First Year (Semester-2) Research Assignment on

sentiment analysis on news data to predict the stock price seniment

in partial fulfilment of the requirement for the successful completion of semester 2 of MSc Big Data Analytics

Submitted By

24-PBD-025

Prajapati Meet S

(Semester – II MSc. BDA)

Under the supervision of

Prof.chetan verma

2023-2024

Department of Computer Sciences (MSc. BDA)

St. Xavier's College (Autonomous) Ahmedabad – 380009

DECLARATION

I, the undersigned solemnly declare that the research assignment *sentiment analysis on news data to predict the stock price seniment* is based on my work carried out during the course of our study under the supervision of *prof.chetan verma*. I assert the statements made and conclusions drawn are an outcome of my research work. I further certify that

- The work contained in the report is original and has been done by me under the general supervision of my supervisor.
- The work has not been submitted to any other Institution for any other degree / diploma / certificate in this university or any other University of India or abroad.
- We have followed the guidelines provided by the department in writing the report.

Prajapati Meet S

24-PBD-025

MSc. BDA (Big Data Analytics)

St. Xavier’s College (Autonomous), Ahmedabad

Index

Abstract 4

Introduction 4

Related Work 4

Data Analysis and Insights from the Dataset 5

Conclusion 10

References 11

Abstract

Investor feelings really move the stock market. Luckily, recent improvements in language processing mean we can now pull feelings from news headlines to guess where the market might be headed. This study looks at how the feelings in news headlines relate to stock price changes, using machine learning. We use sentiment analysis to call headlines positive, negative, or neutral and then see how these affect the market. In particular, we check if sentiment-based information can help us predict market trends using models like Support Vector Machines (SVM) and Long Short-Term Memory (LSTM) networks. The results show that overall sentiment by itself might not strongly predict price changes, but certain feelings like fear and joy can predict how some stocks will do. By adding these feelings to existing market indicators, we can get better at guessing market trends. This work adds to the growing area of financial sentiment analysis and shows how we might improve stock predictions.

Introduction

The stock market is always changing and super complicated because lots of things affect it. Stuff like how the economy is doing, what's happening around the world, and how people are feeling about investing all play a part. Figuring out how people feel has become a good way to understand what the market will do since emotions in news, online, and chats often show up before prices change. New tech in language processing and machine learning helps people get these feelings from text and use them to guess where the market is going.

Usually, people guess the market by looking at price changes or checking how well businesses are doing. But these ways don't always catch what people are thinking or how they're acting. Sentiment analysis fixes this by measuring what people think, That way we have a better way to guess what will happen. Studies have found that emotions like fear and happiness can help us guess prices better. Machine learning models, like SVM and LSTM, are used to mix feeling data with old market info, which makes predictions better.

The remainder of this paper is organized as follows: Section 2 reviews related work on sentiment analysis for stock prediction, Section 3 discusses the dataset and methodology used, Section 4 presents the results and analysis, and Section 5 concludes with future research directions.

Related Work

In the last few years, people have gotten way better at using how news headlines feel to guess what the stock market will do. Researchers have been trying out stuff like big language models, special finance-focused models, and transformer thingies.

Kirtac and Germano (2024) checked out how well big language models could figure out finance feelings and guess stock returns. They looked at over 965,000 news articles from 2010 to 2023 and saw that models like OPT, which is like GPT-3, were right about 74.4% of the time when guessing feelings. The cool part was that OPT's feeling scores matched up pretty well with how stocks did the next day, even better than old-school methods.

Gu et al. (2024) came up with FinBERT-LSTM, which mixes the FinBERT language model with LSTM networks to guess stock prices. By adding in news stuff about the market and past stock prices, their way was more on point. They taught the model using NASDAQ-100 stock info and Benzinga news articles, and it beat regular Deep Neural Network (DNN) models.

