

Group 6 Project Report

Ketan Kapse and Sai Abhishree Pappusetty

Introduction

Financial news data based on stock markets, companies' performance reports, and government policies, etc. are extremely important for various reasons. For investors, this news is crucial in assessing the performance and potential of stocks, bonds, or other investment vehicles. It provides insights into profitability, growth potential, and risks associated with investment options. It plays a key role for helping people in analyzing market trends and forecasting future market movements. It also provides valuable insights for traders, analysts, and financial institutions in making informed decisions. However, it requires the analysis of multiple news sources and headlines and it can take a person a while before they can reliably arrive at the conclusion for the overall sentiment of a particular event, such as whether a company is expected to foresee a downturn in sales because there will be multiple conflicting news from different sources.

The aim of our project is to analyze financial news and reliably predict the sentiment that they carry. This can help investors to quickly gauge the overall sentiment of an event they are interested in and can help them make informed choices.

For this project, we have employed a mixture of classical ML and transformer based models. The classical models include Logistic Regression, Multinomial Naive Bayes and Gradient Boosted Classifier. The transformer models that we used for this project are pretrained Distil-Bert (distilbert-base-uncased) and roBERTa (roberta-large) [1],[2] obtained from HuggingFace which we have meticulously finetuned into classifying the sentiments of news data into three labels: positive, neutral and negative. We have performed a comparative analysis of all these models at the end of this report.

Description of the Dataset:

Our dataset is a combination of two datasets namely the financial_phrasebank dataset and the FiQA dataset obtained from HuggingFace. It is a collection of sentences extracted from English language financial news articles, organized and categorized based on sentiment polarity by human annotators.

It consists of approximately 12000 samples of text and their corresponding sentiments.

Description of the NLP Models:

We employed three classical models, namely Logistic Regression, MultinomialNB and Gradient Boosting Classifier. Logistic Regression models naturally establish linear decision boundaries. In sentiment analysis, this can be advantageous when the relationship between features and sentiment is somewhat linear or can be approximated as such. It separates classes effectively in high-dimensional spaces.

The Multinomial Naive Bayes (MultinomialNB) classifier stands out for its simplicity and efficiency in classification tasks [3]. Operating on the principles of Bayes' theorem, this probabilistic algorithm assumes independence among features, making it particularly adept at handling text data. Its effectiveness lies in its ability to work well even with high-dimensional feature spaces by considering the occurrences of words or features while assuming they are independent of each other.

Sentiment analysis might involve complex nonlinear relationships between words and sentiments. GBC, with its ability to create strong predictive models by combining weak ones, can capture such intricate patterns in the data, which might be challenging for linear models like Logistic Regression or Naive Bayes, hence it acts as a comparative

model for the other two models. For our training, we had the number of weak estimators as 300, and the max depth of the model was set to 10.

We have developed two BERT based classifiers, namely DistilBert and roBERTa for detecting the sentiment in the text that we have. The reasoning behind using these BERT models is because they are pre-trained on large amounts of text data and can capture complex contextual information within sentences. In financial news, where context plays a crucial role in determining sentiment, their bidirectional nature helps in understanding the nuanced meaning of words and phrases. By fine-tuning on this dataset, the models can learn to comprehend financial jargon, industry-specific sentiments, and economic nuances, leading to more accurate sentiment analysis.

DistilBert after training on our dataset achieved a F1 score of 0.9602 and roBERTa achieved 0.94.

Experimental Setup

Our experimental setup involves the utilization of the approximately 12000 samples in our dataset, stored in a pandas dataframe, and their corresponding sentiments to train the models for sentiment analysis.

We then preprocess the data to improve the models' performances by performing the following operations on the textual data:

1. Lowercasing:
 - Conversion of all text to lowercase. This standardizes the text by treating words with different cases as identical.
2. Tokenization:
 - Segmentation of the text into smaller units, such as words or sentences.
3. Removing Punctuation:

- Elimination of punctuation marks from the text.
- 4. Stopword Removal:
 - Extraction of common stopwords, such as from the text.
- 5. Lemmatization:
 - Reduction of words to their base or root form.
- 6. Normalization:
 - Additional transformations to standardize text, such as handling contractions (e.g. converting "can't" to "cannot").

This preprocessed textual data is then tokenized and fed into our models. The methods used for tokenization differ from model to model.

