NLP + Finance Algorithmic Trading

By Ryerson, Aviv, Brody, Hari, Dev and Diego

Introduction

NLP + Finance

Predicting stock market prices using sentiment analysis on Twitter.

This project focuses on the intersection of AI and finance, predicting stock market prices using sentiment analysis on Twitter.

By analyzing sentiment associated to tweets the project gives traders an extra edge by indicating stock price indications.

How did we solve?

By utilizing models and different NLP techniques that preprocess data and accurately/efficiently predict fluctuations, we were able to solve the problem in multiple different ways.

- Tokenization
- Stop word removal
- Lemmatization
- Bag of words
- Padding
- Attention Mask
- Sentiment Analysis
- Logistic Regression
- Long Short-Term Memory(LSTM)
- BERT Model

Text Filtering

Tokenization: Separates each word and punctuation in a string of text.

I woke up this morning at noon
>>I, Woke, Up, This, Morning, At, noon.

Lemmatization: Turning each word into its simplest form. Running>Run
Stop word Removal: Removing words that does not help in terms of getting a sentiment. Such as pronouns and connections words.
After those processes, the input(text) would be simple and ready for it to be translated into numbers.

Text to Numbers

Bag of Words

A method to transfer texts into numerical data by recording the frequency of each word.

Padding

Adding a set of numbers to data, making it able to be processed. Does it have any meaning itself.

Attention Mask

Tells the computer to ignore (not to process) the paddings.

Models Used

Sentiment Analysis and Logistic Regression

Sentiment Analysis

Sentiment analysis trains a model to classify news from various sources as neutral, negative, or positive.

- Purpose: Helps predict stock price fluctuations based on news sentiment:
  - Positive news: Can lead to increased investment and price rises
  - Negative News: Can lead to decreased investment and price drops.
- Techniques Involved:
  - Tokenization:
  - Logistic Regression:
- Used in LSTM, BERT

Logistic Regression

Logistic regression is a type of model used to predict the probability of a particular outcome

- Also used in sentiment analysis by classifying news.
- Function:
  - Analyzes relationships between input variables and binary/categorical variables
- Purpose: Helps in making informed decisions or predictions based on historical data patterns like stocks

Models Used

LSTM and BERT

LSTM

LSTM Models use Bag of Words and Sentiment Analysis to predict stock price changes.

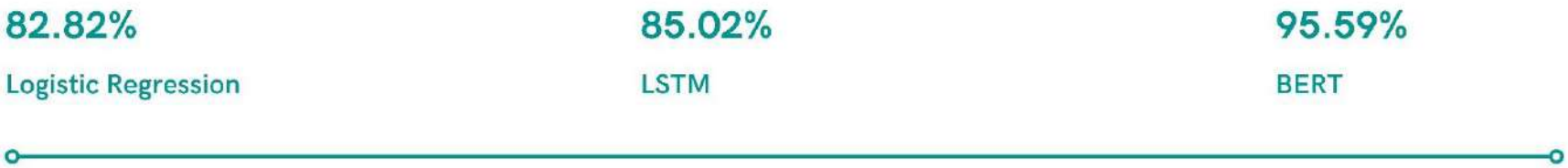
- Financial Sentiment Analysis:
  - Data is preprocessed through padding and one-hot encoding.
  - Model is trained and evaluated on this processed data
- Hyperparameter Optimization:
  - Improves model performance
- Purpose:
  - Can accurately classify sentiments and is used in finance to properly predict stock fluctuations.