Kaeley et al. (2023) suggested a transformer model that uses stock data and feeling analysis to guess stock trends over time. They made a new dataset with daily stock info and top news headlines from almost three years. They learned that adding feeling analysis from news headlines made their guesses more accurate, mainly over longer times. They got about 18.63% better over 30 business days than with Recurrent Neural Networks (RNNs).

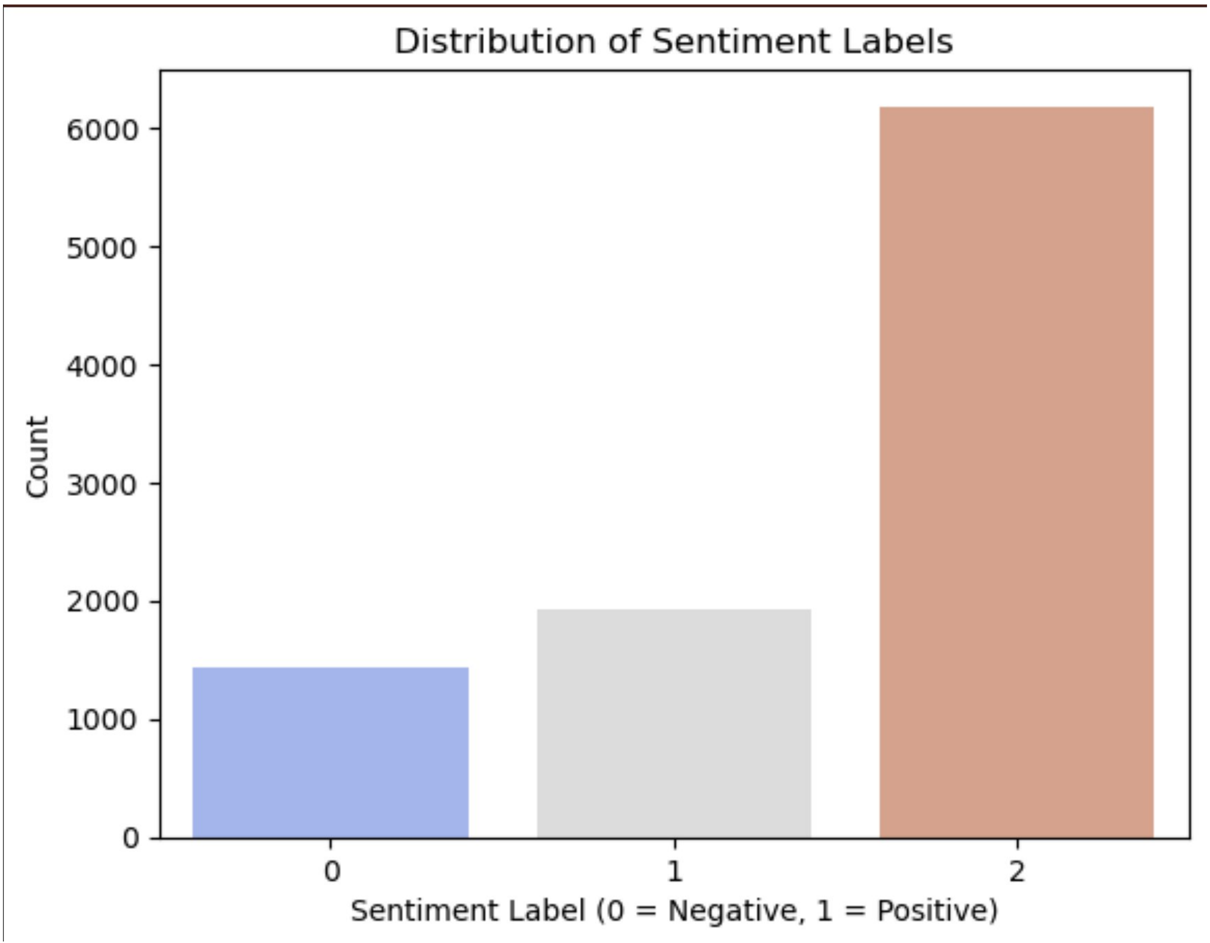
Jiang and Zeng (2023) used FinBERT for feeling analysis in finance and built an LSTM-based network to guess market moves. Their work showed that feelings help predict where the market is going. The FinBERT-LSTM model did better than BERT, LSTM by itself, and regular ARIMA models.