For the classical ML models, we are splitting the dataset using `train_test_split` from `sklearn` into train and test sets with a 80:20 split. We then vectorize the data using TF-IDF or count vectorizer and fit it onto the models.

For the BERT models, we chose to utilize stratified k-fold cross validation which divides the training data into k-folds and then the models train on k-1 folds and then validate on the last fold. The training data is tokenized with their respective tokenizers from HuggingFace. Each sample's tokens are mapped to their respective indices in the tokenizer vocabulary and special tokens like [CLS], [SEP] are added. They are also padded in order to ensure uniform length.

Since this is a multi-class classification problem, we chose to utilize the CrossEntropy loss from `pytorch`. This loss measures the deviance of the predicted sentiment from the actual sentiment and then penalizes the model accordingly. The optimizer function that we used was Adam with weight decay. The learning rate was set to $2e-5$ with betas = (0.9, 0.999) and epsilon $1e-5$ for both the models. The batch size we used was 16 for both the models. The number of folds for the cross validation was 3, with 3 training epochs per fold.

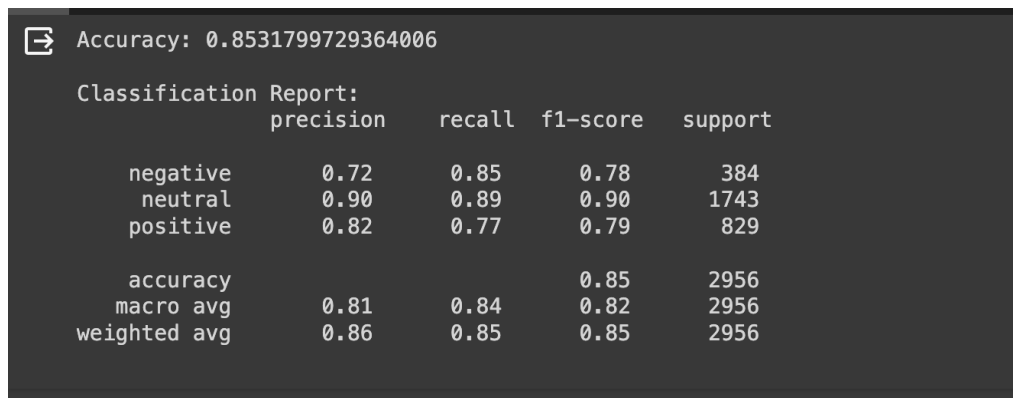
All of the models are saved as either pickle files or as safetensors.

Finally, we utilized streamlit to apply our saved models and generate a UI within which a user can select the models to use and input the text that they want to generate sentiments for in a text field that is provided. The output is the predicted sentiment, i.e. positive, negative or neutral for the given text, along with the calculated probabilities for each sentiment.

Results and Observations:

Classical Models:

1) MultinomialNB:



```
➞ Accuracy: 0.8531799729364006

Classification Report:
      precision    recall  f1-score   support

negative      0.72      0.85      0.78        384
neutral      0.90      0.89      0.90       1743
positive      0.82      0.77      0.79        829

accuracy      0.85      0.85      0.85       2956
macro avg      0.81      0.84      0.82       2956
weighted avg   0.86      0.85      0.85       2956
```