BERT

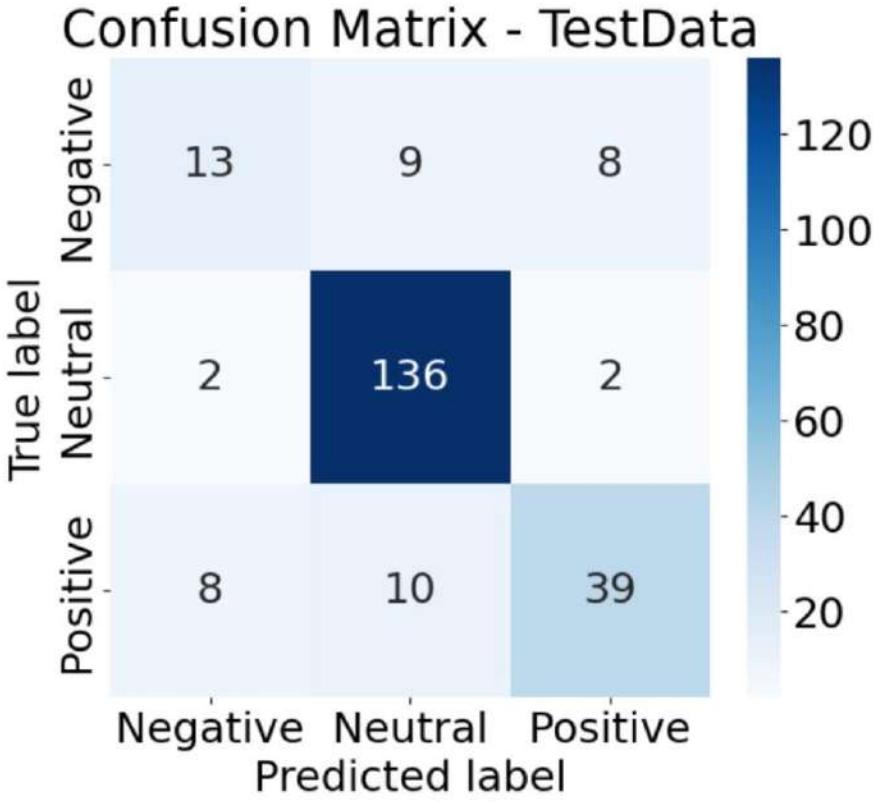
An advanced NLP model that excels in understanding the context of words within sentences.

- Tuned for Sentiment Analysis and Classification:
  - BERT Models can process and analyze data quickly and effectively.
- Techniques Used:
  - Tokenization
  - Padding
  - Other NLP Techniques
- Benefits:
  - Advanced NLP model that can quickly, accurately, and efficiently predict changes.