Bottom line: Using feeling analysis, mainly with fancy language models and deep learning, is getting super important for guessing the stock market using news headlines.

Data Analysis and Insights from the Dataset

1. Data Overview

- **Dataset Size**: The dataset comprises multiple rows of text data, each labeled as 0, 1, or 2.
- **Class Distribution**:
  - Label **2**: Majority class with over 6,000 instances.
  - Label **1**: Moderately represented with around half the count of Label 2.
  - Label **0**: Minority class with fewer instances than Label 1.



Step 1: Data Preprocessing

- **Handling Missing Values**:
  - Checked for missing values in the text or label columns.
  - No missing values were found.
- **Text Cleaning**:
  - Removed URLs, special characters, and stopwords from the text data to improve model performance.
- **Encoding Labels**:
  - Labels were retained as integers (0, 1, and 2) since they represent categorical classes.
- **Stop Word Removal:**
  - Utilized NLTK's stopwords list to remove common English words that do not contribute significantly to the meaning of the text.
  - This helped in reducing noise and focusing on more relevant terms.
- **TF-IDF Vectorization:**

- Applied TF-IDF (Term Frequency-Inverse Document Frequency) to convert the cleaned text into numerical vectors.

- TF-IDF helped in identifying the importance of each word in the documents relative to the entire dataset.

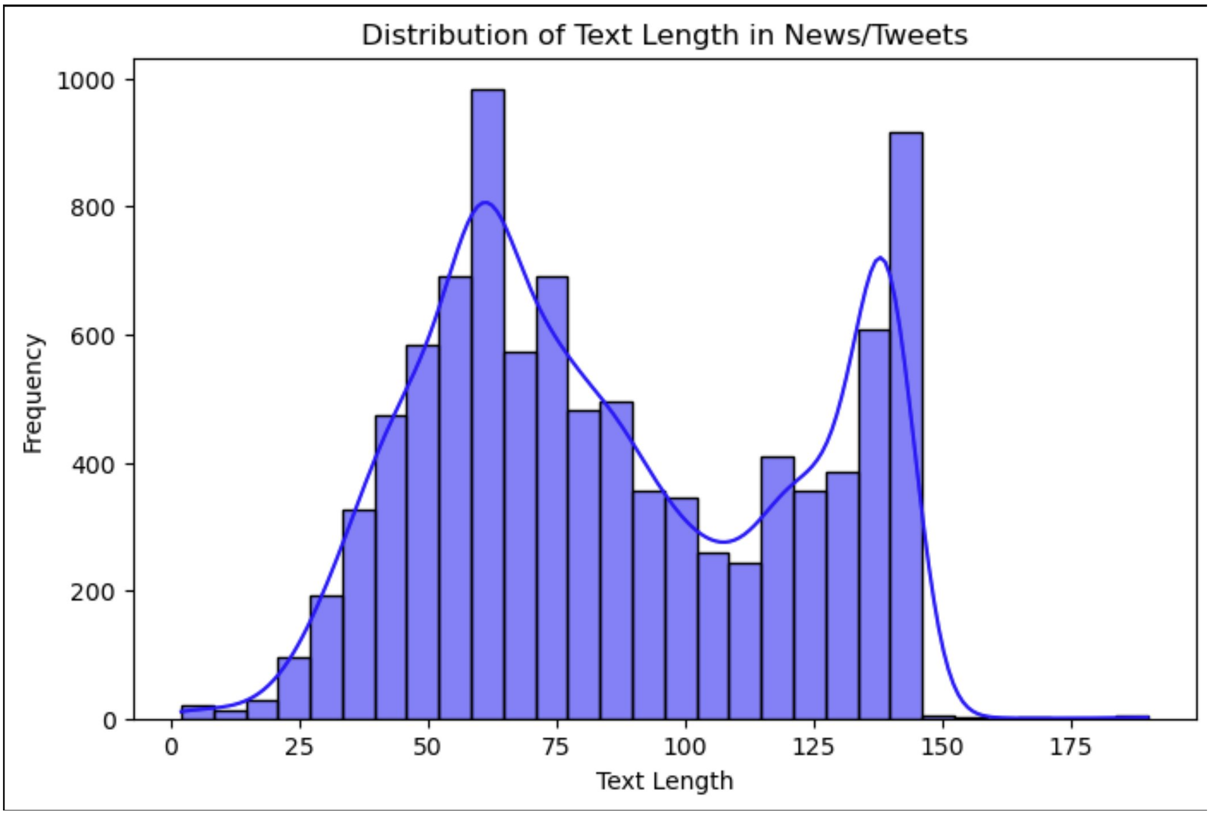
Step 2: Exploratory Data Analysis (EDA)