Figure 1. Classification Report of MultinomialNB

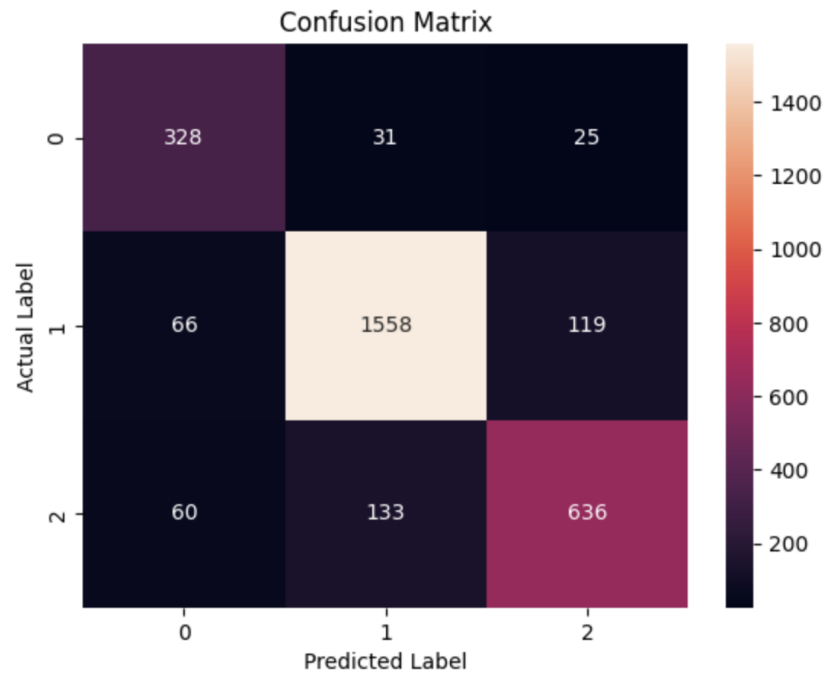


Figure 2. Confusion Matrix for Multinomial NB

The accuracy on the test set was 0.8531.

2) Logistic Regression:

```

Accuracy: 0.9465493910690121

Classification Report:

```

	precision	recall	f1-score	support
negative	0.92	0.88	0.90	384
neutral	0.95	0.98	0.97	1743
positive	0.95	0.90	0.92	829
accuracy			0.95	2956
macro avg	0.94	0.92	0.93	2956
weighted avg	0.95	0.95	0.95	2956

Figure 3. Classification report for log reg.

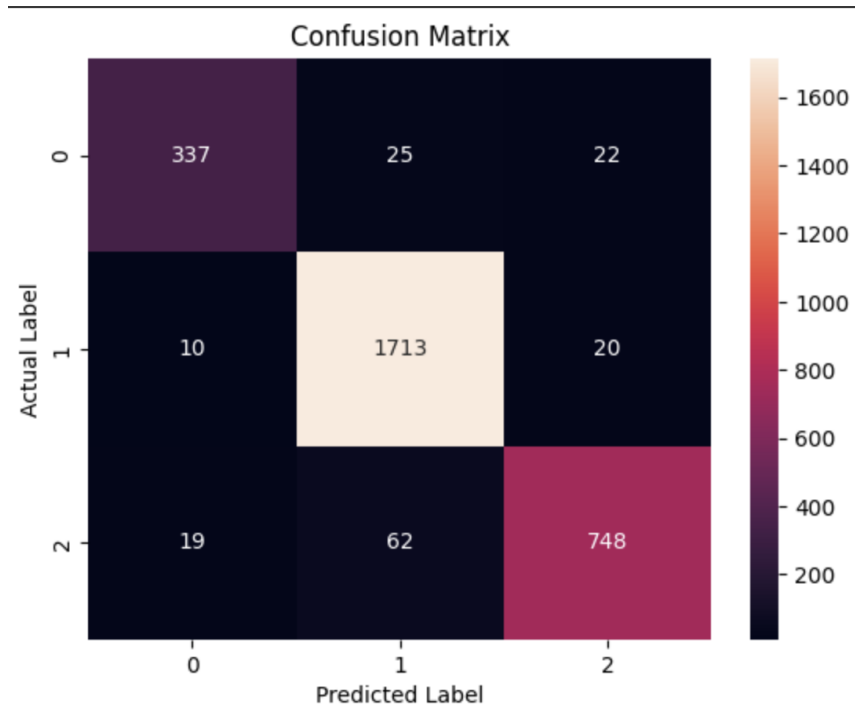


Figure 4. Confusion Matrix for log reg.

The accuracy observed on the test set is 0.9465.

3) GBC:

Classification Report:				
	precision	recall	f1-score	support
0	0.96	0.86	0.91	243
1	0.94	0.98	0.96	1223
2	0.94	0.89	0.91	520
accuracy			0.94	1986
macro avg	0.95	0.91	0.93	1986
weighted avg	0.94	0.94	0.94	1986

