Twitter Financial News Classification

Abstract

Sentiment analysis on Twitter financial news is crucial for understanding market trends and investor sentiments. In this research, we explore six distinct methodologies to predict sentiment in the Twitter Financial News dataset, leveraging a modest set of 32 labelled examples. Our experimentation spans conventional techniques like TF-IDF and advanced deep learning models including distilBERT and LLM Text Generated Model. Additionally, we introduce innovative approaches such as Data Augmentation and Combined Technique Model, which amalgamates BERT with augmentation strategies. Through rigorous evaluation against benchmark datasets, we identify the Combined Technique Model as the most effective in sentiment prediction, showcasing superior performance across specified metrics. This study not only contributes to advancing sentiment analysis in financial domains but also underscores the significance of integrating diverse methodologies for robust predictive outcomes.

Dataset Background

The Twitter Financial News dataset is a curated collection of English-language tweets centred around financial topics. This dataset is carefully hand-labelled for sentiment analysis purposes, categorising sentiments into three distinct groups: 'Bearish' (LABEL_0), 'Bullish' (LABEL_1), and 'Neutral' (LABEL_2). Acquired via the Twitter API, it is tailored for multi-class classification challenges. The dataset is partitioned into two segments: the training set comprises 9,938 tweets (making up 80% of the dataset), while the validation set includes 2,486 tweets (accounting for 20%). Each category encapsulates a broad spectrum of financial contexts, as illustrated in Figure 1.

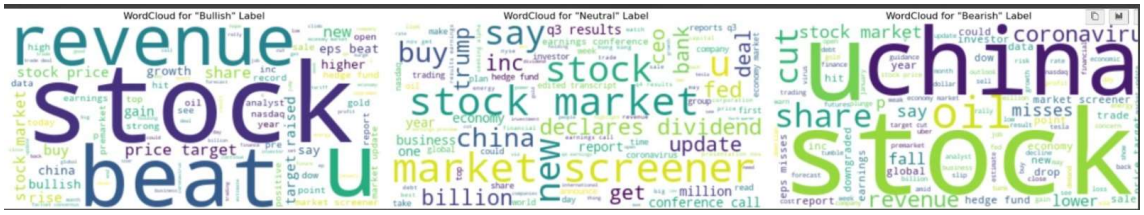
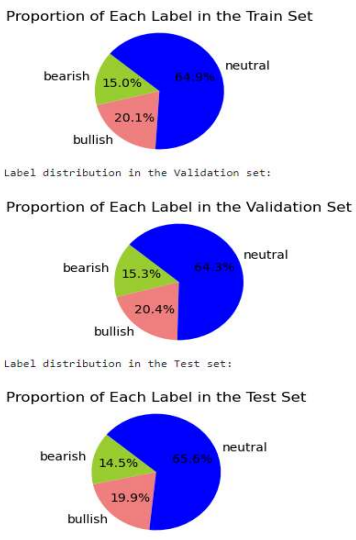


Figure 1: Wordcloud of labels Bullish, Neutral & Bearish

The dataset we used comprised only a training and validation set, and we observed an imbalance in the labels, with a dominance of tweets classified as neutral.



To adapt the dataset for our modelling needs, we created a test dataset from the existing training set.

For data preprocessing, we employed a package called Ekphrasis[1], specifically developed for social media text and online communication. We selected this tool for its effectiveness in processing social media texts and accurately capturing informal

[1] Cbaziotis ekphrasis [Source code repository]. GitHub. https://github.com/cbaziotis/ekphrasis

language, abbreviations, hashtags, and emoticons, enabling a more thorough text analysis.

The dataset we used comprised only a training and validation set, and we observed an imbalance in the labels, with a dominance of tweets classified as neutral.
To adapt the dataset for our modelling needs, we created a test dataset from the existing training set.

Figure 2: Proportion of text in each label

Methodology:

Preprocessing:

Initially, we carefully examined the dataset and addressed any disparities in the frequency of each sentiment. Subsequently, we streamlined the text to enhance comprehensibility. This involved converting everything to lowercase, eliminating unnecessary symbols, and identifying and removing infrequent emoticons and hashtags. However, as these elements were rare, we opted to exclude them. For data preprocessing, we employed a package called Ekphrasis, specifically developed for social media text and online communication. We selected this tool for its effectiveness in processing social media texts and accurately capturing informal language, abbreviations, hashtags, and emoticons, enabling a more thorough text analysis.
We attempted to rebalance the dataset due to a larger presence of neutral labels, but this did not improve predictions, so we decided to keep the original distribution.

Random Classifier and Baseline Model:

Our initial approach involved employing a random classifier to gauge its predictive performance. As expected, it achieved an accuracy of approximately 33%. Following this, we utilised a Spacy matcher to identify word patterns associated with specific sentiments. This approach yielded a slightly improved accuracy compared to the random classifier, but still not comparable to more advanced techniques.

BERT Models:

To simulate a scenario where labelled data is scarce, we selected only 32 labelled examples and treated the rest as unlabelled. Initially, we compared a baseline TF-IDF model with a BERT model that required no additional preprocessing. For the BERT model, we utilised a pre-trained distilBERT architecture to compute embeddings, followed by using the [CLS] token with self-attention scores to make predictions through a single Dense layer with softmax activations.

In an effort to enhance performance, we experimented with integrating an RNN layer on top of BERT. However, upon evaluation, we observed a degradation in prediction accuracy.

Data Augmentation:

Following the initial modelling with a restricted dataset, we investigated various methodologies to expand the available data.

Translation and Backtranslation:

In this approach, we translated the text data into French and then back into English.

Word2Vec Similarity:

Utilising this method, we synthesised additional data points by replacing some terms with their nearest words using a pretrained word2vec model.

Synonym Replacement:

Similarly, we augmented the data by generating synonyms using the 'wordnet' library from NLTK.

Random Swap:

Lastly, we employed a technique involving the random swapping of a fixed number of words in the sentence, altering their order in the process.

After augmenting the dataset using standard techniques, we conducted another round of modelling with the augmented data. This approach demonstrated a noticeable improvement in the results.

LLM text Generation:

For a more sophisticated approach to data augmentation, we employed text generation using Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG). Initially, we engaged in prompt engineering to fine-tune the prompt for enhanced quality, relevance, and coherence. Subsequently, we supplied the optimised prompt along with the 32 labelled examples to the LLM, enabling it to conduct Retrieval-Augmented Generation (RAG). We stored the newly labelled examples and ran another BERT model on the augmented data.

Optimal technique:

For the Optimal technique we opted for combining the augmented data with LLM text generation with the augmented data with standard techniques. After trying several combinations for the data we obtained the highest accuracy combining all the generated datasets in a single augmented dataset.

State of the Art Model

Analysing the sentiment of financial tweets has numerous business uses. Companies leverage this information to monitor their public image, while investors pay attention to sudden sentiment changes as indicators of possible risks or market instability. In the realm of algorithmic trading, traders integrate sentiment data into their strategies, making trades based on specific rules that respond to shifts in sentiment. The State-of-the-Art model (SoA) we have chosen for benchmarking our research is a paper written by Stanford University titled "Tweet Sentiment Analysis to Predict Stock Market[2]." This study utilised the same Twitter corpus in their model and incorporated a Finetuned FinancialBERT (FinBERT), yielding an accuracy and F1 score of 0.843. We believe this paper aligns well with our study, and our goal is to either match or closely approach these performance metrics.

Results

32 observations:

Model	Accuracy	Precision	Recall	F1-Score

[2] Palomo, C. Tweet Sentiment Analysis to Predict Stock Market. Stanford University.

Baseline TF-IDF	44.81	30.61	36.46	30.32
distilBERT model	43.17	43.67	47.61	40.88
distilBERT with RNN	36.31	39.32	40.95	34.00
Data Augmentation	48.16	46.91	50.35	43.59
LLM Text Generated	50.92	44.28	48.26	44.13
Combined Technique	54.06	44.57	48.07	43.54

While working on the challenge with limited labelled data (just 32 labelled examples) we found that data augmentation and LLM Text Generated approaches give better results compared to a baseline TF-IDF approach and distilBERT model.

Between Data augmentation and LLM Text Generated approaches, Data augmentation performed better in terms of precision and recall, while the LLM Text Generated one performed better in terms of accuracy and F1-score. Data augmentation typically focuses on increasing the diversity of the training data by applying various transformations or generating synthetic samples. This approach tends to improve precision and recall by exposing the model to a wider range of data instances, thereby enhancing its ability to discriminate between classes and reducing the likelihood of false positives and false negatives. On the other hand, Language Model (LM) Text Generated approaches leverage the power of pre-trained language models to generate additional training data. While this approach may improve accuracy and F1-score by providing additional high-quality data instances, it might not necessarily enhance precision and recall to the same extent as data augmentation. This could be because LM-generated text may not cover all possible variations in the data distribution or may introduce noise that affects precision and recall.

Combining data augmentation and LM Text Generated approaches resulted in higher accuracy, but improvements in precision, recall, and F1-score were limited. While the additional diversity from data augmentation and high-quality data from LM Text Generated helped boost accuracy, they may not have fully addressed issues related to false predictions or class imbalance.

Entire dataset:

Model	Accuracy	Precision	Recall	F1-Score
distilBERT model	86.85	83.73	81.27	82.36
Distilled tiny-bert	64.45	55.75	60.58	56.95
SoA	84.3	N/A	N/A	84.3

The distilBERT model trained on the full dataset achieves high accuracy and balanced precision and recall, indicating its effectiveness in correctly classifying instances across all classes. Model performance is in line with the State-of-the-Art results.

The distilled Tiny-BERT model demonstrates lower overall performance compared to DistilBERT, with notably lower accuracy and precision. The lower precision suggests a higher rate of false positives, while the lower recall indicates difficulty in correctly identifying positive instances. At the same time, the inference speed of this model is much faster than the teacher's (distilBERT) model.

Discussion & Conclusion

In this study, we investigated several distinct methodologies for sentiment analysis on Twitter financial news. Our experimentation revealed promising results with data augmentation and LLM Text Generated approaches, showing improvements in predictive performance compared to baseline techniques. While data augmentation techniques enhanced precision and recall by increasing data diversity, LLM Text Generated approaches contributed to higher accuracy and F1-score by generating high-quality data instances.

Furthermore, combining data augmentation with LLM Text Generated approaches resulted in further improvements in accuracy, although challenges related to false predictions and class imbalance persisted.

Evaluation of models trained on the entire dataset highlighted the effectiveness of distilBERT in achieving high accuracy and balanced precision and recall.

Overall, our study underscores the significance of integrating diverse methodologies and addressing class distribution issues to achieve robust predictive outcomes in sentiment analysis of financial news on Twitter. Future research directions could explore further improvements in class rebalancing techniques, more advanced preprocessing steps including emoticons and hashtags or more complicated model architectures for prediction.