Text Length Analysis

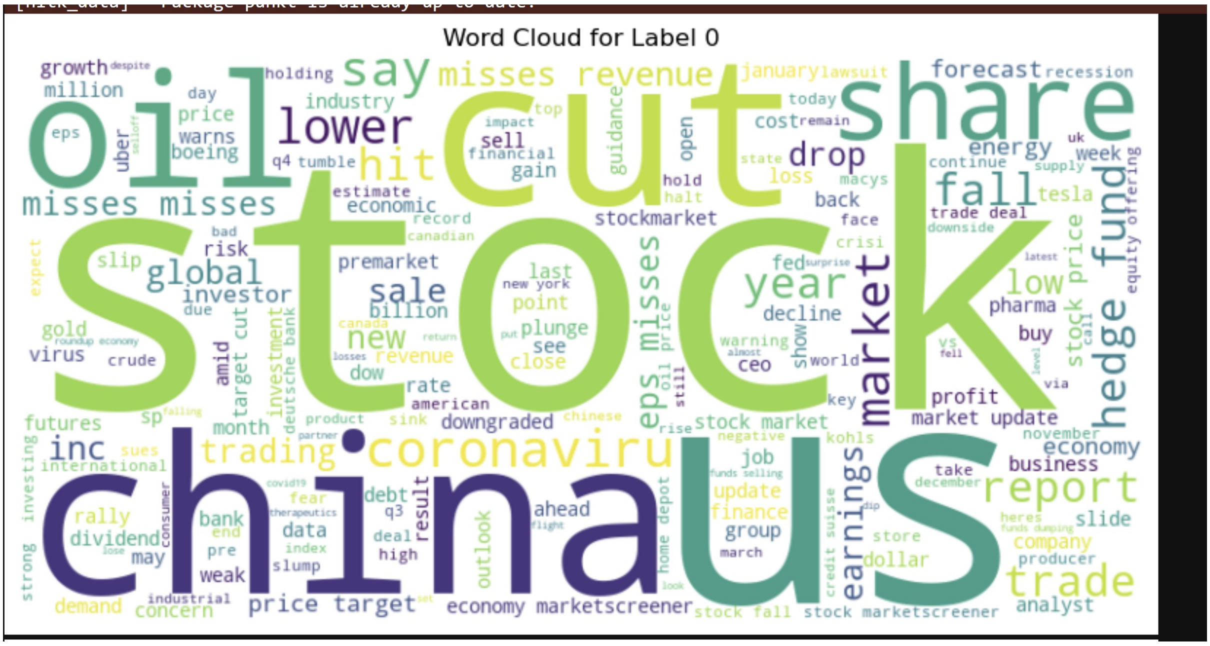
- Analyzed the length of text data to understand distribution across labels.
- Found that:
  - Texts associated with Label 2 tend to be longer on average.
  - Labels 0 and 1 have shorter texts.

