Contents

Introduction

With the rise in popularity of social networks, social websites can increasingly provide multimodal data like images, text, and video. Social big data has greatly benefited from the exponential expansion of user-generated material on social media platforms, especially critical text and reviews (K., 2020). Text analytics, specifically sentiment analysis, has revolutionized our ability to glean actionable insights from unstructured data. This form of Natural Language Processing (NLP) focuses on text sentiment analysis, facilitating a nuanced understanding of public opinion, consumer sentiment, and financial market trends. Based on their analysis of stock trends from 1992 to 2017, Agarwal et al. (2019) discovered that sentiment analysis of stock opinions from online sources can help investors make investment decisions and possibly even forecasts. Therefore, in this study, we focus on the comparison of FinBERT, BERT, and traditional models to evaluate the performance in sentiment analysis of financial news data.

Pre-trained BERT and FinBERT

BERT (Bidirectional Encoder Representations from Transformers) and FinBERT are both pre-trained language models used for various NLP tasks such as text classification, sentiment analysis, and named entity recognition (Devlin et al., 2019). While BERT, developed by researchers at Google AI Language, is a general-purpose model trained on the English open-source corpus BooksCorpus and Wikipedia, amounting to a total of 3.3 billion words. FinBERT is a specialized version of BERT that has been fine-tuned on a large financial dataset (including financial news articles or reports) to cater to finance-related tasks (Araci, 2019). The fine-tuning process enables FinBERT to better understand the unique vocabulary, jargon, and context found in financial texts. As a result, FinBERT is better suited for finance-related NLP tasks, such as sentiment analysis of financial news or entity recognition in financial documents (Huang, 2022).

BERT is pre-trained on two unsupervised tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP) (Devlin, et al., 2019). In the MLM task, BERT learns to predict the missing words in a sentence by masking a certain percentage of the input token. MLM encourages the model to learn deep semantic representations of words and sentences which is useful for sentiment analysis. To use BERT for specific NLP tasks such as classification, it must be fine-tuned on labelled data using TensorFlow, Torch, and Transformers libraries.

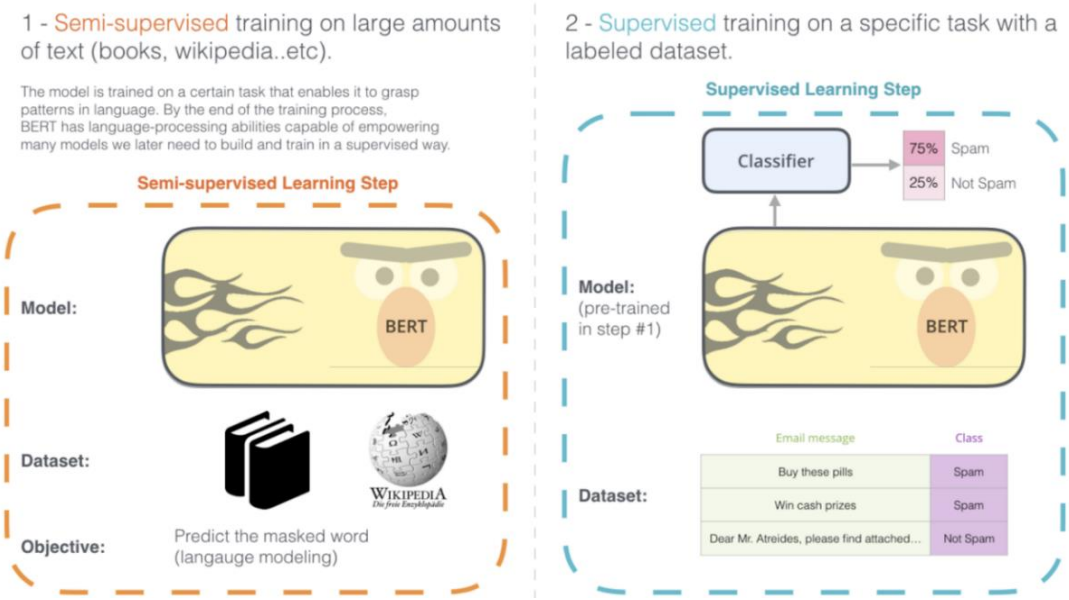


Figure 1. (Alammar, 2021)

Figure 1 shows the two steps of how BERT is developed. We can download the model pre-trained in Step 1 (trained on unannotated data) and only worry about fine-tuning it for Step 2.

Fine-tuning involves training the pre-trained BERT model on a specific task for a few epochs while updating the model weights using task-specific training data. The fine-tuned model can then be used to make predictions on new, unseen data (Devlin, et al., 2019).

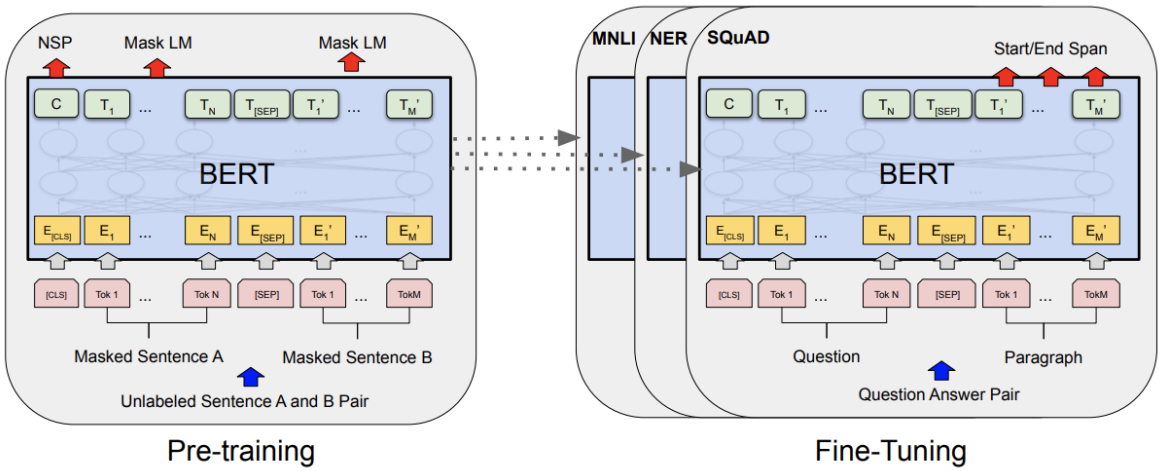


Figure 2. (Alammar, 2021)

Figure 2 shows BERT's learning process, where two main stages are involved: pre-training and fine-tuning. Except for the final output layers, both stages use identical structures. The pre-trained model parameters serve as a starting point for models tailored to different specific tasks. During the fine-tuning stage, all these parameters undergo adjustments. BERT uses special symbols like [CLS] and [SEP]. [CLS] is placed before each input example, and [SEP] is used as a separator token to divide questions from answers, for instance**.** (Devlin, et al., 2019).

Traditional NLP models have limitations in handling complex language structures and generating coherent text (Mikolov, et al., 2013). BERT uses a bidirectional training mechanism that allows it to understand the context of words in a sentence by considering both left and right contexts simultaneously (Peters, et al., 2018). It captures complex linguistic nuances through rich, contextualized word representations, unlike models that rely on basic word embeddings (Araci, 2019). In terms of model architecture, BERT leverages self-attention mechanisms in the transformer architecture to capture long-range dependencies and relationships between words in a sentence (Vaswani, et al., 2017). It is pre-trained on unsupervised tasks before being fine-tuned on specific tasks, while models such as SVM are directly trained on labelled data for a specific task (Howard & Ruder, 2018).

Methodology

To ascertain the efficacy of the models in accurately discerning the sentiment of stock market news, we opted for two disparate datasets to scrutinize the generalization capacity and performance of the models across diverse contexts. The chosen datasets, designated as Financial PhraseBank (dataset_1) and FiQA (dataset_2), were procured from the Hugging-Face repository and encompass financial sentences coupled with sentiment labels. These datasets, after necessary EDA (APPDENDIX 2), cleaning, including symbol and numeric conversion, URL and phone-numbers removal, and lemmatization, were divided for model training and testing to assess the effectiveness of four models: the Bert-base-uncased BERT model, FinBERT provided by Huang et. al (2020), SVM, and Logistic Regression. In assessing the performance of classification models, two pivotal metrics were employed: Overall Accuracy and Macro F1-score. The Macro F1-score is computed by obtaining individual F1-scores for each class and subsequently averaging them. Given that our dataset_2 exhibits a significant label imbalance — with approximately 60% of sentences classified as neutral — the Macro F1-score offers a valuable measure of classification efficacy. The model was optimized with the best hyperparameters, and

any alteration could yield different results. Also, due to limited GPU, model complexity was necessarily restricted.

Evaluation

The classification outcomes of all models on both datasets are presented in Table 1. On the metrics measured, it is evident that the state-of-the-art models, namely BERT and FinBERT, considerably outperform the traditional models for all three classes on both datasets. This is demonstrated by an approximately 8% enhancement in accuracy rate and around 10% improvement in Macro F1-score, thereby showing the effectiveness of pre-training in language models. However, the dataset_2 results expose challenges with the Negative class for all models (APPENDIX 1), resulting from class imbalance.

Table 1: Experimental Result

	First dataset		Second dataset	
Model	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score
Bert	0.83	0.82	0.77	0.68
Finbert	0.82	0.80	0.79	0.72
SVM	0.75	0.69	0.68	0.59
Logistic Regression	0.74	0.65	0.7	0.58

Comparing these two transformer-based models (APPENDIX 1), FinBERT's accuracy on both datasets differed from BERT by only 1%. Yet, FinBERT always surpasses BERT in terms of precision for the 'Positive' categories, implying a greater proficiency in classifying relevant instances. Consequently, a fintech manager's choice between these models should rely on their task-specific needs. If high precision in identifying positive instances, for instance, in the detection of profitable investments or opportunities based on positive reviews, FinBERT could be more advantageous. On the other hand, if identifying negative instances is important, for example, to identify potential risks for proactive risk-management, and spot market overreactions, potentially unveiling opportunities to acquire undervalued stocks, BERT, with its excellent performance on identifying negative text and overall F1-score, might prove to be the preferred choice.

The traditional machine learning models display lower overall performance, but they exhibit competitive results in certain areas. For instance, Logistic Regression shows a high recall for the Neutral class (0.96/0.88) on both datasets, which suggests its potential strength in identifying neutral instances, despite its lower overall scores.

Conclusion

Our research findings show that FinBERT and BERT outperform traditional models in various aspects. However, FinBERT and BERT exhibit similar performance in analyzing financial text, with FinBERT demonstrating superior accuracy in identifying positive sentiment compared to BERT, consistent with the original research. Financial text, with its frequent updates and abundance, reflects market dynamics. Quick and accurate judgments of change support decision-making as market sentiment affects mid and small-cap stocks and can predict stock price direction.

References

Alammar, J., 2021. The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning. [Online] Available at: https://jalammar.github.io/illustrated-bert/

Araci, D., 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Howard, J. & Ruder, S., 2018. Universal Language Model Fine-tuning for Text Classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics , Volume 1, pp. 329-339.

HUANG, A. H., 2022. FinBERT: A Large Language Model for Extracting Informationfrom Financial Text. Contemporary Accounting Research, 00(00), pp. 1-36.

K., Y.R. (2020). Deep Learning-based Aspect-Level Sentiment Analysis of User-Generated Content. Journal of Advanced Research in Dynamical and Control Systems, 12(SP4), pp.1457–1465. doi:https://doi.org/10.5373/jardcs/v12sp4/20201624

Mikolov, T., Chen, K., Corrado, G. & Dean, J., 2013. Efficient Estimation of Word Representations in Vector Space.

Peters, M. E. et al., 2018. Deep contextualized word representations.

Satapathy, R., Cambria, E. and Hussain, A. (2017). Sentiment analysis in the bio-medical domain: techniques, tools, and applications. Cham: Springer.

Vaswani, A. et al., 2017. Attention is All you Need. Advances in Neural Information Processing Systems 30 (NIPS 2017).

Yang, Y., UY, M.C.S. and Huang, A. (2020) ‘FinBERT: A Pretrained Language Model for Financial Communications’. doi: https://doi.org/10.48550/arXiv.2006.08097 (Accessed: 12 May 2023).

