

Annotation of Stock Tweets for Market Trends

Vaishnav Mandlik
University of Arizona
vaishnavm@arizona.edu

Abstract

Understanding sentiment on social media, especially around stock market trends, is becoming increasingly relevant for financial forecasting. In this study, we annotated a dataset of stock-related tweets with three sentiment categories: *bullish* (optimistic), *bearish* (pessimistic), and *neutral* (mixed or objective). To ensure high-quality labels, we assessed inter-annotator agreement using metrics like Cohen’s Kappa, Scott’s Pi, and observed agreement percentages. We also explored the performance of zero-shot and few-shot classification models on the annotated data, comparing their effectiveness in handling this type of financial text.

1 Introduction

In financial Industry Sentiment is the most important in share market. Public sentiment towards certain stocks and market movements can be either positive or negative, and it is frequently expressed on social media sites like Twitter. Real-time sentiment analysis and recording can provide analysts and investors with insightful information.

Although machine learning models have demonstrated potential for automating sentiment analysis, the calibre of annotated datasets is a major determinant of their effectiveness. Annotation is an essential step since stock-related tweets frequently contain context-dependent language and nuanced clues. In this work, a dataset of tweets about stocks is annotated, and the effectiveness of machine learning models on these labels is assessed.

My approach involves annotating tweets into three categories—*bullish*, *bearish*, and *neutral*. I evaluated the reliability of the annotation process, explore the challenges in achieving agreement, and assess how zero-shot and few-shot learning models perform on this data.

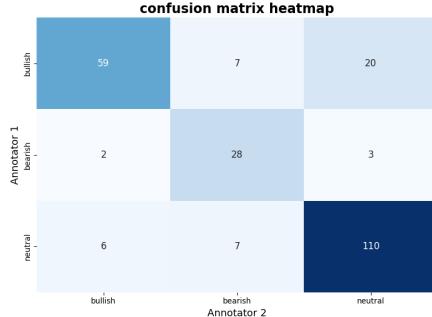


Figure 1: Confusion matrix

2 Methods

2.1 Dataset Creation and Annotation

The dataset consisted of stock-related tweets that were annotated by two human annotators. To ensure consistency, we provided clear definitions and examples for each sentiment category:

- **Bullish:** for positive outlook on the market.
- **Bearish:** If tweet is for negative market perspective.
- **Neutral:** If lack of sentiment.

Annotation guidelines included specific criteria to resolve ambiguity, such as prioritizing the dominant sentiment if multiple interpretations were possible. After independent labeling, disagreements then I produced a final label.

2.2 Measuring Annotator Agreement

To assess annotation quality, we calculated:

- **Cohen’s Kappa:** A score of 0.69 indicated substantial agreement.
- **Scott’s Pi:** A reliability score of 0.69 supported the consistency of annotations.

The confusion matrix (Figure 1) revealed that most disagreements involved tweets classified as