Figure 5. Classification report for GBC

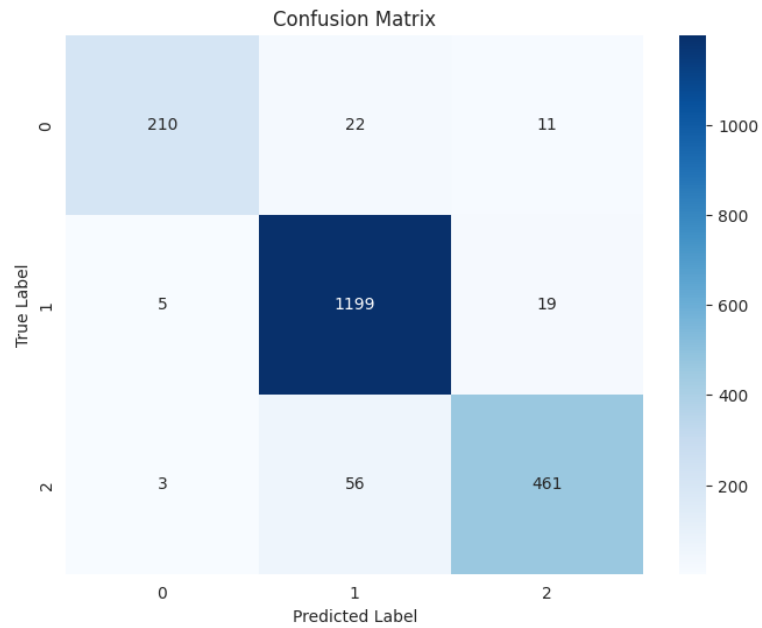


Figure 6. Confusion Matrix for GBC

The observed test set accuracy was 0.94.

Transformer Models:

4) DistilBert:

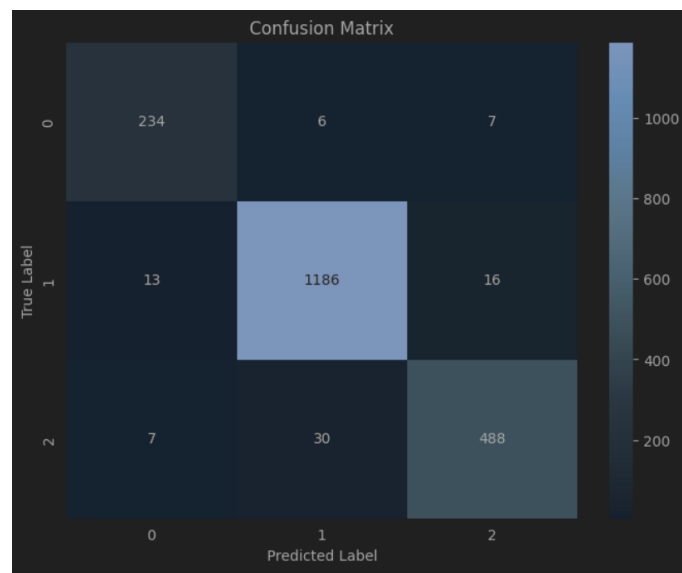


Figure 7. Confusion Matrix for DistilBert

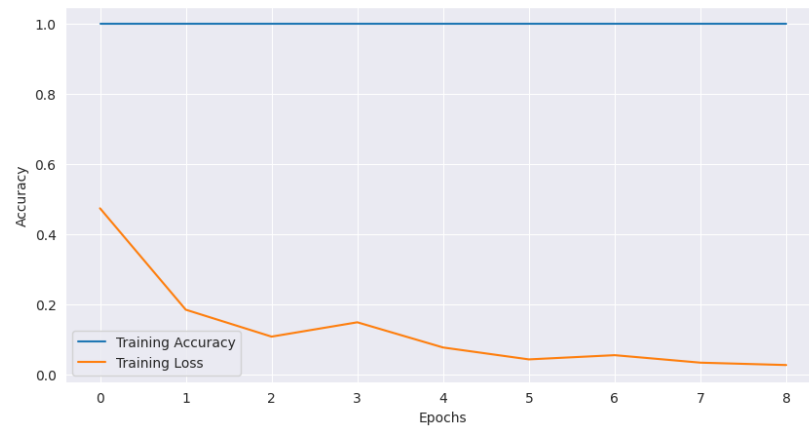


Figure 8. Training Accuracy and Training Loss

Classification Report for Holdout Set:				
	precision	recall	f1-score	support
0	0.92	0.95	0.93	247
1	0.97	0.98	0.97	1215
2	0.95	0.93	0.94	525
accuracy			0.96	1987
macro avg	0.95	0.95	0.95	1987
weighted avg	0.96	0.96	0.96	1987

Figure 9. Classification report for distilbert on the holdout.

The model has a F1 score of 0.9602 on the holdout set after training.

5) roBERTa:

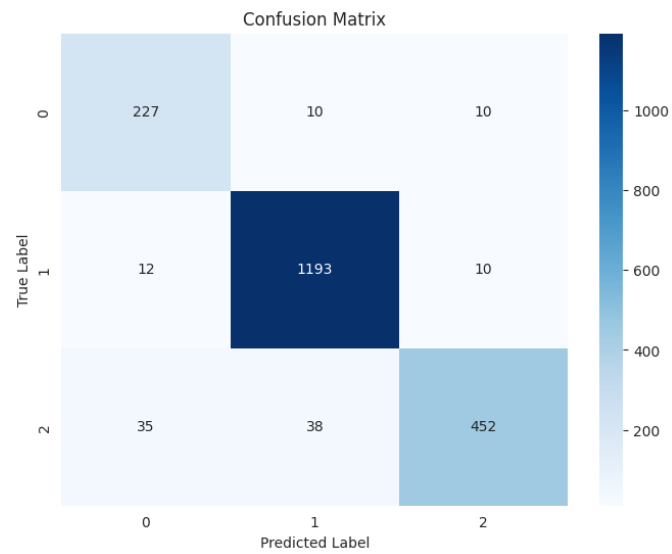


Figure 10. Confusion Matrix

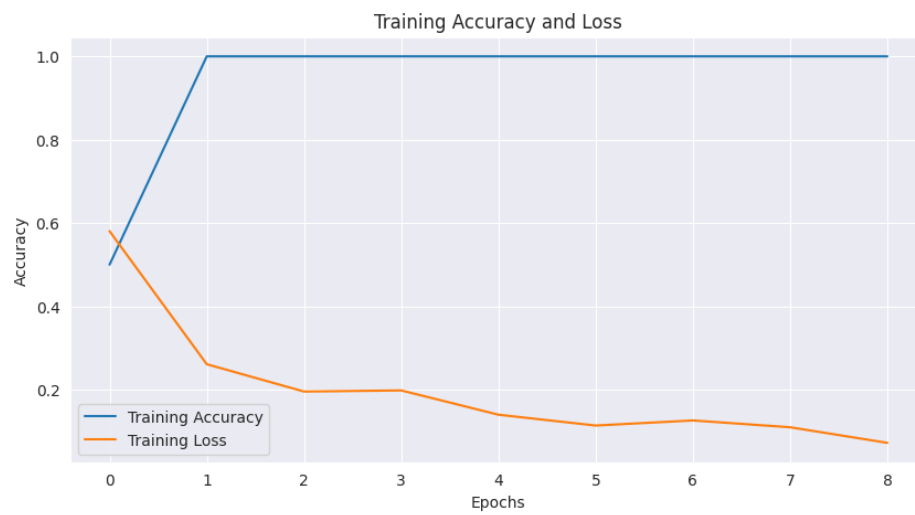


Figure 11. Training accuracy and loss for roberta

Classification Report for Holdout Set:					
	precision	recall	f1-score	support	
0	0.83	0.92	0.87	247	
1	0.96	0.98	0.97	1215	
2	0.96	0.86	0.91	525	
accuracy			0.94	1987	
macro avg	0.92	0.92	0.92	1987	
weighted avg	0.94	0.94	0.94	1987	

Figure 12. Classification report for roberta on the holdout set.

STREAMLIT UI

Financial Sentiment Prediction

Select a Model

Distilbert-Base

DistilBert

Enter text for prediction:

Warren Buffett's Berkshire Hathaway continues to sell HP shares, reducing stake to 5.2%

Predict

Predicted Label: The Predicted Sentiment is Neutral. (ID = 1) Label Probabilites: [0.006544056348502636, 0.9892507791519165, 0.004205229692161083]

Figure 13. Streamlit



Figure 14. Selecting Models

As can be seen in the images above, the UI provides a drop down list of all the models that we have available and allows the user to select from them and make predictions on the text that they input. The predicted label along with the probabilities for all the other labels are also presented to the user.

Summary and Conclusions

In this project, we trained a variety of models for sentiment analysis - Logistic Regression, Multinomial NB, GBC, DistilBert and roBERTA on the financial sentiment dataset we obtained from HuggingFace. We evaluated our models on their performance using F1 scores. Furthermore, these models were then deployed using a streamlit application that allows users to input their own text and get the sentiments using the trained models.

References

[1] Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *arXiv preprint arXiv:1910.01108* (2019).

[2] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019).

[3]<https://medium.com/@evertongomede/understanding-multinomial-naive-bayes-classifier-fdbd41b405bf>