APPENDIX 1: EXPERIMENT RESULT IN DETAIL

Table 2: Experimental Result of Logistic Regression on dataset_1

	Precision	Recall	F1-score
Neutral	0.73	0.96	0.83
Negative	0.84	0.44	0.58
Positive	0.75	0.41	0.53
		Macro F1 average	0.65

Table 3: Experimental Result of SVM on dataset_1

	Precision	Recall	F1-score
Neutral	0.78	0.89	0.83
Negative	0.72	0.59	0.65
Positive	0.67	0.52	0.59
		Macro F1 average	0.69

Table 4: Experimental Result of Finbert on dataset_1

	Precision	Recall	F1-score
Neutral	0.88	0.86	0.87
Negative	0.79	0.79	0.79
Positive	0.73	0.76	0.74
		Macro F1 average	0.80

Table 5: Experimental Result of Bert on dataset_1

	Precision	Recall	F1-score
Neutral	0.88	0.88	0.88
Negative	0.83	0.82	0.83
Positive	0.73	0.74	0.74
		Macro F1 average	0.82

Table 6: Experimental Result of Logistic Regression on dataset_2

	Precision	Recall	F1-score
Neutral	0.7	0.88	0.78
Negative	0.43	0.17	0.25
Positive	0.76	0.65	0.7
		Macro F1 average	0.58

Table 7: Experimental Result of SVM on dataset_2

	Precision	Recall	F1-score
Neutral	0.72	0.78	0.75
Negative	0.32	0.23	0.27
Positive	0.75	0.79	0.74
		Macro F1 average	0.59

Table 8: Experimental Result of Finbert on dataset_2

	Precision	Recall	F1-score
Neutral	0.82	0.88	0.85
Negative	0.51	0.43	0.47
Positive	0.85	0.82	0.84
		Macro F1 average	0.72

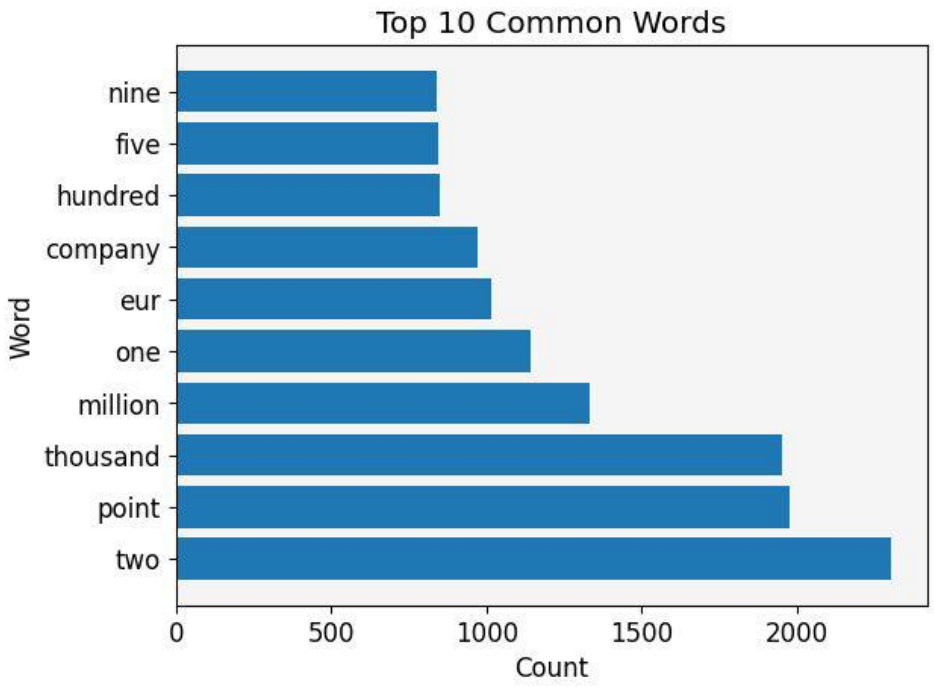
Table 9: Experimental Result of Bert on dataset_2

	Precision	Recall	F1-score
Neutral	0.82	0.88	0.85
Negative	0.47	0.37	0.42
Positive	0.8	0.77	0.78
		Macro F1 average	0.68

APPENDIX 2: EXPLORATORY DATA ANALYSIS OF DATASET_1

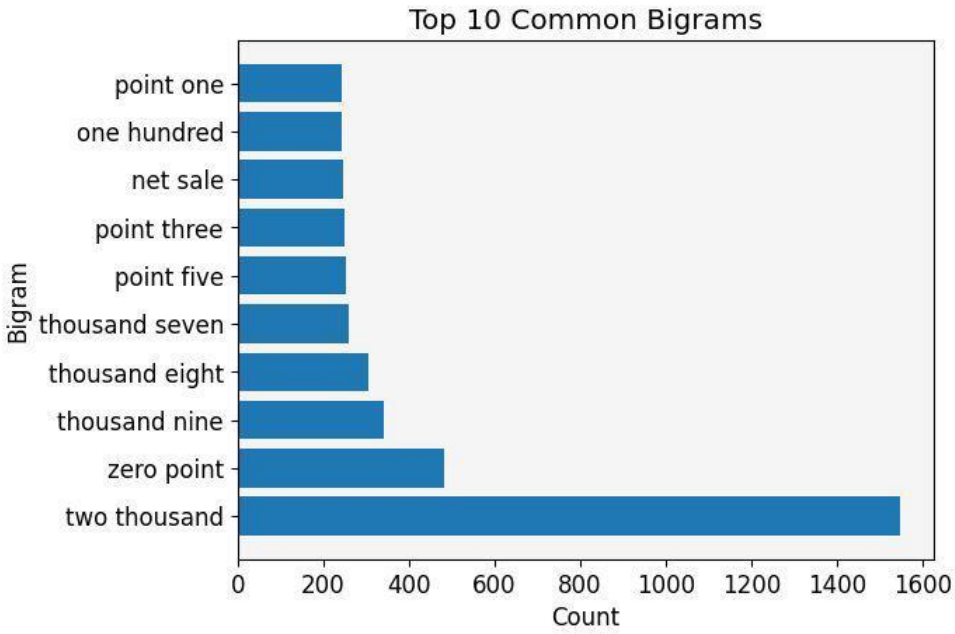
APPENDIX 2.1 Top 10 Common Words

The plot shows the top 10 most common words in the dataset. The most frequently occurring word is 'two' with over 2000 counts, followed closely by 'point' and 'thousand' with around 2000 counts each. The remaining words have much lower counts, ranging from 500 to 1000.



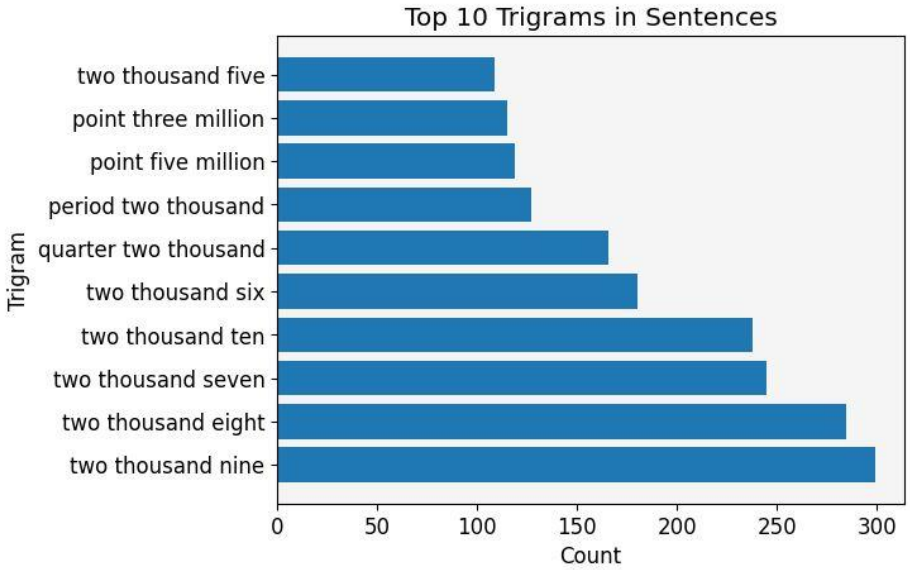
APPENDIX 2.2 Top 10 Common Bigrams

The bar plot illustrates the top 10 most common bigrams in the dataset, with 'two thousand' being the most frequent one with approximately 1500 counts. 'Zero point' and 'thousand nine' are the next two most common bigrams with around 500 and 400 counts, respectively. The presence of these bigrams suggests that the dataset might be related to years or financial transactions, where phrases such as 'two thousand' and 'zero point' are commonly used. The rest of the bigrams are less frequent, with counts ranging from 100 to 300.



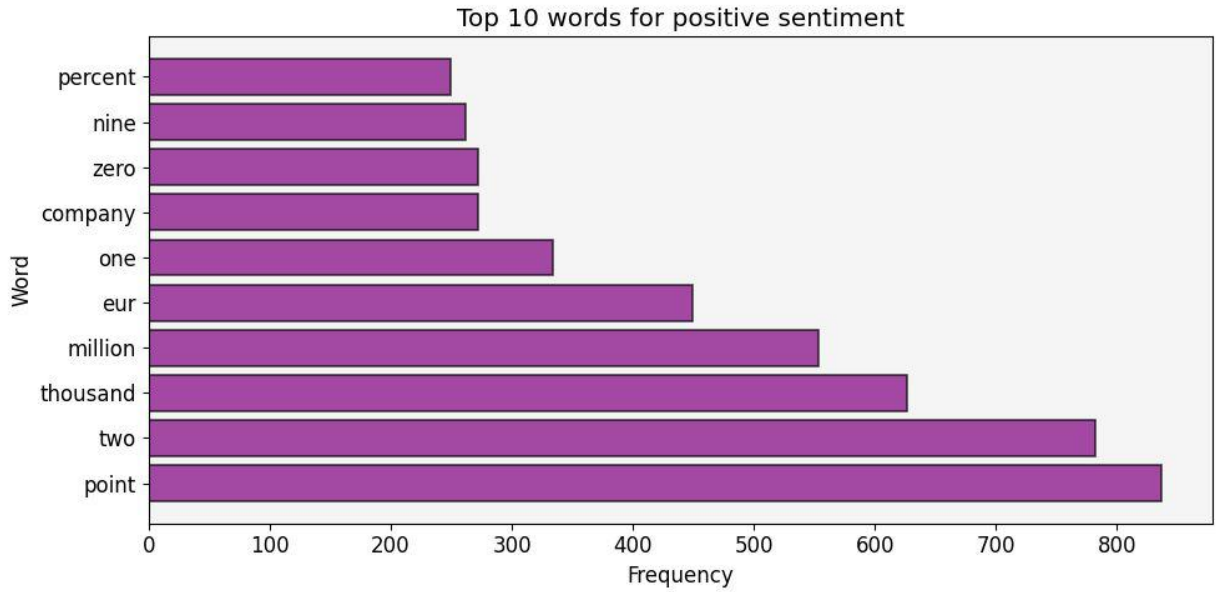
APPENDIX 2.3 Top 10 Common Trigrams

The bar plot displays the top 10 most common trigrams in the dataset, with 'two thousand nine' being the most frequent one with around 300 counts. 'Two thousand eight' and 'two thousand seven' are the next two most common trigrams with around 270 and 250 counts, respectively. The high frequency of these trigrams suggests that the dataset might be related to financial or economic data where years and quarters are commonly used. The remaining trigrams have lower frequencies, ranging from 100 to 200 counts.