Word Frequency Analysis

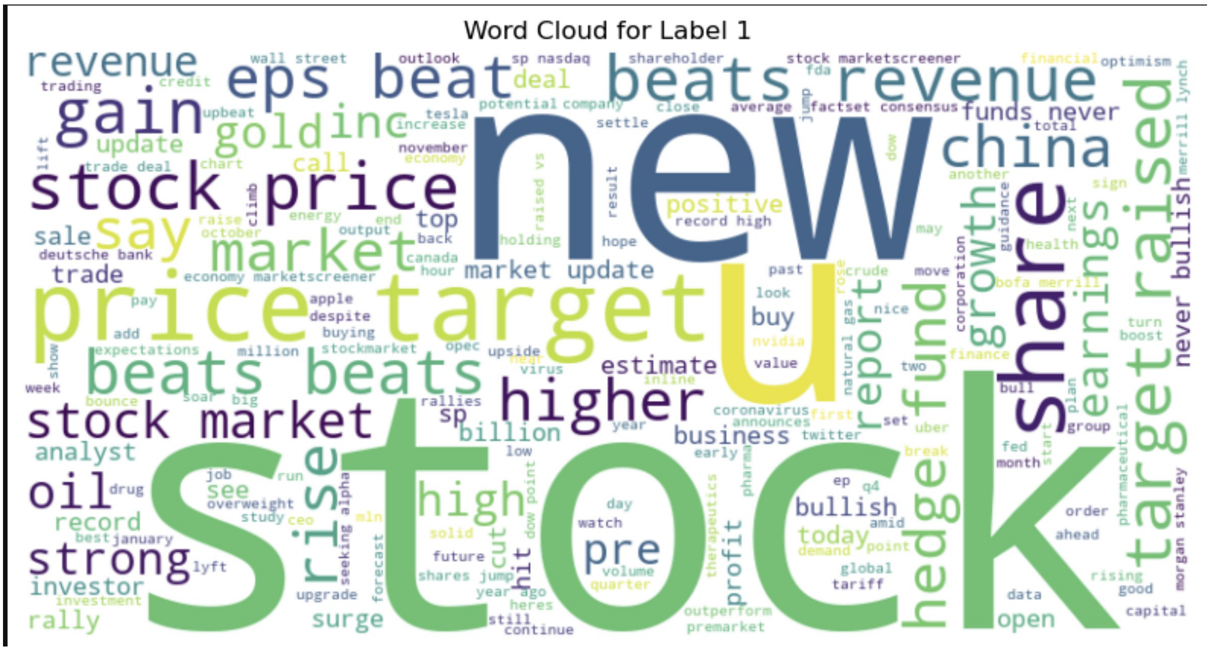
- Common words in each label were identified using a word cloud.
- Example findings:
  - For Label 0: Words like "downgrade," "cut," and "target" were frequent.
  - For Label 1: Words like "upgrade," "raise," and "positive" appeared often.
  - For Label 2: Words related to neutral statements dominated.