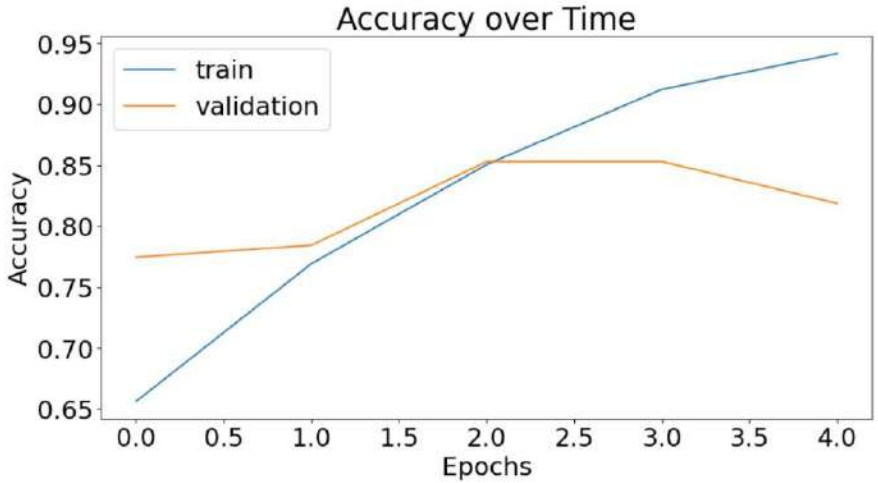
Different Levels of Accuracy

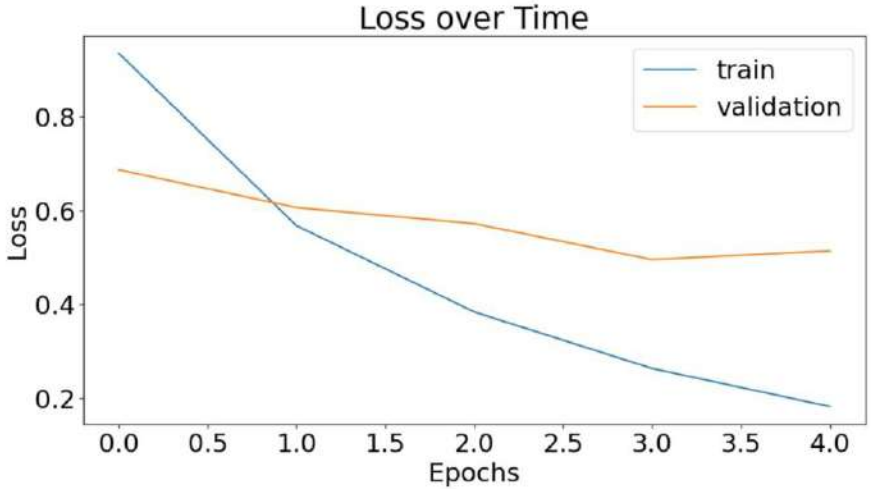


Logistic Regression



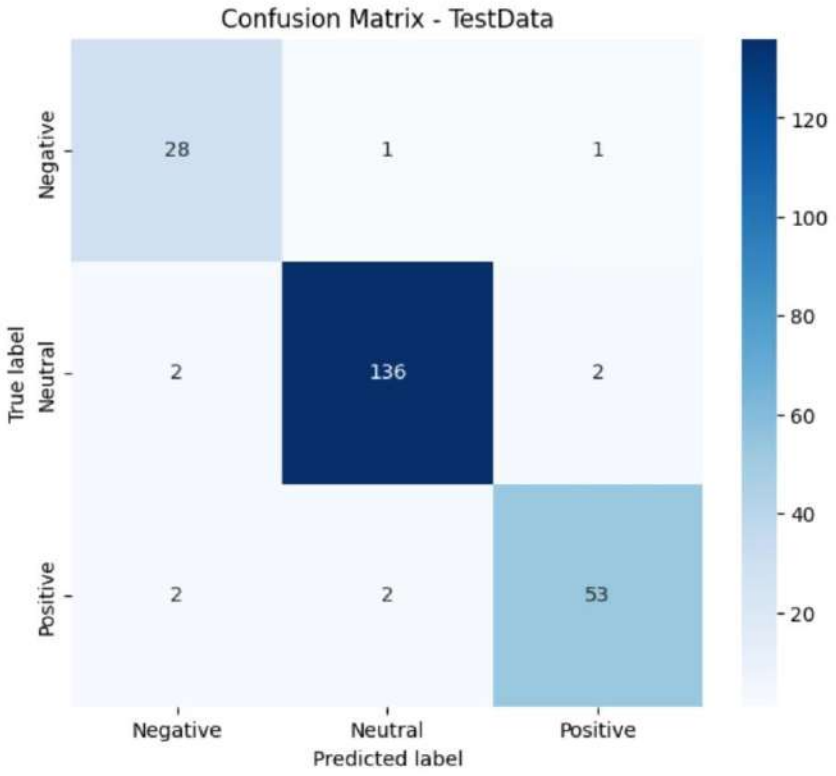
LSTM Graphs

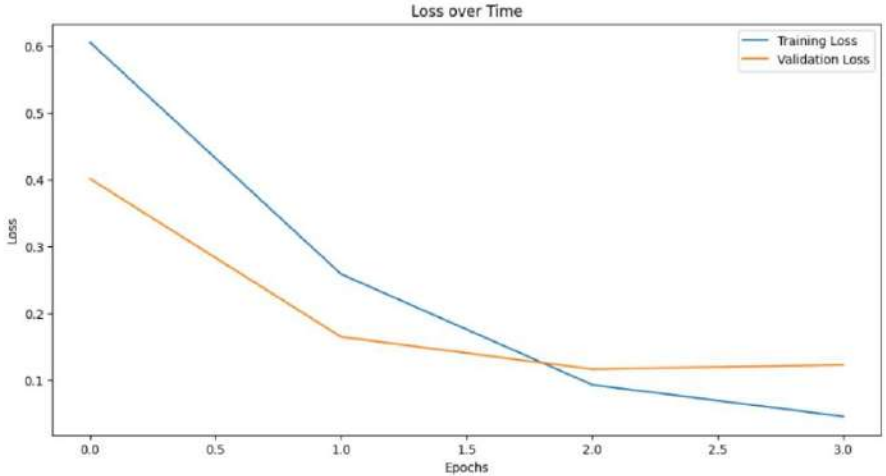

Accuracy over Time
train
validation
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Epochs


Loss over Time
train
validation
Loss
0.8
0.6
0.4
0.2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Epochs

BERT

Graphs


Confusion Matrix - TestData
True label
Negative
Neutral
Positive
28
1
1
2
136
2
2
2
53
120
100
80
60
40
20
Negative
Neutral
Positive
Predicted label


Loss over Time
Training Loss
Validation Loss
Loss
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Epochs

Opinion

Overall

We believe for the purpose of our project, the BERT model is the best. This is evident from its accuracy when compared to the other models that we tried such as, the LSTM model and the linear regression model. As it is 11% and 14% better than the linear regression and LSTM models respectively.

Improvement

In Performance

Training

Data

More data means more examples for our model to learn from. By collecting a larger dataset of financial news articles and their sentiment labels, we can help our model recognize patterns more accurately.

Model

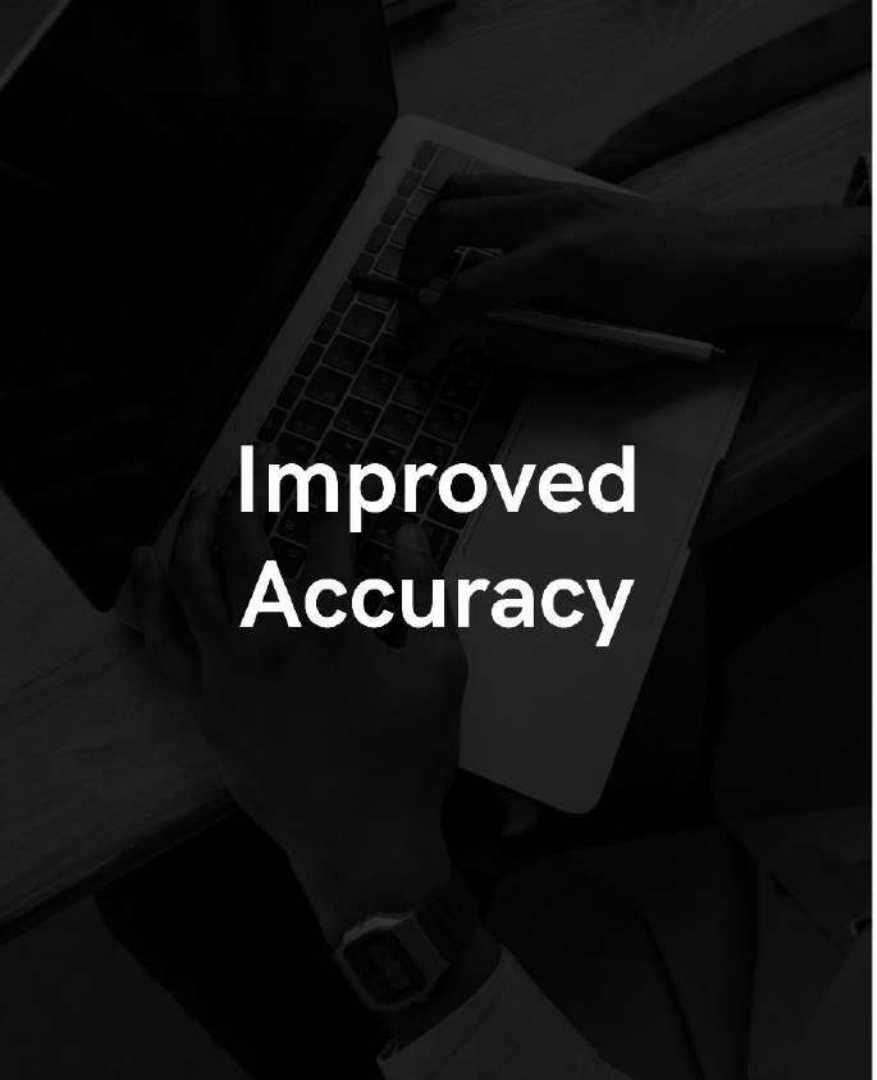
Architectures

Just like there are many ways to solve a math problem, there are different types of neural networks we can use. Experimenting with various architectures or combining multiple models can lead to better overall performance.

Hyper

Parameters

Hyperparameters are like the settings of our model that we can adjust. By tweaking these values and using techniques like cross-validation, we can find the optimal configuration for our specific task.



Improved Accuracy

Clean and preprocess data

The quality of our data directly impacts the quality of our results. By removing irrelevant information and fixing errors in our text data, we can help our model focus on what's important.

Pre-trained Word Embeddings

Words have complex meanings and relationships that can be hard for a model to learn from scratch. Using pre-trained word embeddings gives our model a head start by leveraging existing knowledge about language.

Analyze Mistakes

Understanding where our model goes wrong is key to making it better. By examining the examples it misclassifies, we can gain insights into its weaknesses and how to address them.

Conclusion

In conclusion, this week through NLP in finance has taken us from basic models to advanced techniques like BERT. We have evaluated how to preprocess financial news data, build sentiment analysis models, and evaluate their performance.

In every single technique, from simple logistic regression to complex neural networks, we've seen how different approaches can be more or less accurate, and work more or less efficiently. While our BERT model showed promising results, there's always room for improvement.

Hence, the exploration in the future will include expanding our dataset, and fine-tuning our models. Through this future exploration, the aim is to achieve maximum accuracy to predict market movements and invest accordingly.

Thank you!

We appreciate your attention and time.

Special thank you to all instructors for their time, dedication and efforts for us this week.

INSPIRIT AI