APPENDIX 2.4 Top 10 Words for Positive Sentiment

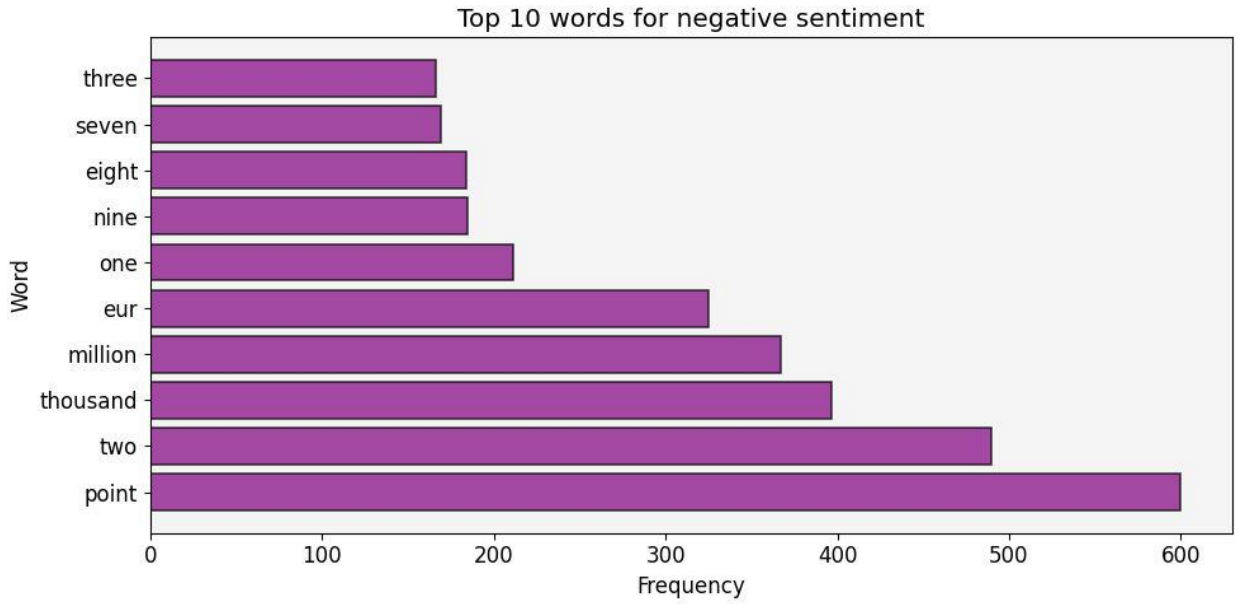
The bar plot illustrates the top 10 most common positive sentiment words in the dataset, with 'point' being the most frequent one with over 800 counts. 'two' and 'thousand' are the next two most common words with around 800 and 700 counts, respectively. The rest of the words are



less frequent, with counts ranging from 100 to 500.

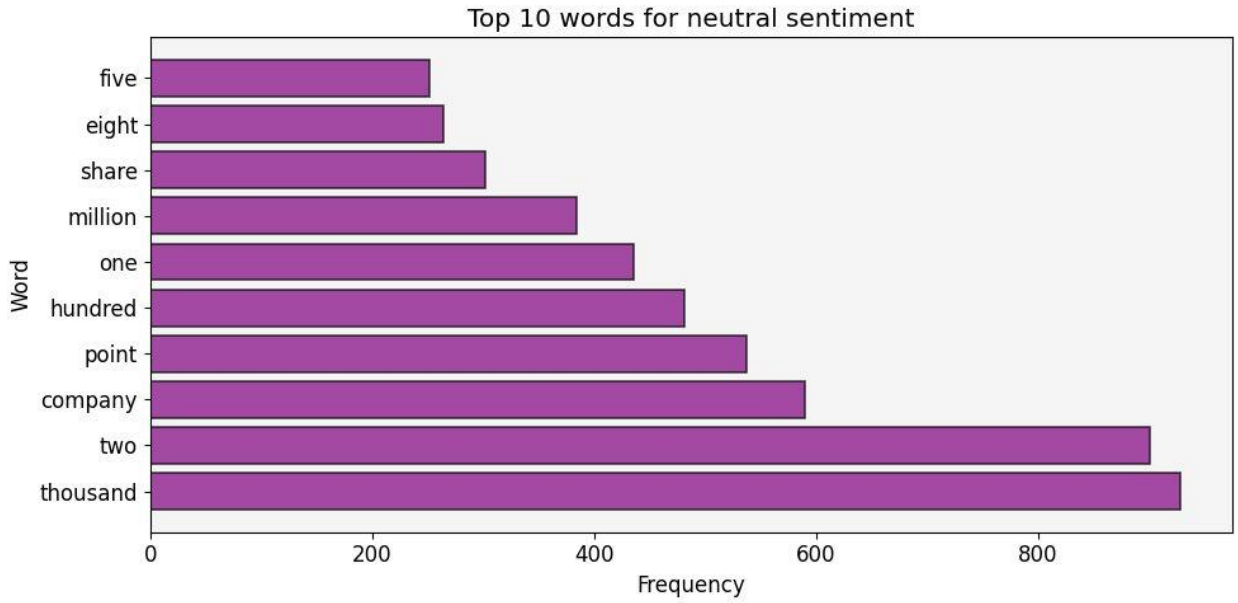
APPENDIX 2.5 Top 10 Words for Negative Sentiment

The plot displays the top 10 most common negative sentiment words in the dataset. The most frequently occurring word is 'point' with around 600 counts, followed by 'two' and 'thousand' with around 500 and 400 counts, respectively. The remaining words have lower counts, ranging from 100 to 300.



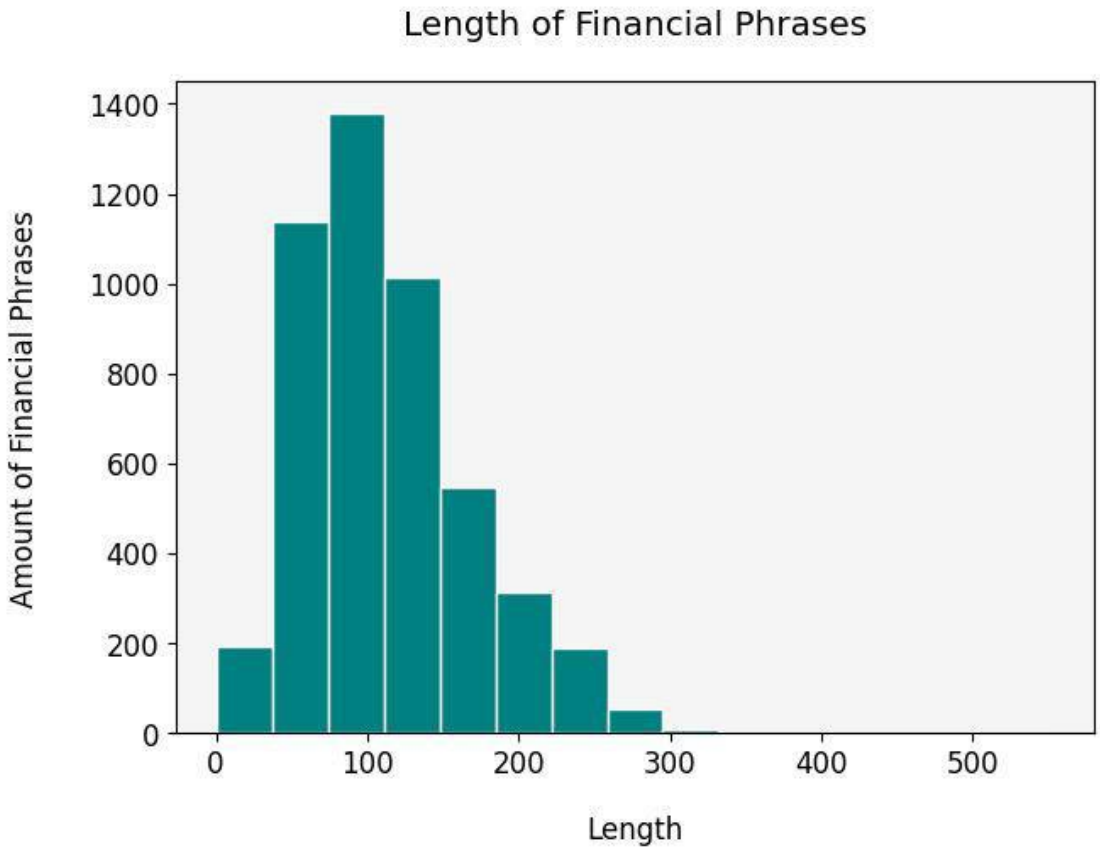
APPENDIX 2.6 Top 10 Words for Neutral Sentiment

The plot displays the top 10 most common neutral sentiment words in the dataset. The most frequently occurring word is 'point' with around 600 counts, followed by 'two' and 'thousand' with around 500 and 400 counts, respectively. The remaining words have lower counts, ranging from 100 to 200.



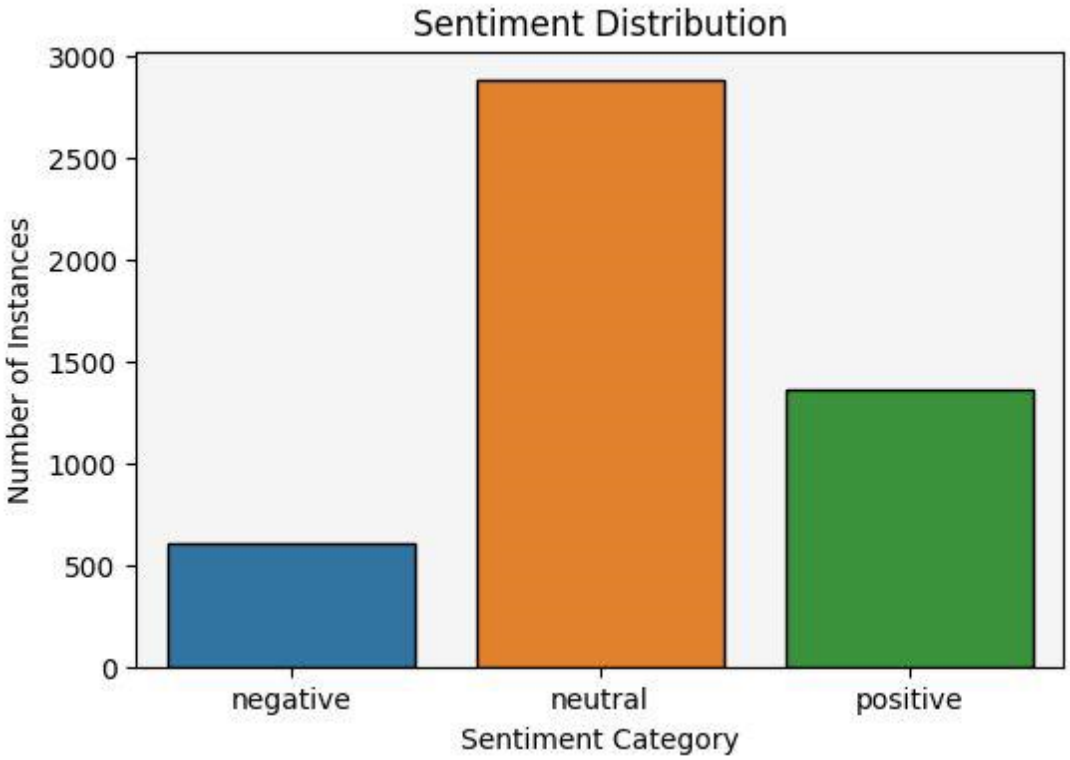
APPENDIX 2.7 Length of Financial Phrases

The histogram shows a distribution of financial phrases by their length. The most common length is around 100 with 1400 phrases, followed closely by length of around 50 with 1200 phrases. The third highest peak occurs at around 120 with 1000 phrases. The distribution shows a long-tail pattern with decreasing frequency as the length increases.



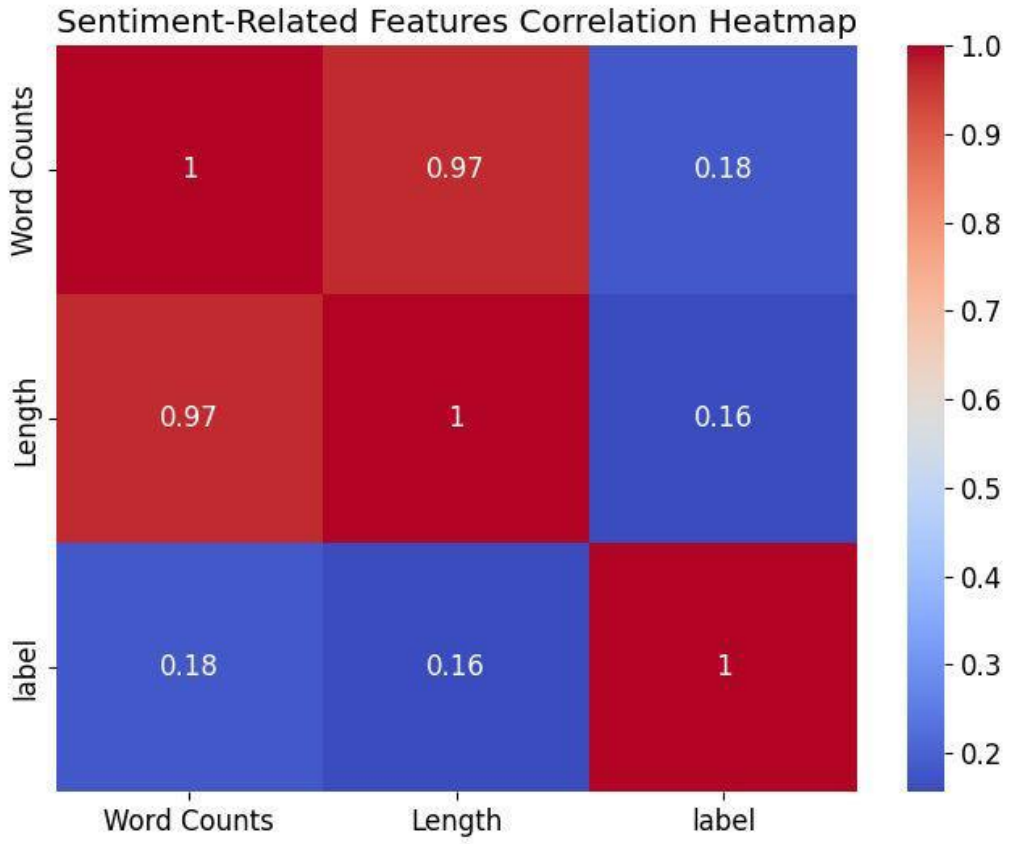
APPENDIX 2.8 Sentiment Distribution

The barplot shows the distribution of sentiment categories in the dataset, with 'neutral' being the most frequent sentiment category with around 3000 instances. 'Positive' and 'negative' sentiments are the next two most common categories, with around 1300 and 500 instances, respectively. The plot suggests that the majority of instances in the dataset are classified as neutral.



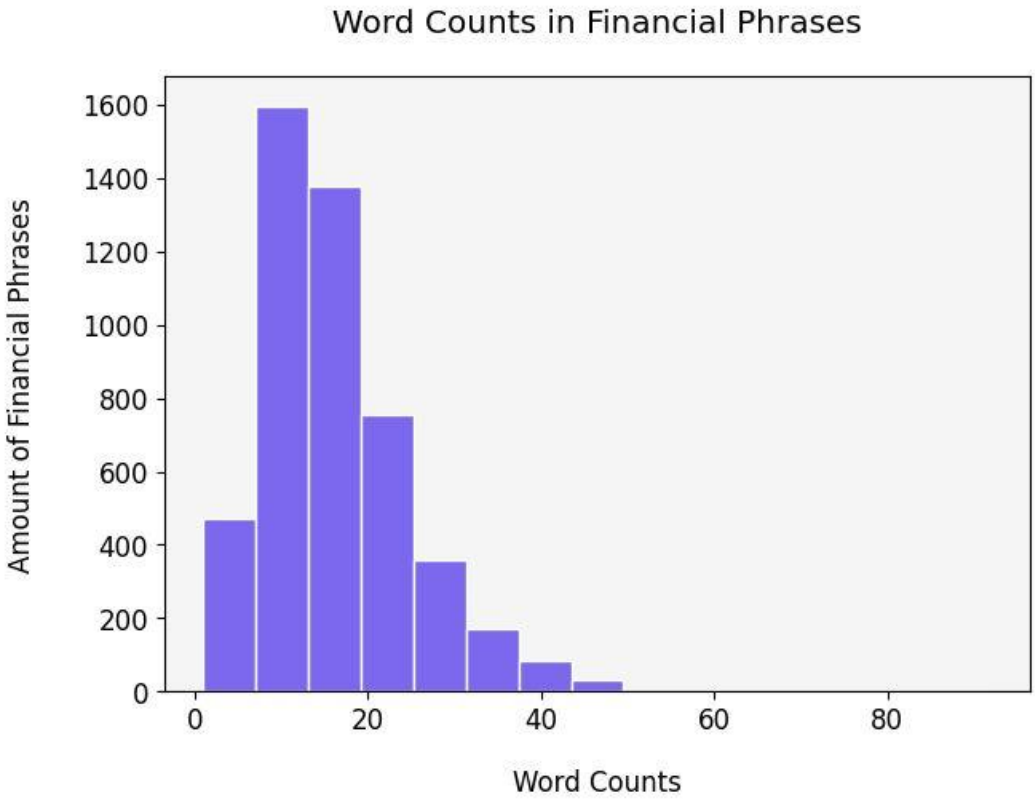
APPENDIX 2.9 Sentiment-Related Features Correlation Heatmap

The plot shows the correlation between three features in the dataset: 'word counts', 'length', and 'label'. In this case, the correlation between 'word counts' and 'length' is very high at 0.97, suggesting that the length of the financial phrases is highly correlated with the number of words used in the phrase. The correlation between 'word counts' and 'label' is lower at 0.18, indicating a weak positive correlation, while the correlation between 'label' and 'length' is also weak at 0.16. 'Word counts' refers to the number of words in a financial phrase, while 'length' refers to the number of characters in the phrase.



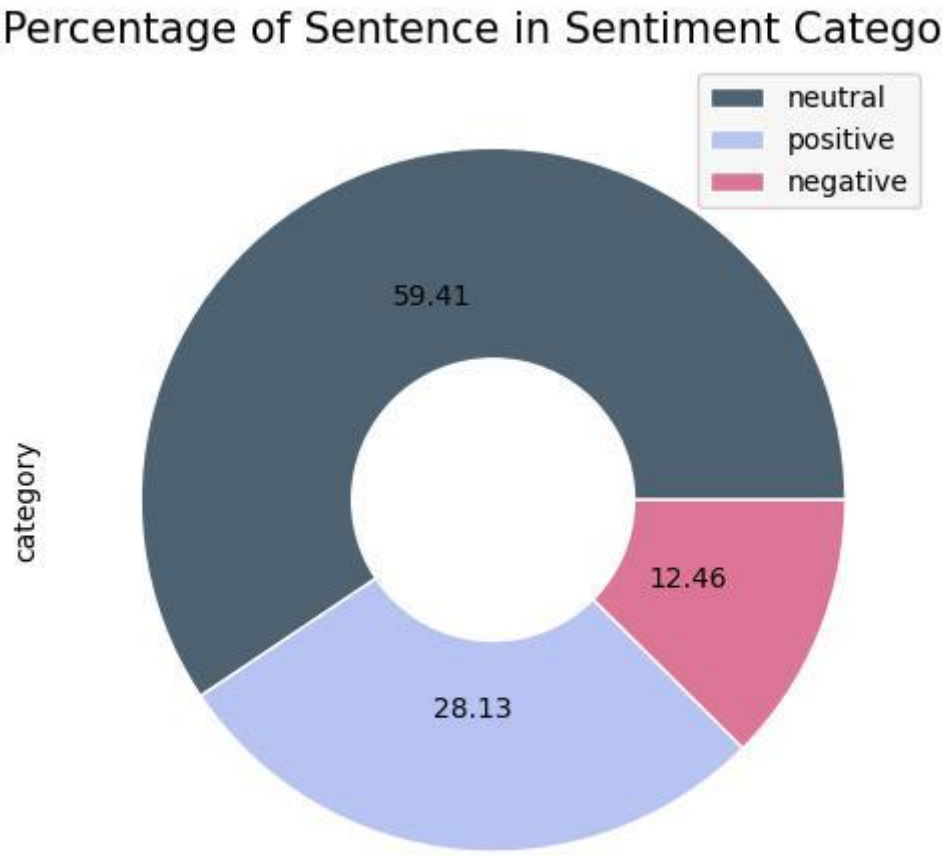
APPENDIX 2.10 Word Counts in Financial Phrases

The histogram represents the distribution of financial phrases by their word count. The highest frequency occurs at word counts of around 6-13 with 1600 phrases, followed by word counts of around 13-20 with 1400 phrases. The third highest peak occurs at word counts of around 20-26 with 750 phrases. The distribution shows a long-tail pattern with decreasing frequency as the word count increases.



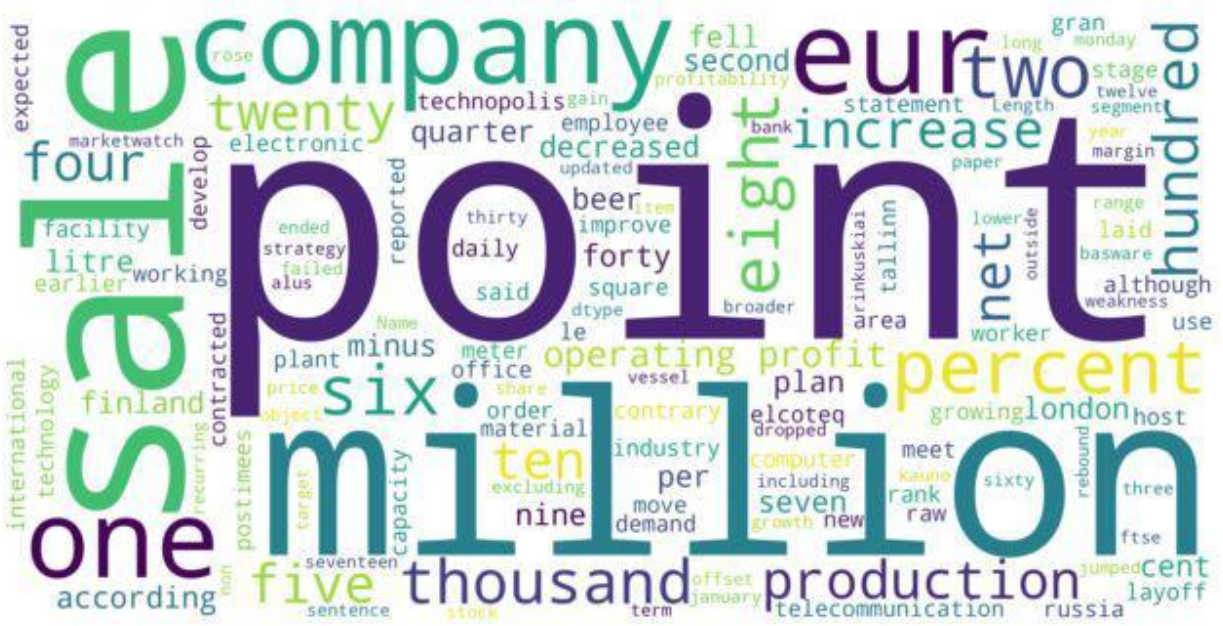
APPENDIX 2.11 Percentage of Sentence in Sentiment Category

The pie chart shows the distribution of sentiment categories in the dataset. The majority of sentences are labelled as neutral (59.41%), followed by positive (28.13%) and negative (12.46%). This indicates that the dataset is imbalanced towards neutral sentiment. It is important to note that imbalanced datasets can impact the performance of sentiment analysis models and should be taken into consideration during model development and evaluation.



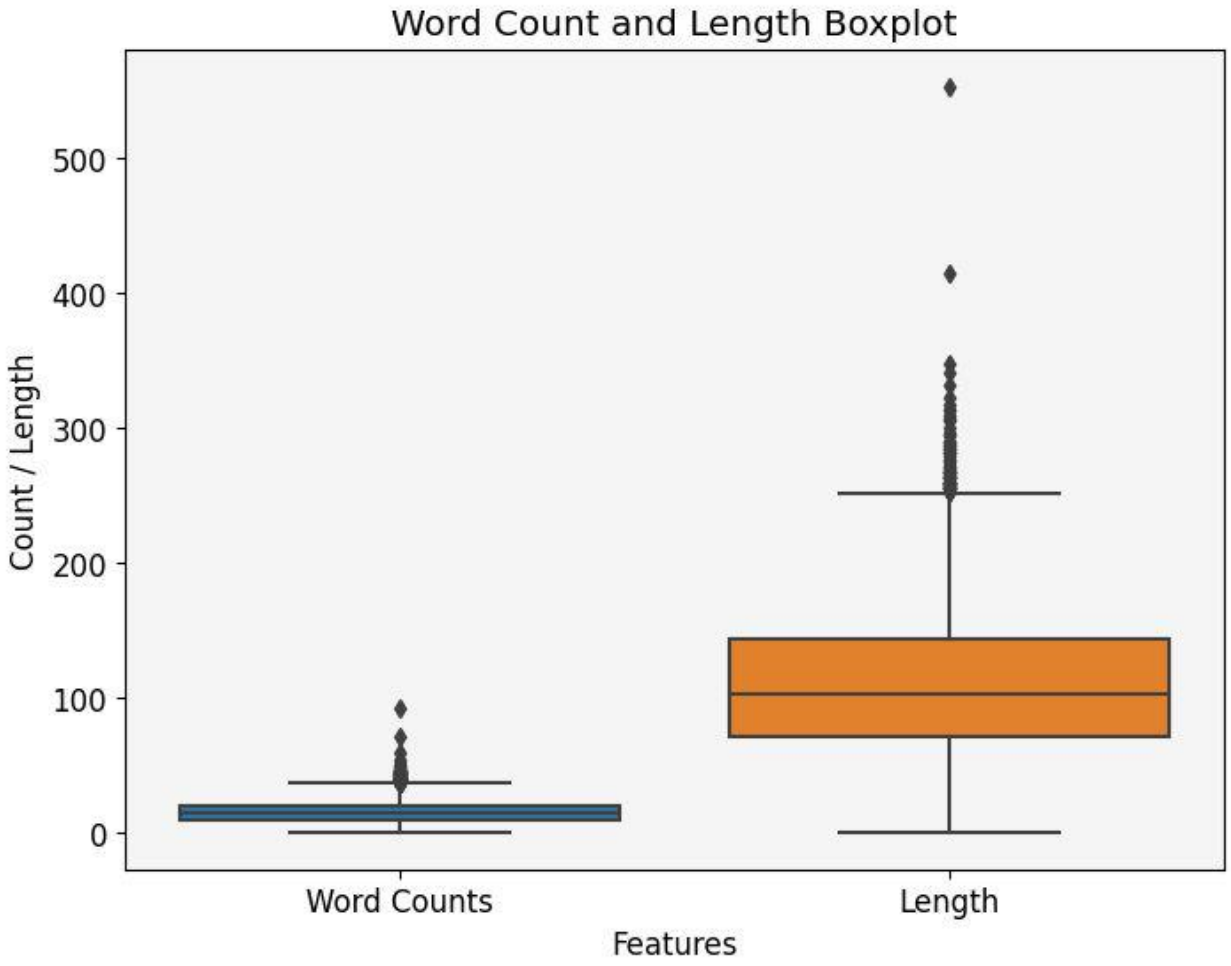
APPENDIX 2.12 Word Cloud

The word cloud can provide a quick overview of the most common words in a dataset. The size of the words indicates their frequency, with larger words appearing more often. In this dataset, the most common word is 'point', followed by 'sale', 'million', 'eur', 'company', 'one', and others.



APPENDIX 2.13 Word Count and Length Boxplot

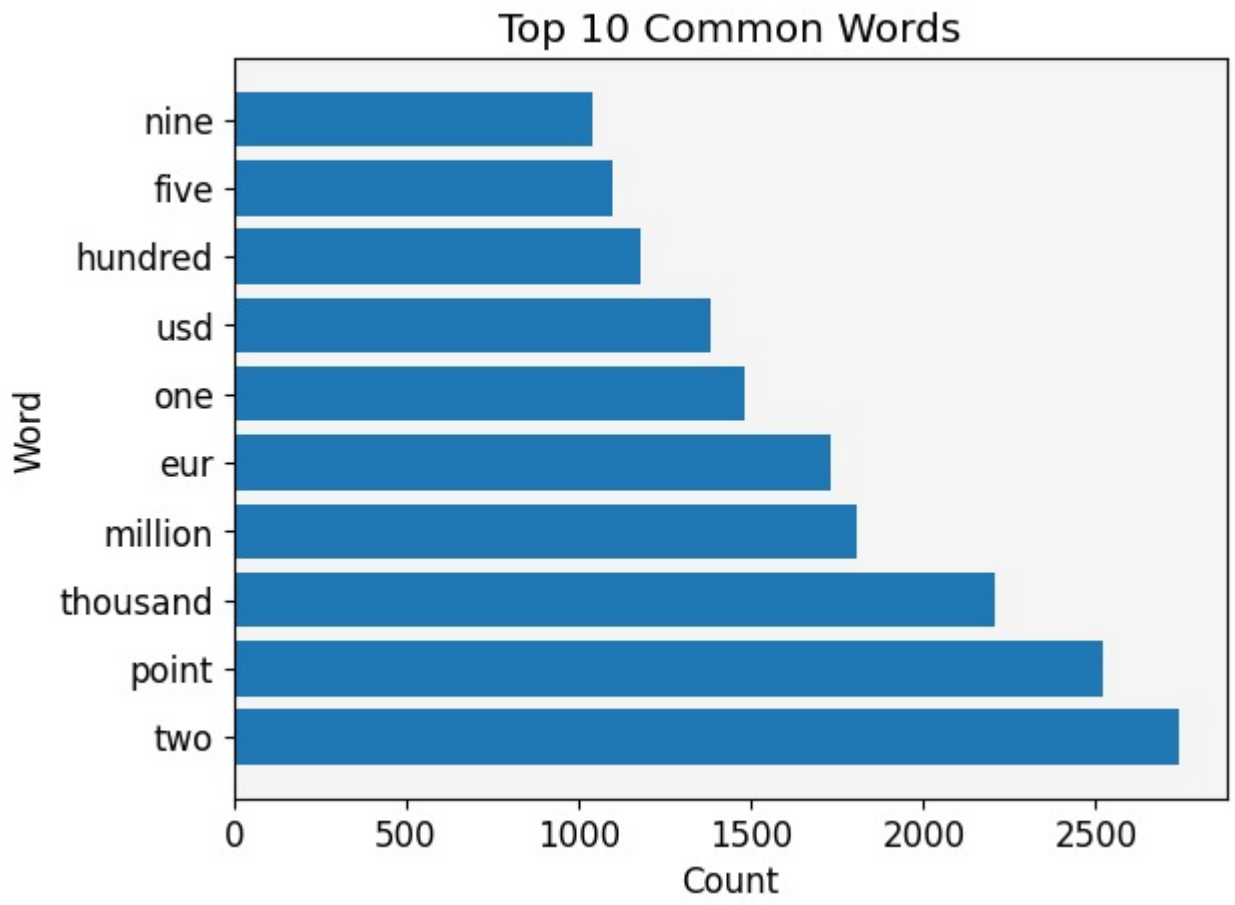
The boxplot shows the distribution of word counts and length in the dataset. The mean word count is around 16, with a minimum of 1 and maximum of 92. The mean sentence length is around 113, with a minimum of 1 and maximum of 553. Both distributions show a large spread of data, with many outliers. The boxplot also indicates that the distribution of word counts is slightly more skewed than the distribution of length.



APPENDIX 3: EXPLORATORY DATA ANALYSIS OF DATASET_2

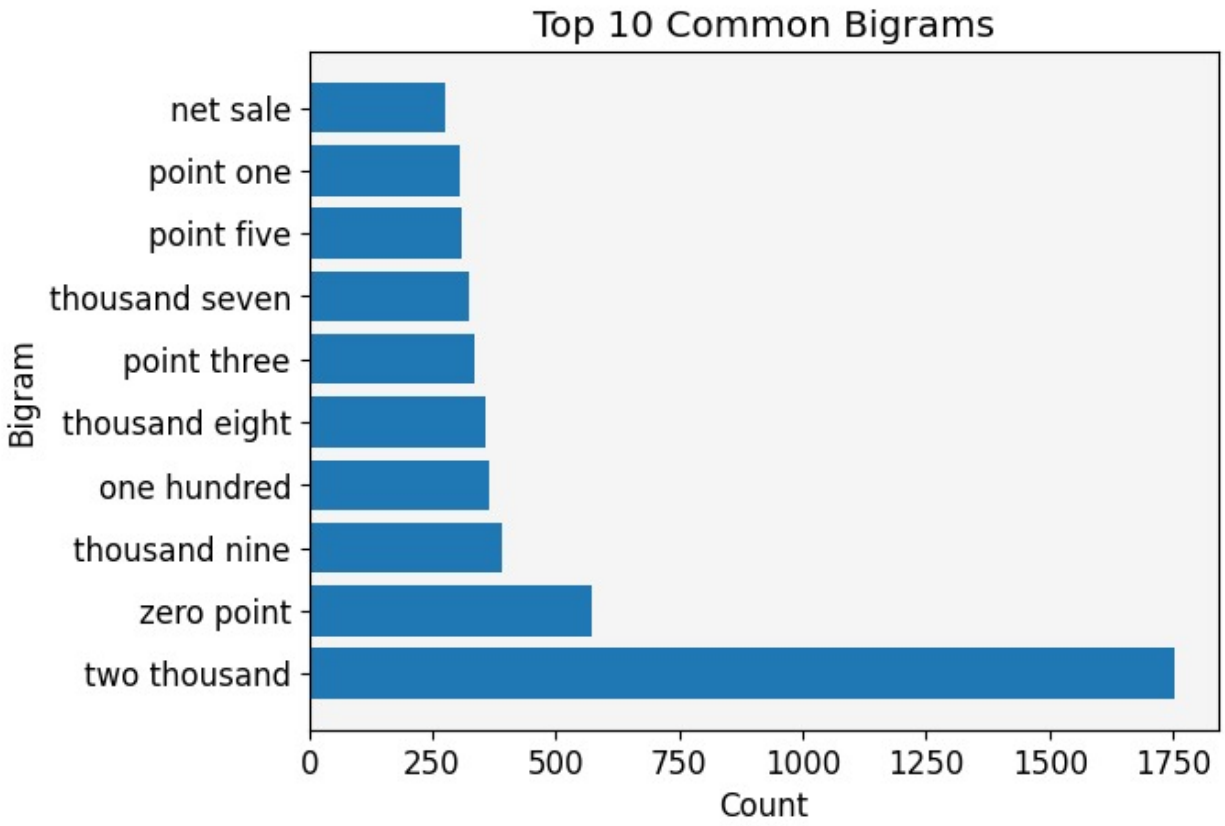
APPENDIX 3.1 Top 10 Common Words

The plot shows the top 10 most common words in the dataset. The most frequently occurring word is 'two' with over 2500 counts, followed closely by 'point' and 'thousand' with around 2500 counts and 2200 counts each. The remaining words have much lower counts, ranging from 1000 to 2000.



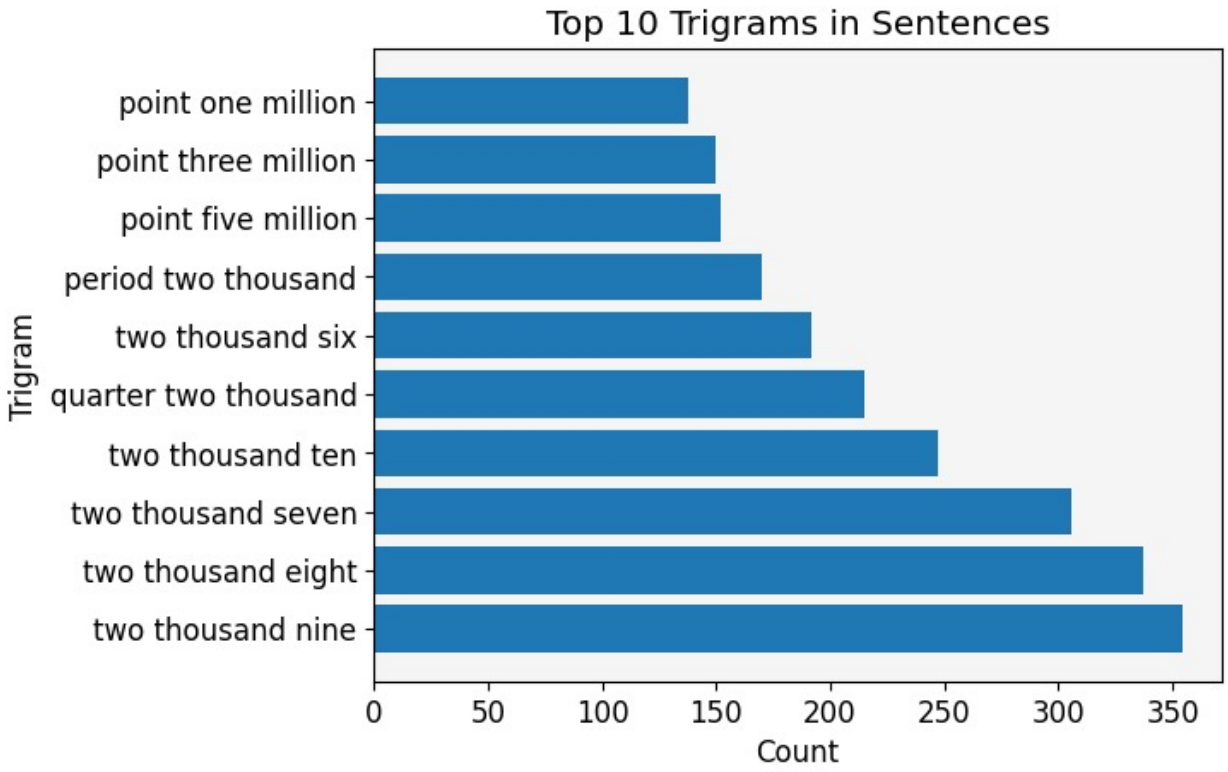
APPENDIX 3.2 Top 10 Common Bigrams

The bar plot illustrates the top 10 most common bigrams in the dataset, with 'two thousand' being the most frequent one with approximately 1750 counts. 'zero point' and 'thousand nine' are the next two most common bigrams with around 600 and 400 counts, respectively. The presence of these bigrams suggests that the dataset might be related to years or financial transactions, where phrases such as 'two thousand' and 'zero point' are commonly used. The rest of the bigrams are less frequent, with counts ranging from 250 to 350.



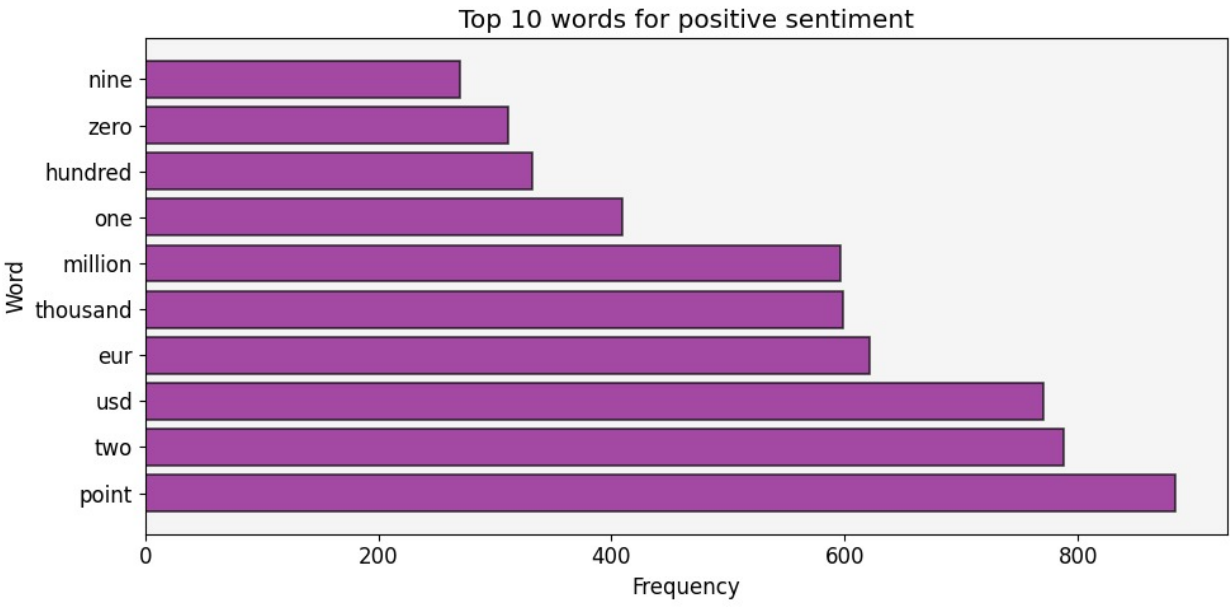
APPENDIX 3.3 Top 10 Common Trigrams

The bar plot displays the top 10 most common trigrams in the dataset, with 'two thousand nine' being the most frequent one with over 350 counts. 'Two thousand eight' and 'two thousand seven' are the next two most common trigrams with around 350 and 300 counts, respectively. The high frequency of these trigrams suggests that the dataset might be related to financial or economic data where years and quarters are commonly used. The remaining trigrams have lower frequencies, ranging from 100 to 250 counts.



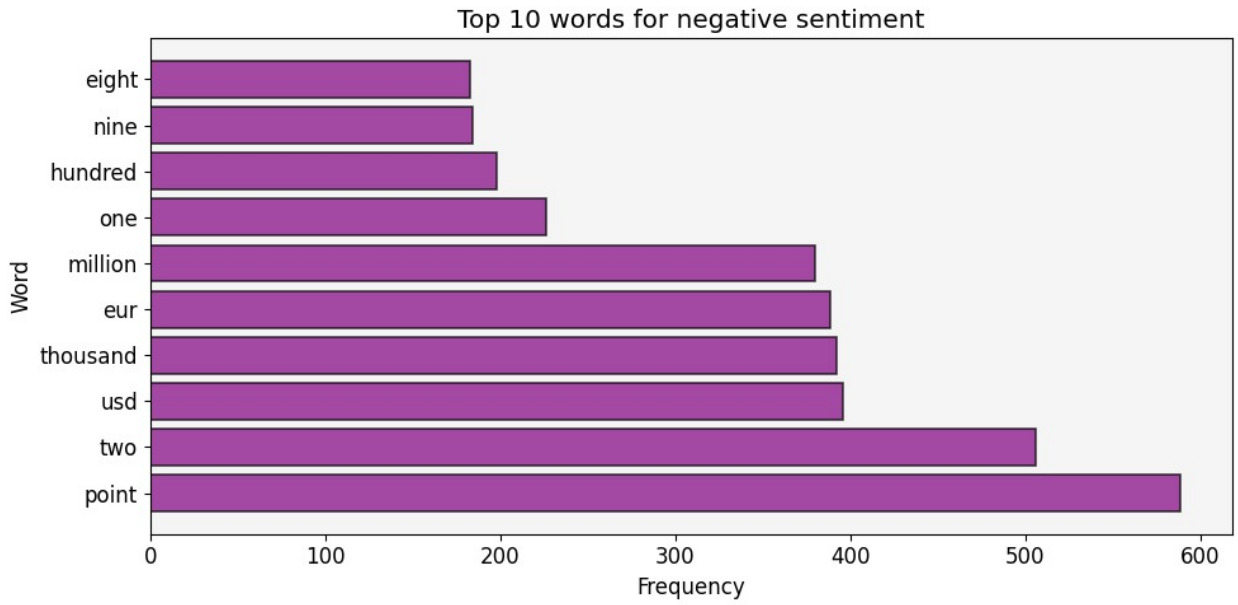
APPENDIX 3.4 Top 10 Words for Positive Sentiment

The bar plot illustrates the top 10 most common positive sentiment words in the dataset, with 'point' being the most frequent one with over 800 counts. 'two' and 'usd' are the next two most common words with around 800 counts, respectively. The rest of the words are less frequent, with counts ranging from 200 to 650.



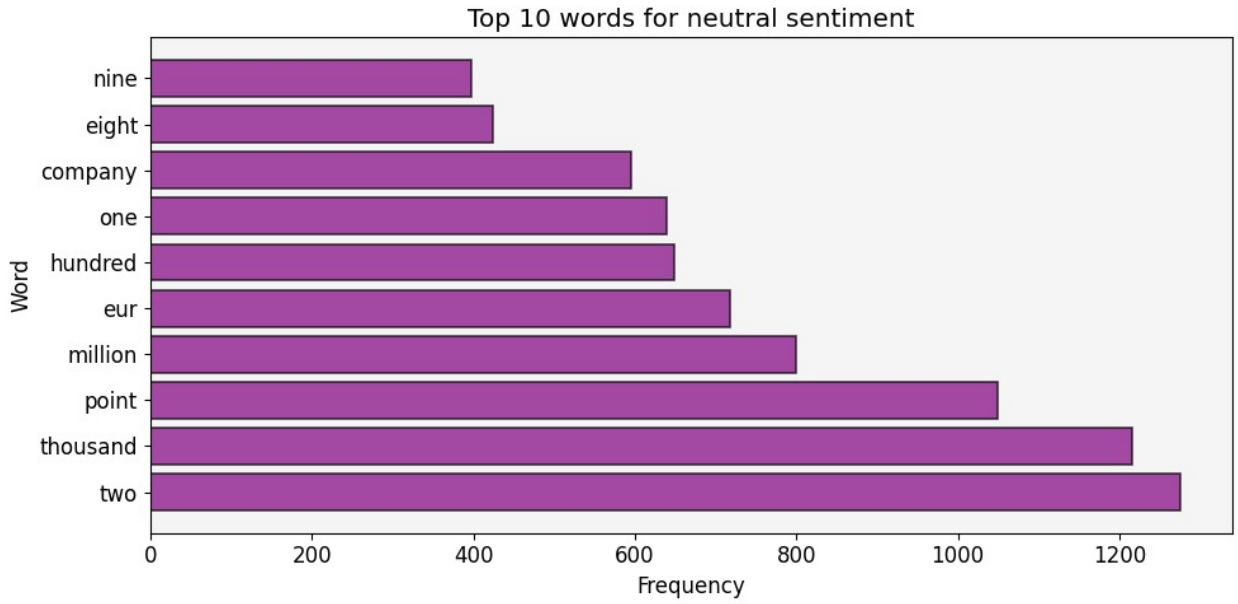
APPENDIX 3.5 Top 10 Words for Negative Sentiment

The plot displays the top 10 most common negative sentiment words in the dataset. The most frequently occurring word is 'point' with around 600 counts, followed by 'two' and 'usd' with around 500 and 400 counts, respectively. The remaining words have lower counts, ranging from 150 to 400.



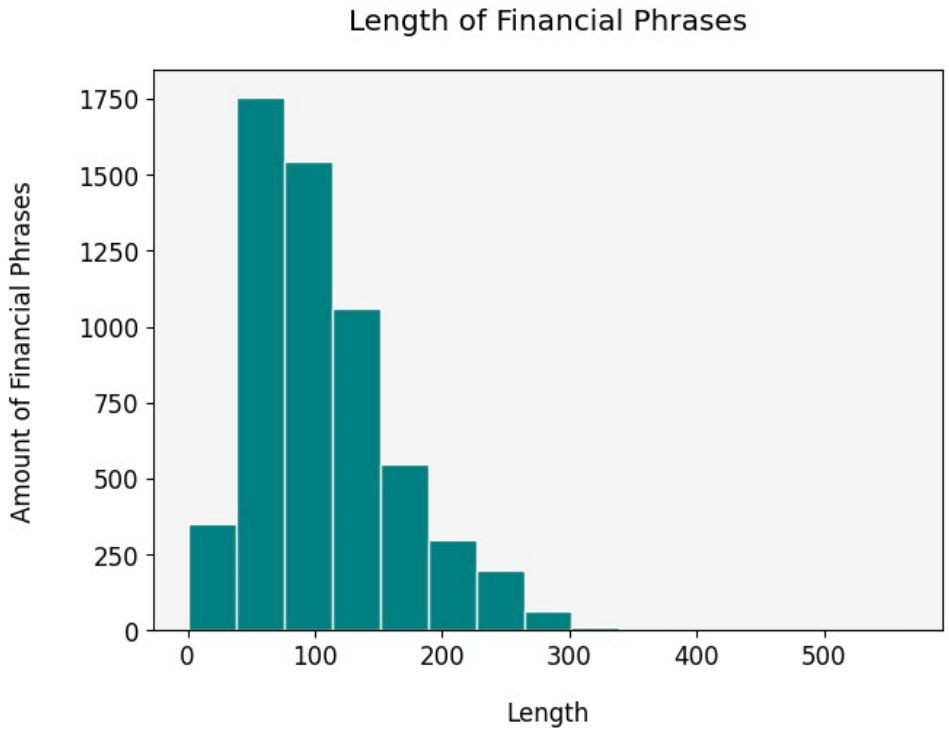
APPENDIX 3.6 Top 10 Words for Neutral Sentiment

The plot displays the top 10 most common neutral sentiment words in the dataset. The most frequently occurring word is 'two' with over 1200 counts, followed by 'thousand' and 'point' with around 1200 and 1000 counts, respectively. The remaining words have lower counts, ranging from 300 to 800.