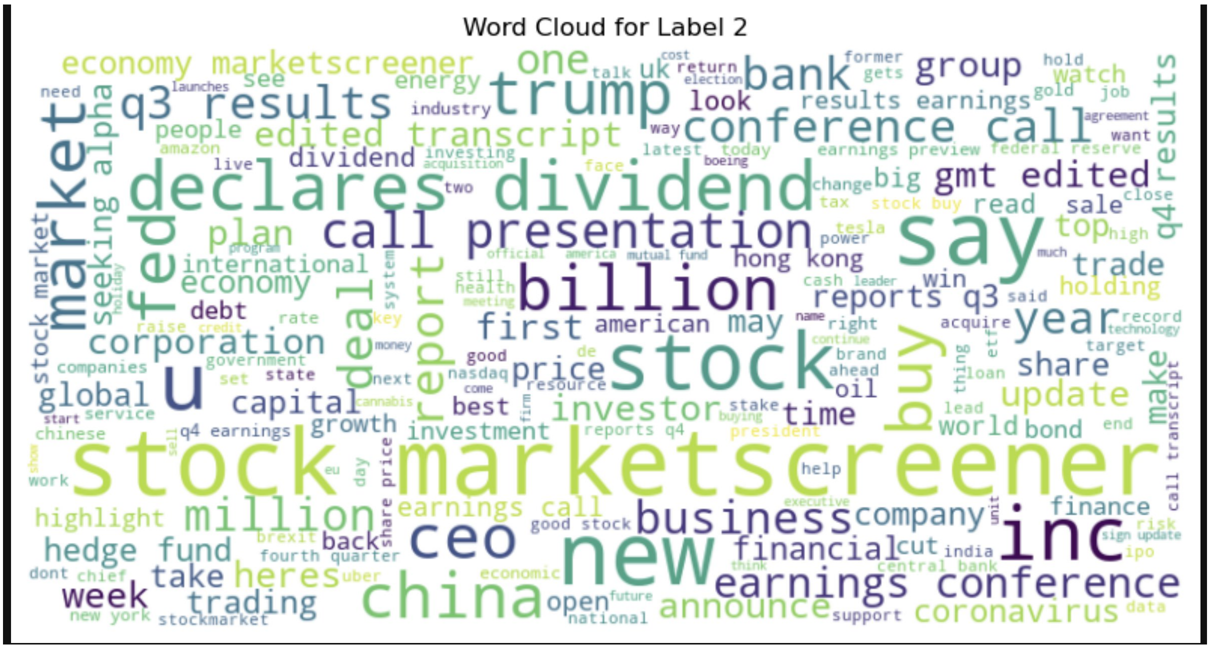
```
[nltk_data]   Package punkt is already up-to-date!
```