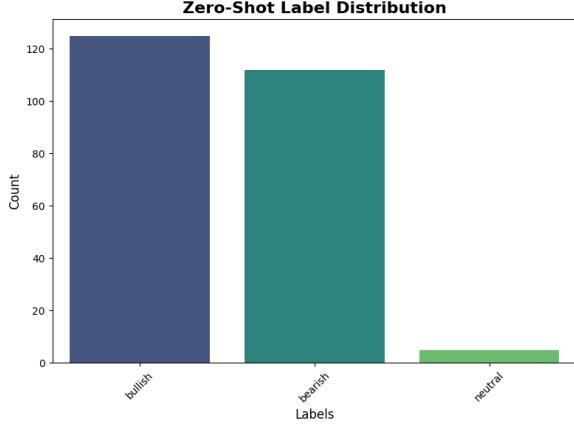


Figure 2: Zero-Shot

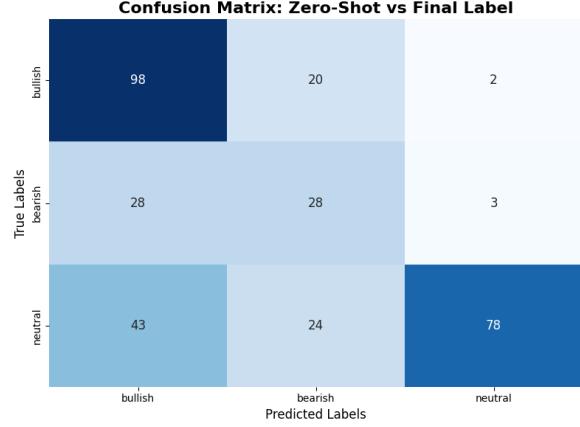


Figure 3: Zero-shots vs Final Label

neutral, reflecting the inherent ambiguity in this category.

3 Results and Analysis

3.1 Exploring the Dataset

The final label distribution showed an expected bias toward *bullish* sentiment, comprising 34.3% of tweets. *Neutral* and *bearish* sentiments accounted for 52.1% and 13.6%, respectively.

3.2 Zero-Shot and Few-Shot Classification

I applied two machine learning approaches to classify the tweets:

1. **Zero-Shot Classification:** Using the `facebook/bart-large-mnli` model, tweets were classified without task-specific training.
2. **Few-Shot Classification:** Fine-tuned a gpt-based model with a small set of annotated examples.

3.3 Results and Evaluation

The dataset's label distribution revealed an expected bias toward *neutral* tweets (52%), followed by *bullish* (34%) and *bearish* (14%) sentiments (Figures 1 and 2). This reflects the prevalence of mixed or factual statements in financial discussions.

3.4 Model Performance

Zero-Shot vs. Final Labels

Zero-shot classification achieved moderate accuracy (63%) against human labels, with the following performance:

- *Bullish*: Precision 0.39, Recall 0.47, F1-score 0.43.

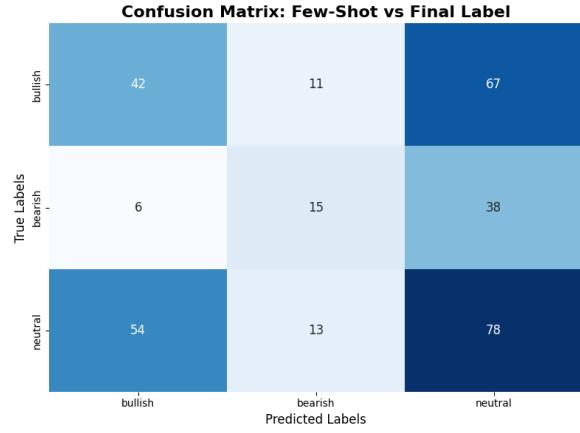


Figure 4: Few-shot vs Final Label

- *Bearish*: Precision 0.58, Recall 0.82, F1-score 0.68.
- *Neutral*: Precision 0.94, Recall 0.54, F1-score 0.68.

Few-Shot vs. Final Labels

Few-shot classification underperformed, with an accuracy of 42%. Performance metrics include:

- *Bullish*: Precision 0.38, Recall 0.25, F1-score 0.31.
- *Bearish*: Precision 0.41, Recall 0.35, F1-score 0.38.
- *Neutral*: Precision 0.43, Recall 0.54, F1-score 0.48.

Zero-shot classification demonstrated reasonable performance without requiring additional training, making it practical for quick deployment. Few-shot learning, while promising, struggled with the limited training examples provided in this study. The

results emphasize the importance of high-quality labeled data in improving model performance.

4 Implications for Financial NLP

1. **Market Trend Analysis:** Detect sentiment shifts to predict market movements.
2. **Investment Strategies:** Use real-time sentiment insights to inform trading decisions.
3. **Enhanced Decision-Making:** Combine sentiment analysis with other financial indicators for a holistic market view.

5 Conclusion

This study highlights the importance of annotated datasets and the role of advanced NLP models in financial sentiment analysis. Zero-shot learning showed moderate success, while few-shot learning requires further refinement. By expanding datasets and improving contextual understanding, social media sentiment analysis can become a powerful tool for market prediction.