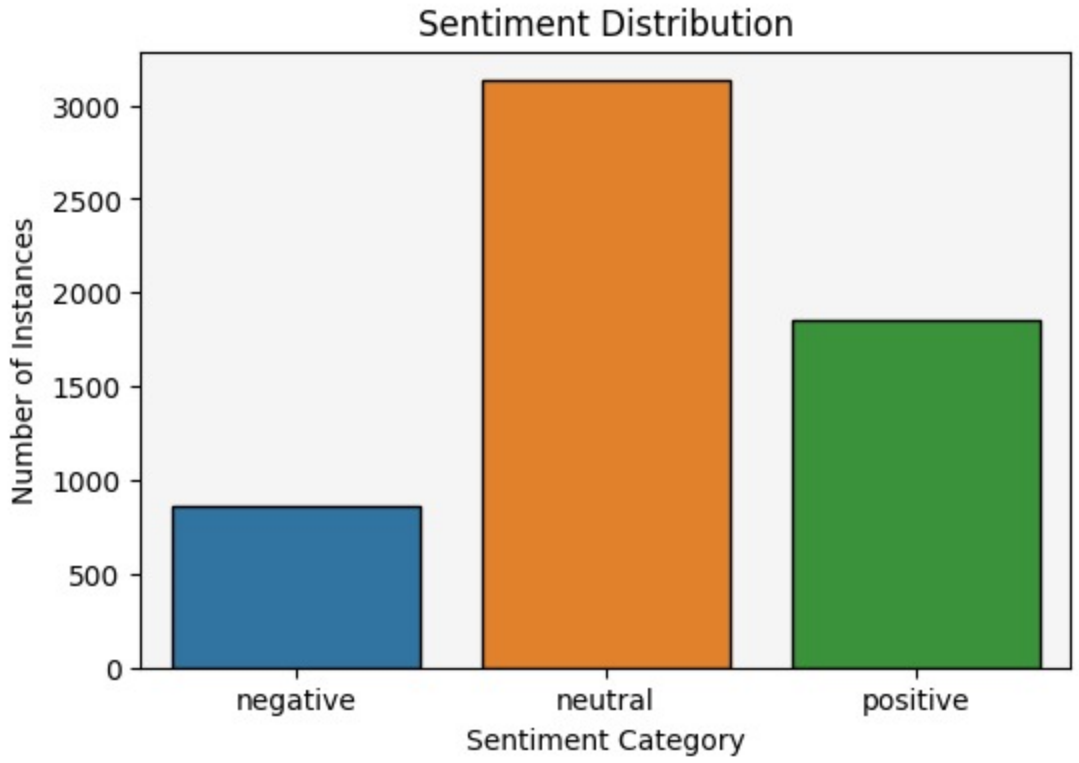
APPENDIX 3.7 Length of Financial Phrases

The histogram shows a distribution of financial phrases by their length. The most common length is around 50 with 1750 phrases, followed closely by length of around 100 with 1500 phrases. The third highest peak occurs at around 120 with 1000 phrases. The distribution shows a long-tail pattern with decreasing frequency as the length increases.



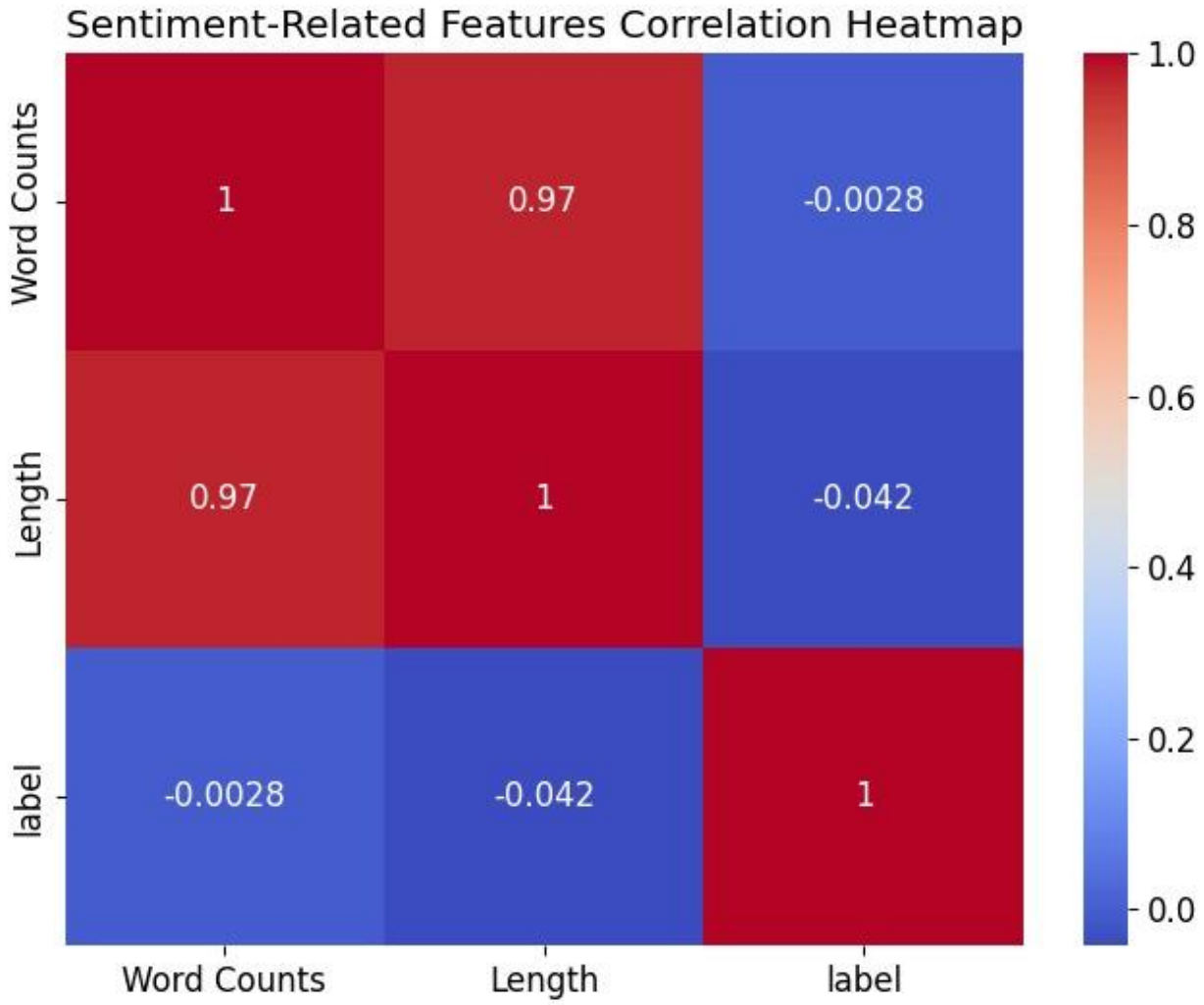
APPENDIX 3.8 Sentiment Distribution

The barplot shows the distribution of sentiment categories in the dataset, with 'neutral' being the most frequent sentiment category with around 3000 instances. 'Positive' and 'negative' sentiments are the next two most common categories, with around 2000 and 850 instances, respectively. The plot suggests that the majority of instances in the dataset are classified as neutral.



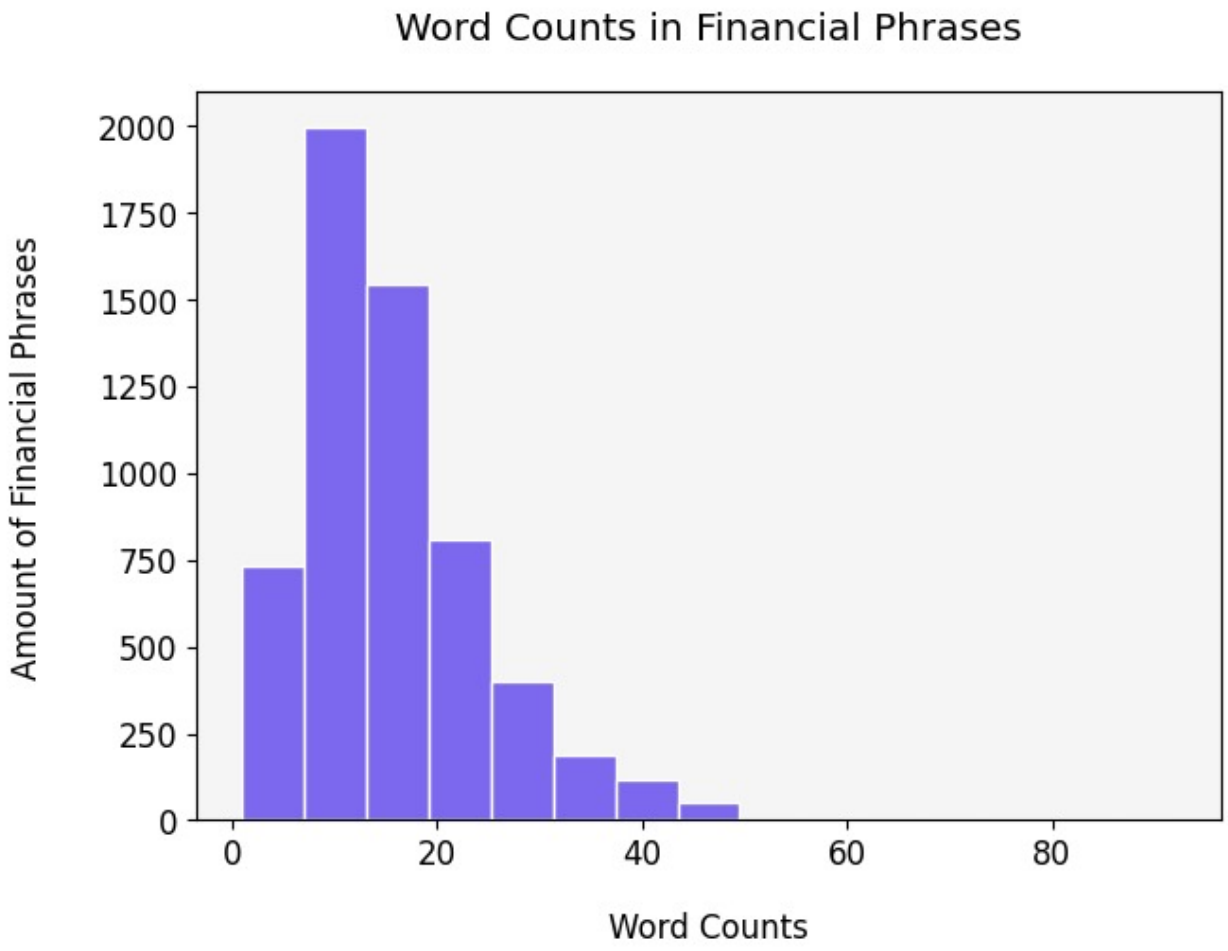
APPENDIX 3.9 Sentiment-Related Features Correlation HeatMap

The plot shows the correlation between three features in the dataset: 'word counts', 'length', and 'label'. In this case, the correlation between 'word counts' and 'length' is very high at 0.97, suggesting that the length of the financial phrases is highly correlated with the number of words used in the phrase. The correlation between 'word counts' and 'label' is lower at -0.0028, indicating a weak negative correlation, while the correlation between 'label' and 'length' is also weak at -0.042.



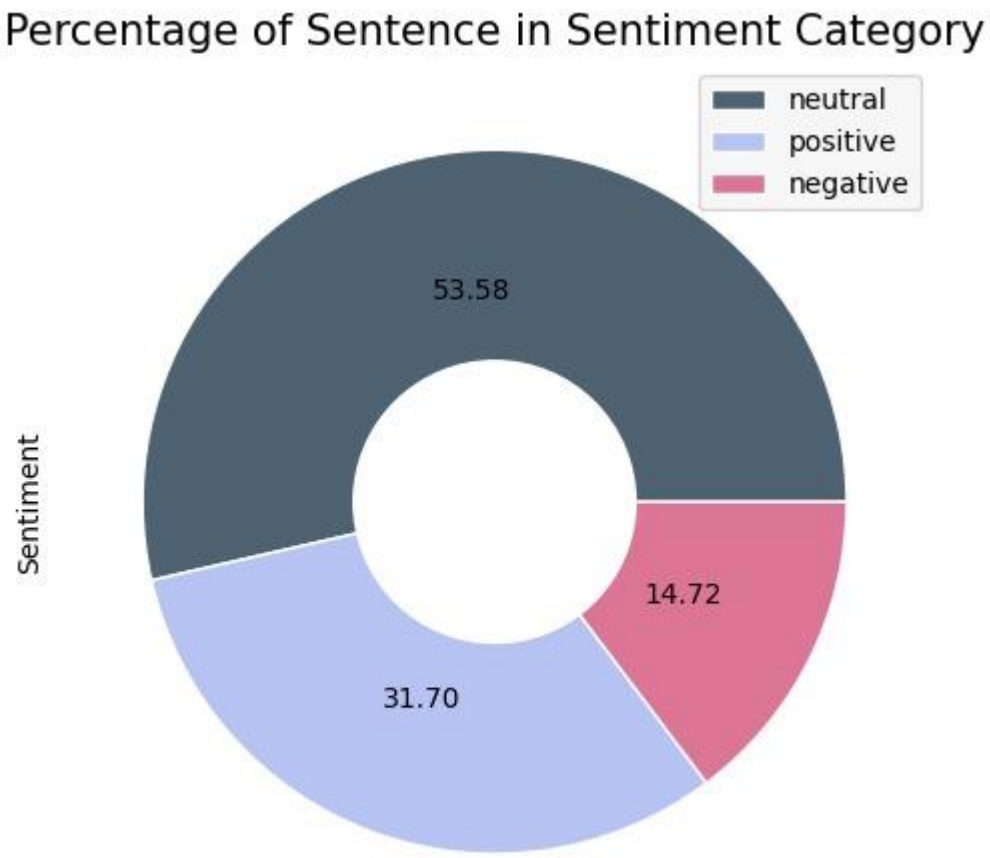
APPENDIX 3.10 Word Counts in Financial Phrases

The histogram represents the distribution of financial phrases by their word count. The highest frequency occurs at word counts of around 6-13 with 2000 phrases, followed by word counts of around 13-20 with 1500 phrases. The third highest peak occurs at word counts of around 20-26 with 750 phrases. The distribution shows a long-tail pattern with decreasing frequency as the word count increases.



APPENDIX 3.11 Percentage of Sentence in Sentiment Category

The pie chart shows the distribution of sentiment categories in the dataset. The majority of sentences are labeled as neutral (53.58%), followed by positive (31.70%) and negative (14.72%). This indicates that the dataset is also imbalanced towards neutral sentiment. It is important to note that imbalanced datasets can impact the performance of sentiment analysis models and should be taken into consideration during model development and evaluation.

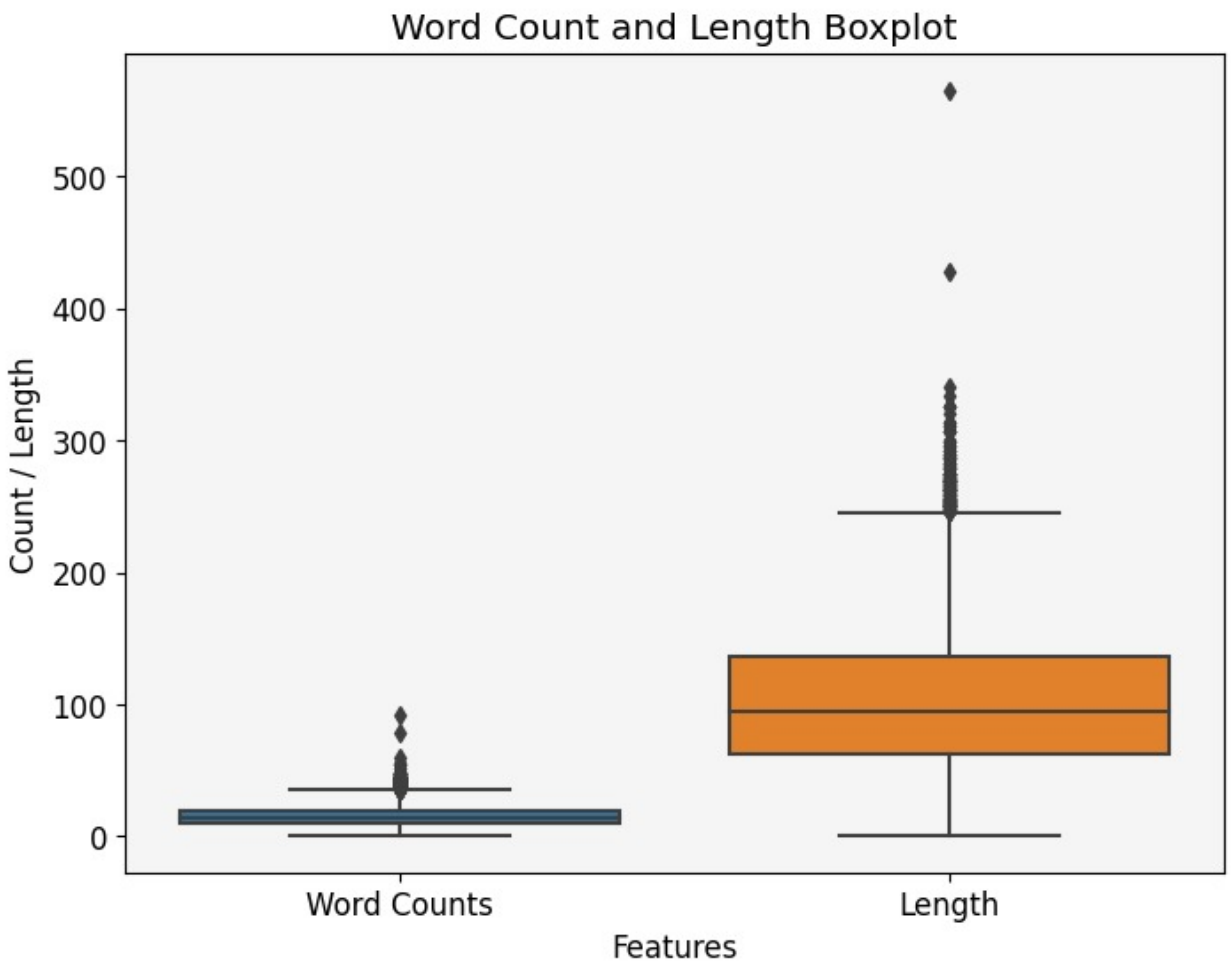


APPENDIX 3.12 Word Cloud

The word cloud can provide a quick overview of the most common words in a dataset. The size of the words indicates their frequency, with larger words appearing more often. In this dataset, the most common word is 'point', followed by 'two', ‘one’, 'eur', 'million', ‘usd’, 'five', and others.

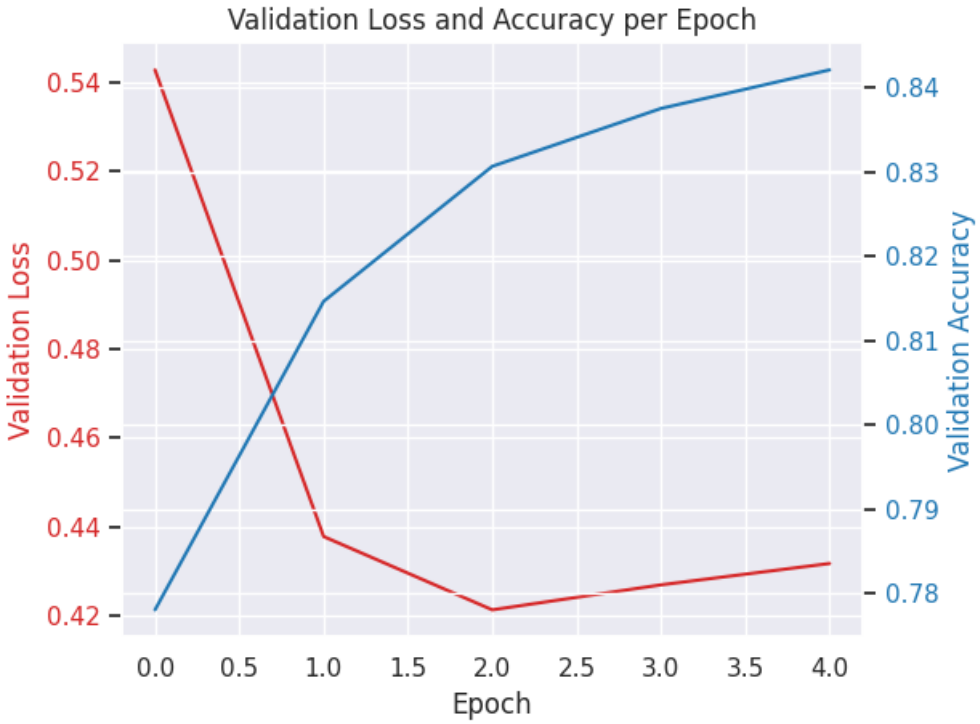
APPENDIX 3.13 Word Count and Length Boxplot

The boxplot shows the distribution of word counts and length in the dataset. The mean word count is around 16, with a minimum of 1 and maximum of 92. The mean sentence length is around 107, with a minimum of 1 and maximum of 565. Both distributions show a large spread of data, with many outliers. The boxplot also indicates that the distribution of word counts is slightly more skewed than the distribution of length.

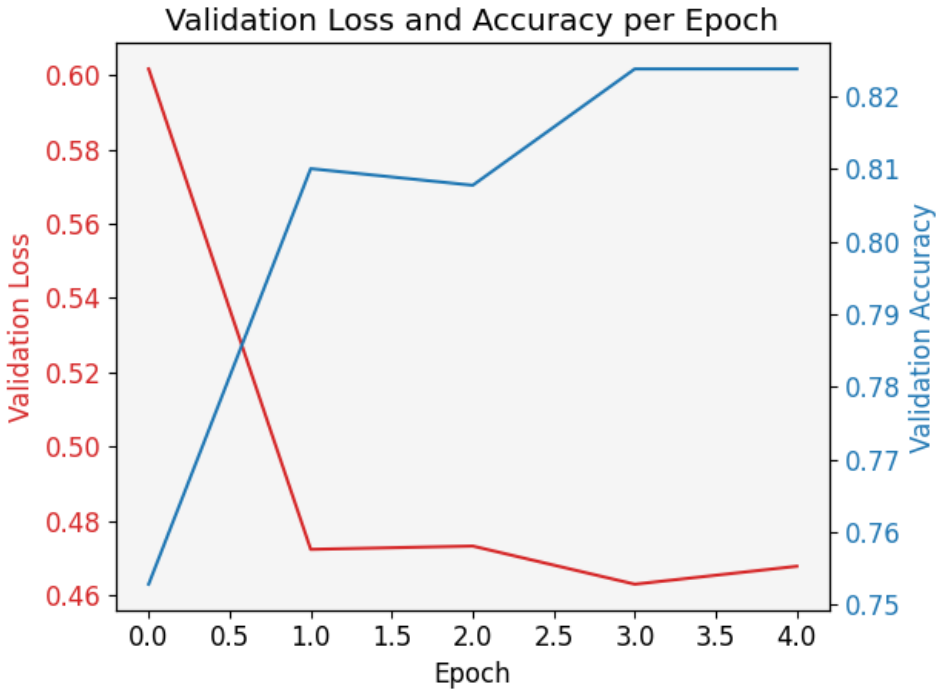


APPENDIX 4: TRAINING RESULT OF BERT AND FINBERT

APPENDIX 4.1: TRAINING RESULT OF FINBERT ON DATASET_1



APPENDIX 4.2: TRAINING RESULT OF BERT ON DATASET_1



APPENDIX 4.3: TRAINING RESULT OF FINBERT ON DATASET_2

APPENDIX 4.4: TRAINING RESULT OF BERT ON DATASET_2