Sentiment Analysis of Financial News Using CRISP-DM and Logistic Regression

Aditya Rajpurohit

October 8, 2024

Abstract

In this study, we employ the CRISP-DM framework to perform sentiment analysis on financial news articles. The analysis aims to classify news into positive, neutral, and negative sentiment categories using natural language processing techniques. We trained a Logistic Regression model using TF-IDF vectorization and fine-tuned it through hyperparameter optimization. The model achieved an accuracy of 75.9%, making it a viable tool for understanding market sentiment and supporting decision-making in finance.

1 Introduction

Sentiment analysis is a valuable tool in financial markets, enabling investors and analysts to gauge public opinion and market trends through news articles. This study leverages sentiment analysis to classify financial news using the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology. We use a labeled dataset containing news articles and sentiment labels, aiming to develop a model that predicts the sentiment with high accuracy.

2 Related Work

Previous research in sentiment analysis for finance has employed techniques ranging from simple bag-of-words models to complex deep learning algorithms. This study focuses on a balanced approach using TF-IDF vectorization for feature extraction and Logistic Regression as the classifier, optimized through hyperparameter tuning. The CRISP-DM framework, known for its flexibility, guides the entire process from data preparation to deployment.

3 Methodology

3.1 CRISP-DM Overview

The CRISP-DM framework consists of six phases:

- **Business Understanding:** Define the project's objectives and requirements.
- **Data Understanding:** Collect and explore the dataset.
- **Data Preparation:** Preprocess and clean the text data.
- **Modeling:** Build and train the sentiment classification model.
- **Evaluation:** Assess the model's performance.
- **Deployment:** Save the model for real-world usage.

3.2 Data Description

The dataset contains two columns:

- **Sentiment:** The sentiment label (positive, neutral, negative).
- **News Text:** The content of the financial news articles.

The dataset has been preprocessed by removing special characters, converting text to lowercase, and splitting into training and testing sets.

3.3 Data Preparation

Data preparation included:

- **Text Cleaning:** Removing special characters and converting text to lowercase.
- **TF-IDF Vectorization:** Transforming text data into numerical features.
- **Train-Test Split:** Splitting the data into 80% training and 20% testing subsets.

3.4 Modeling

The initial model was trained using Logistic Regression, followed by hyperparameter tuning using Grid Search. The parameter grid explored different values of regularization strength (C) and solver types.

3.5 Model Evaluation

Model evaluation was based on metrics such as accuracy, precision, recall, and F1-score. The best model achieved an accuracy of 75.9%, with the highest performance for neutral sentiment. Table 1 summarizes the evaluation results.

Class	Precision	Recall	F1-Score	Support
Negative	0.77	0.55	0.64	115
Neutral	0.77	0.88	0.82	567
Positive	0.74	0.61	0.67	287
Accuracy	0.76			
Macro Avg	0.76	0.68	0.71	969

Table 1: Classification Report of the Fine-tuned Model

4 Results and Discussion

The model demonstrated robust performance, with a balanced accuracy across the sentiment classes. The precision, recall, and F1-score were highest for neutral sentiment, which aligns with the inherent nature of financial news that tends to be more factual. Future improvements could include exploring more complex models like SVMs or deep learning techniques and integrating word embeddings to capture context better.

5 Conclusion

This study successfully applied the CRISP-DM methodology to build a sentiment analysis model for financial news. The model can be utilized in various financial applications, including market sentiment tracking, investment analysis, and news monitoring. Future work could focus on real-time implementation and further performance improvements through advanced NLP techniques.

6 Future Work

To enhance the model's performance and applicability, the following strategies could be explored:

- Implementing more sophisticated models such as SVMs or deep learning.
- Expanding the dataset to include a broader range of financial news sources.
- Integrating real-time sentiment analysis into a web application.

Acknowledgments

The authors would like to thank [Data Source] for providing the dataset used in this study.

References

[1] Wirth, R., and Hipp, J., "CRISP-DM: Towards a Standard Process Model for Data Mining", Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 2000.

[2] Jurafsky, D., and Martin, J. H., "Speech and Language Processing", 3rd Edition, 2021.

[3] Pedregosa, F., Varoquaux, G., Gramfort, A., et al., "Scikit-learn: Machine Learning in Python", Journal of Machine Learning Research, 2011.