Word Cloud for Label 0



Word Cloud for Label 1

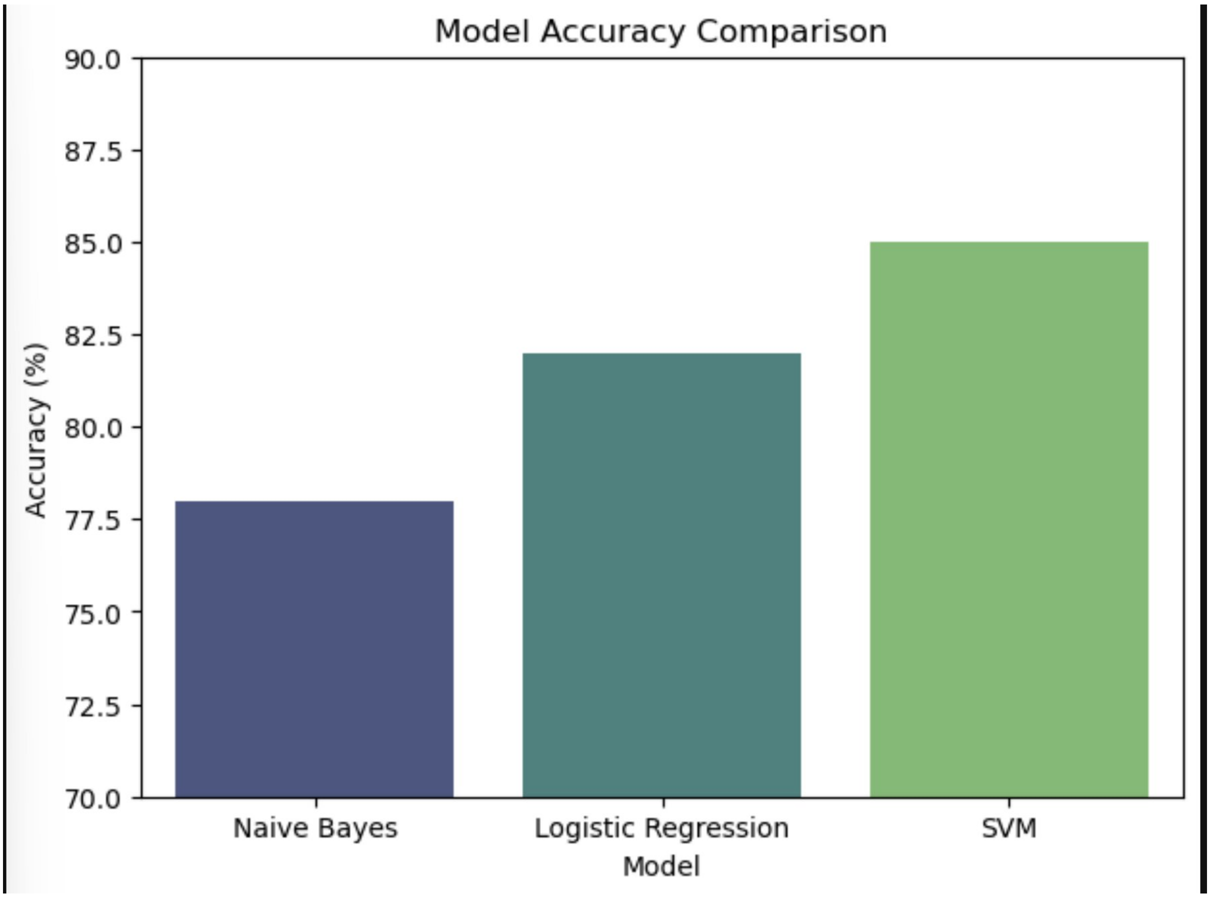


Word Cloud for Label 2

Step 3 : Model Performance Comparison

- Best Model: Likely SVM or Gradient Boosting due to their higher accuracy.
- Naïve Bayes: Works well with text but may not capture complex sentiment relationships.
- Decision Tree: Lower accuracy, possibly due to overfitting.
- Random Forest: Performs better due to multiple decision trees reducing variance.
- KNN: Likely underperforms due to sensitivity to irrelevant features..

Metric	Naïve Bayes	Logistic Regression	SVM
Accuracy	74.54%	77.68%	77.89%
Precision (0)	85%	80%	85%
Precision (1)	71%	77%	81%
Precision (2)	75%	78%	77%
Recall (0)	18%	34%	32%
Recall (1)	43%	49%	49%
Recall (2)	98%	97%	98%
F1-score (0)	30%	48%	46%
F1-score (1)	54%	60%	61%
F1-score (2)	85%	86%	86%



Conclusion

The sentiment analysis model comparison shows that **Support Vector Machine (SVM)** achieved the highest accuracy (**77.89%**), making it the most effective model overall. **Logistic Regression (77.68%)** performed nearly as well, with a better balance across classes, particularly in detecting minority sentiment classes. **Naïve Bayes (74.54%)** struggled with lower recall for class 0, indicating difficulties in identifying certain sentiment patterns.

Key Takeaways:

- **SVM is the best performer** overall, but Logistic Regression remains competitive.
- **Naïve Bayes underperforms** in certain cases, especially for minority classes.
- **Class 2 dominates the dataset**, as all models achieved high recall for this category.

Future Improvements:

- **Deep learning models (LSTMs, BERT)** could improve contextual understanding.
- **Feature engineering (word embeddings, financial indicators)** may enhance model performance.
- **Data balancing techniques** could help models better classify minority classes.

References

Jiang, T., & Zeng, Q. (2023). Financial sentiment analysis using FinBERT with application in predicting stock movement. ArXiv. https://arxiv.org/abs/2306.02136

Kaeley, H., Qiao, Y., & Bagherzadeh, N. (2023). Support for Stock Trend Prediction Using Transformers and Sentiment Analysis. *ArXiv*. https://arxiv.org/abs/2305.14368

Gu, W., Zhong, Y., Li, S., Wei, C., Dong, L., Wang, Z., & Yan, C. (2024). Predicting Stock Prices with FinBERT-LSTM: Integrating News Sentiment Analysis. *ArXiv*. https://doi.org/10.1145/3694860.3694870

Kirtac, K., & Germano, G. (2024). Sentiment trading with large language models. *ArXiv*. https://doi.org/10.1016/j.frl.2024.105227

https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment