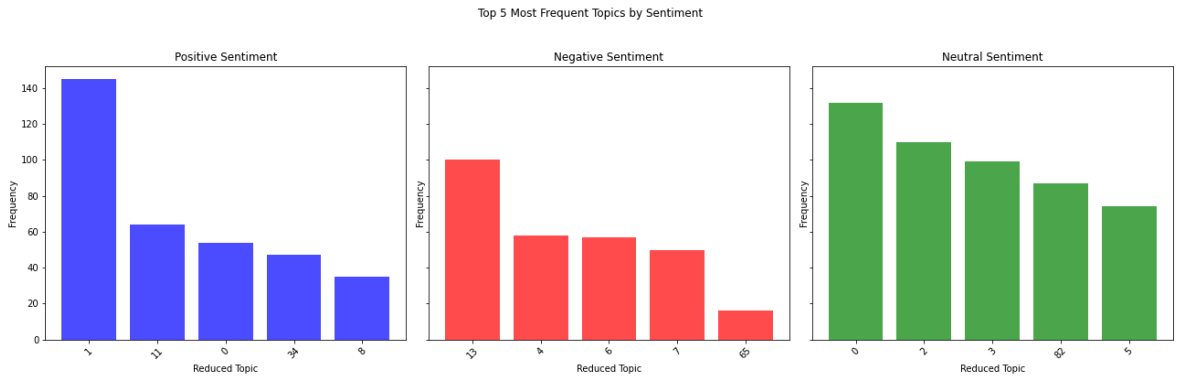

Figure 8: Histograms of top 5 topics by sentiment.

The analysis highlighted overlaps in topic-sentiment associations, especially between neutral and positive sentiments, emphasizing the subtleties in distinguishing between informational content and sentiment expression. This observation, alongside the identification of critical topics within negative sentiments—like sales decline and layoffs—underscores the necessity for refined topic consolidation and enhanced qualitative analysis to ensure comprehensive coverage and balanced representation in synthetic data generation.

This streamlined exploration into linguistic patterns and topic modelling with BERTopic reveals critical insights into the dataset's sentiment composition and thematic structure, guiding the synthetic data generation process focusing on accuracy, relevance, and balance.

Description of Challenges

The first major challenge was finding a high-quality dataset to synthesize the keywords and topics for our synthetic data generation. This foundational dataset needed not only to span three sentiment categories—negative, neutral, and positive—but also be related to the financial domain and have expertly labelled sentiments. With businesses often guarding such a valuable asset from the public eye, finding it was a daunting task. Our diligent exploration on Google Scholar ultimately led us to a goldmine: a dataset expertly annotated as part of a study by Buechel and Hahn (2013), “Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts”. What stood out was his dataset being described by the authors as having been labelled by “Three of the annotators were researchers and the remaining 13 annotators were master's students at Aalto University School of Business” (Buechel & Hahn, 2013) thereby aligning perfectly with our stringent quality benchmarks for sentiment analysis.

Our second significant challenge involved crafting high-quality synthetic data both cost-effectively and efficiently. Traditional methods of applying Large Language Models (LLM) to tasks often require prompt engineering, described as “the process of structuring input text for LLMs and is a technique integral to optimizing the efficacy of LLMs.” (Reynolds et al., 2023).

Although effective at optimizing content output, prompt engineering is rather challenging to automate and would require a human in the process. Moreover, it is a skill which needs to be developed in order to guide the LLM's output as it does not always match the prompter's expectation. As such, this factor can constrain the speed and scalability of synthetic data production as well as impact the quality of the data itself.

The third major hurdle we faced was fine-tuning the parameters of our DSPy text generator. In order to create a synthetic dataset of exceptional quality using a human-expert annotated dataset, we needed an accurate method for distilling semantically rich topics and extracting keywords. The topic extractions needed to be semantically meaningful topics, which are often not found with traditional topic modelling techniques such as Latent Dirichlet Allocation and TF-IDF clustering. In these approaches, the topics are mere clusters of vectors translated back to words rather than nuanced, meaningful themes. Therefore, identifying a suitable keyword extraction method in order to incorporate the sentiment-defining words was key to ensuring each generated piece resonated with the intended emotional undertones.

Navigating the challenge of assessing synthetic data quality emerged as our fourth major obstacle. Evaluation methods for LLMs can be distilled into three categories, spanning quantitative, qualitative, and hybrid approaches. Qualitative methods and hybrid human models are too subjective and impractical for the aims of this project given the amount of manual evaluation that would be required. Quantitative methods such as perplexity–defined as the inverse of the geometric mean of prediction probabilities in a given text (Brown et al., 1992)–are not very useful when it comes to measuring synthetic data outside of a supervised learning problem context. This limitation steered us towards formulating a supervised learning problem, with sentiment analysis serving as the ideal framework. The driving factor for selecting sentiment analysis over content text classification is its ability to capture the subtleties of human emotions embedded within the text, offering a more nuanced challenge than mere text classification.

Methodology

The development of our synthetic data generator leveraged DSPy, an advanced framework designed for the algorithmic optimization of LLM prompts and weights (DSPy Documentation). DSPy's innovative approach simplifies language model pipelines into comprehensive text transformation graphs, facilitating the manipulation of objects, methods, and parameters with ease. This not only automated the synthetic data pipeline creation but also significantly enhanced the initial textual parameters' optimization, ensuring the output is both semantically rich and meaningful.

For the creation of high-quality synthetic data within these pipelines, we utilized OpenAI's GPT-4-0125-preview model. Known for its optimization for natural language

processing tasks, GPT-4 stands out not only for its state-of-the-art performance but also for its cost-effectiveness compared to its predecessor, GPT-4. The integration of DSPy's efficient pipeline methods with GPT-4's advanced capabilities enabled us to generate synthetic textual data of very high quality and accuracy, effectively labelled to meet our analytical needs.

After creating the DSPy generator, the next step entailed the integration of a parameter extraction function, with a primary focus on topics and keywords. after segregating a training subset from the principal dataset, we tackled the challenge of identifying semantically rich topics and pertinent keywords. This endeavour led us to deploy BERTopic alongside TF-IDF for their unparalleled efficacy in topic modelling and keyword extraction, respectively.

BERTopic is a state-of-the-art topic modelling technology, acclaimed for its capacity to generate coherent topics that remain robust across diverse benchmarks (Grootendorst, 2022). This model harnesses the power of TF-IDF metrics and BERT (Bidirectional Encoder Representations from Transformers) to delve into the semantic intricacies of texts, thereby ensuring the extraction of topics with high relevance and coherence. The utilization of transformers with bidirectional capabilities in the context of distilling topics is crucial to recognizing the semantics of text before and after a certain n-gram using attention within a context window. BERTopic incorporates representation models which are used to improve topic coherency. By using a GPT-4 representation model BERTopi significantly enhances topic clarity and relevance.

Once we successfully identify topics closely aligned with the sentiments of interest, TF-IDF steps in to pinpoint the most relevant keywords. Its simplicity and efficiency underscore the preference for TF-IDF in this context as a weighting mechanism, notably outperforming alternatives like TextRank (Mihalcea & Tarau, 2004). The final step of this process involves feeding the DSPy generator with the distilled topics, TF-IDF-derived keywords, and additional parameters such as output length and sentiment. DSPy then synthesizes this information, producing synthetic news articles with corresponding sentiments, demonstrating a sophisticated blend of technology and analytics in generating realistic and contextually accurate synthetic data.

These outputs are compiled into a data frame with a text and sentiment column, serving as the foundation for training a BERT sentiment classifier. The process involves employing BERT's tokenizer to prepare input IDs and attention masks, which is crucial for understanding the context of textual data. Following tokenization, a pre-trained BERT model for sequence classification is fine-tuned on this data. Leveraging PyTorch, the model training and optimization are meticulously conducted, with a concerted effort to refine model weights. The objective here is to reduce loss while enhancing the accuracy of sentiment predictions, thereby ensuring the model's adeptness in distinguishing between varying sentiment categories.

In the training phase, the model's performance is rigorously assessed against a validation dataset. This evaluation is crucial for optimizing hyperparameters and mitigating the risk of overfitting, thereby ensuring the model's robustness and reliability. Once training is completed, the model is evaluated on a separate testing data set. This step is instrumental in measuring the effectiveness of generalizing sentiment classification across new, unseen texts. Finally, results are analyzed through accuracy metrics and a detailed classification report. This approach validates the model's performance and offers insights into its predictive capabilities and potential areas for further enhancement in sentiment analysis tasks.

To replicate the experiment and achieve the outlined results, please refer to the zip file attached to this report. Inside, one will find both the datasets used in our analysis and a Jupyter notebook with the complete code. Simply run the notebook following the provided instructions to conduct the experiment and assess the model's performance.

Evaluation

Synthetic data + Original testing set

The model achieves an accuracy of 60.6%, with a notable proficiency in identifying neutral sentiments (precision: 0.61, recall: 0.96). This suggests the model's inclination toward the most prevalent sentiment class within the training data, indicative of potential class imbalance handling issues. The significantly lower recall for positive (0.07) and negative (0.16) sentiments underscores a critical challenge in recognizing less frequent sentiment classes, potentially due to the synthetic data's inability to fully encapsulate the nuances and variability present in the original testing.

Test Accuracy: **0.606**

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.58	0.16	0.26	110
Neutral	0.61	0.96	0.74	570
Positive	0.61	0.07	0.12	289
Accuracy	0.61			969
Macro avg	0.60	0.40	0.37	969
Weighted avg	0.60	0.61	0.50	969

Synthetic data + External testing set

Improving to 69.7% accuracy, the model displays enhanced capability in identifying negative sentiments with increased precision (0.78) and recall (0.75). This improvement suggests that while synthetic data may not fully mimic the original dataset's complexity, it can provide valuable generalization capabilities, especially in distinguishing clearly defined sentiment categories like negative sentiments in varied contexts.

Test Accuracy: **0.697**

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.78	0.75	0.76	84
Neutral	0.59	0.85	0.69	65
Positive	0.81	0.47	0.59	62
Accuracy	0.70			211
Macro avg	0.72	0.69	0.68	211
Weighted avg	0.73	0.70	0.69	211

Original data + Original testing set

Here, the model exhibits its highest accuracy at 83.8%, demonstrating strong performance across all sentiment categories. This indicates the quality of the dataset when applied to the original datas testing set. The balanced precision and recall across sentiment categories reflect a well-trained model capable of nuanced sentiment differentiation

Test Accuracy: **0.838**

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.84	0.85	0.84	110
Neutral	0.91	0.82	0.87	570
Positive	0.73	0.87	0.79	289
Accuracy	0.84			969
Macro avg	0.83	0.84	0.83	969

Weighted avg	0.85	0.84	0.84	969

Original data + External testing set

When applied to the external dataset the accuracy decreases to 67.3%, but remains relatively stable across all sentiment categories. The results indicate that applying an external dataset to a model trained on the original dataset reduces the model's universal compatibility.

Test Accuracy: **0.673**

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.82	0.67	0.74	84
Neutral	0.61	0.66	0.64	65
Positive	0.59	0.69	0.64	62
Accuracy	0.67			211
Macro avg	0.68	0.67	0.67	211
Weighted avg	0.69	0.67	0.68	211

Conclusions

The detailed evaluation of models trained on synthetic and original datasets against both original and external testing sets has yielded several insights. Primarily, the analysis underscored the nuanced capabilities of sentiment analysis models in deciphering the complex landscape of financial news sentiment, revealing both strengths and limitations influenced by the nature of the training data.

From our evaluation, models trained on synthetic datasets demonstrated a strong capability for identifying neutral sentiments, as evidenced by their precision and recall metrics. However, these models exhibited challenges in accurately classifying positive and negative sentiments, highlighted by lower recall rates for positive sentiments. This discrepancy points towards a potential shortfall in the ability of synthetic data to capture and represent the subtleties of sentiment expression found in real-world financial news articles. Conversely, models trained on original data showcased superior performance across all sentiment categories, with particularly robust accuracy and balanced metrics when evaluated on the original testing set. This affirms the intrinsic value of rich, diverse, and accurately labelled original datasets in developing models that are not only precise but also refined in their sentiment analysis capabilities.

When applied to external testing sets, a notable decline in accuracy for models, irrespective of their training data, accentuates the challenges of generalization in sentiment analysis. This phenomenon brings to light the inherent variability and complexity of sentiment expressions across different financial news sources, highlighting the critical need for models to learn from a broad spectrum of data to achieve optimal generalization.

The comparative analysis of sentiment analysis models trained on synthetic versus original datasets brings to light the critical role of data quality, diversity, and representativeness in training effective models. While synthetic data offers a promising path for overcoming the limitations of data scarcity and accessibility, our findings indicate that enhancements in synthetic data generation methodologies are imperative. Specifically, improving the semantic richness and sentiment representation within synthetic datasets could bridge the performance gap observed in comparison to models trained on original datasets.

Overall, the insights gained from this project emphasize the importance of model adaptability and generalization across varied datasets. As models strive to interpret and predict sentiments accurately, incorporating a more comprehensive array of data sources in the training phase could be key to enhancing their efficacy and applicability in real-world scenarios.

Future Work

The primary objective for future work is to enhance the quality of the synthetic dataset, aiming to match or surpass the performance achieved with the original dataset. Despite the innovative approach, the synthetic dataset generated in this project did not achieve comparable accuracy levels, signalling the need for methodological refinements in data generation. DSPy remains a promising tool for this task, suggesting that future enhancements could focus on optimizing the topic and keyword extraction processes.

One strategy involves experimenting with various topic modelling methods and keyword extraction tools on smaller data subsets, subsequently having these subsets evaluated by experts for quality. The sentiment models trained on these expert-rated synthetic datasets could then undergo evaluation against both the original dataset (Financial News Headlines Dataset) and an external dataset, aiming to assess both inter-dataset and external-dataset generalization capabilities. This iterative process, though resource-intensive, is crucial for identifying the most effective data synthesis methodologies. However, the financial constraints posed by the requisite OpenAI API credits emerged as a significant obstacle for our team.

Additionally, the broad training scope of GPT-4, encompassing data beyond financial news, introduces challenges in generating text with the specific nuances of financial discourse. In this context, employing specialized models like BloombergGPT (Wu et al., 2023), which is tailored to financial news, could offer a more precise text generation capability for our needs.

Exploring fine-tuned mixture of expert models, such as Mistral, might also provide a strategic advantage by allowing greater control and specificity in the generated content, potentially aligning more closely with the intricacies of financial news data.

References

Brown, P. F., Pietra, S. D., Pietra, V. J. D., & Mercer, R. L. (1992). The mathematics of statistical machine translation: Parameter estimation. Retrieved from https://aclanthology.org/J92-4003.pdf

Buechel, S., & Hahn, U. (2013). Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts. Retrieved from https://arxiv.org/abs/1307.5336

DSPy Documentation. (n.d.). An introduction to DSPy: Optimizing language model prompts and weights. Retrieved from https://dspy-docs.vercel.app/docs/intro

Financial News Headlines Dataset. Retrieved from https://www.kaggle.com/datasets/ankurzing/sentiment-analysis-for-financial-news

Grootendorst, M. (2022). BERTopic: Leveraging BERT and TF-IDF for Unrivaled Topic Modeling. Retrieved from https://arxiv.org/pdf/2203.05794.pdf

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Texts. Retrieved from https://dl.acm.org/doi/10.5555/1944566.1944608

Reynolds, C. R., Doshi, P., & Nichol, A. (2023). The intricacies of prompt engineering for Large Language Models. Retrieved from https://arxiv.org/pdf/2310.14735.pdf

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A Model Trained on Financial News Data. arXiv. https://arxiv.org/abs/2303.17564