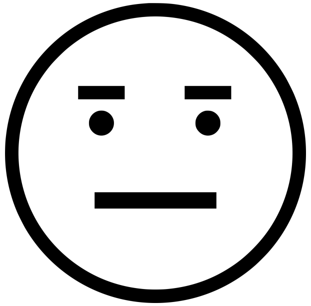
Stock News headlines Sentiment Analysis

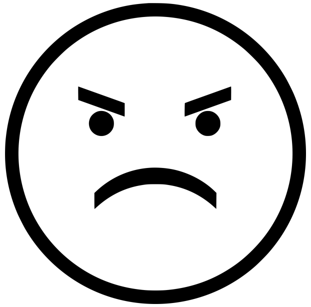
Erika Atoma (CSC-680)

Data Selected

- Collected from EODHD APIs (Stock Fundamental Data API).
- The Datasets were selected from 2021 to 2022 with 5086 entries and 8 features.
- The stocks used were ATVI, CVS, NDAQ, NFLX, NVDA, PLD, SBUX, SCHW, AAPL, TSLA.

Tasks Addressed

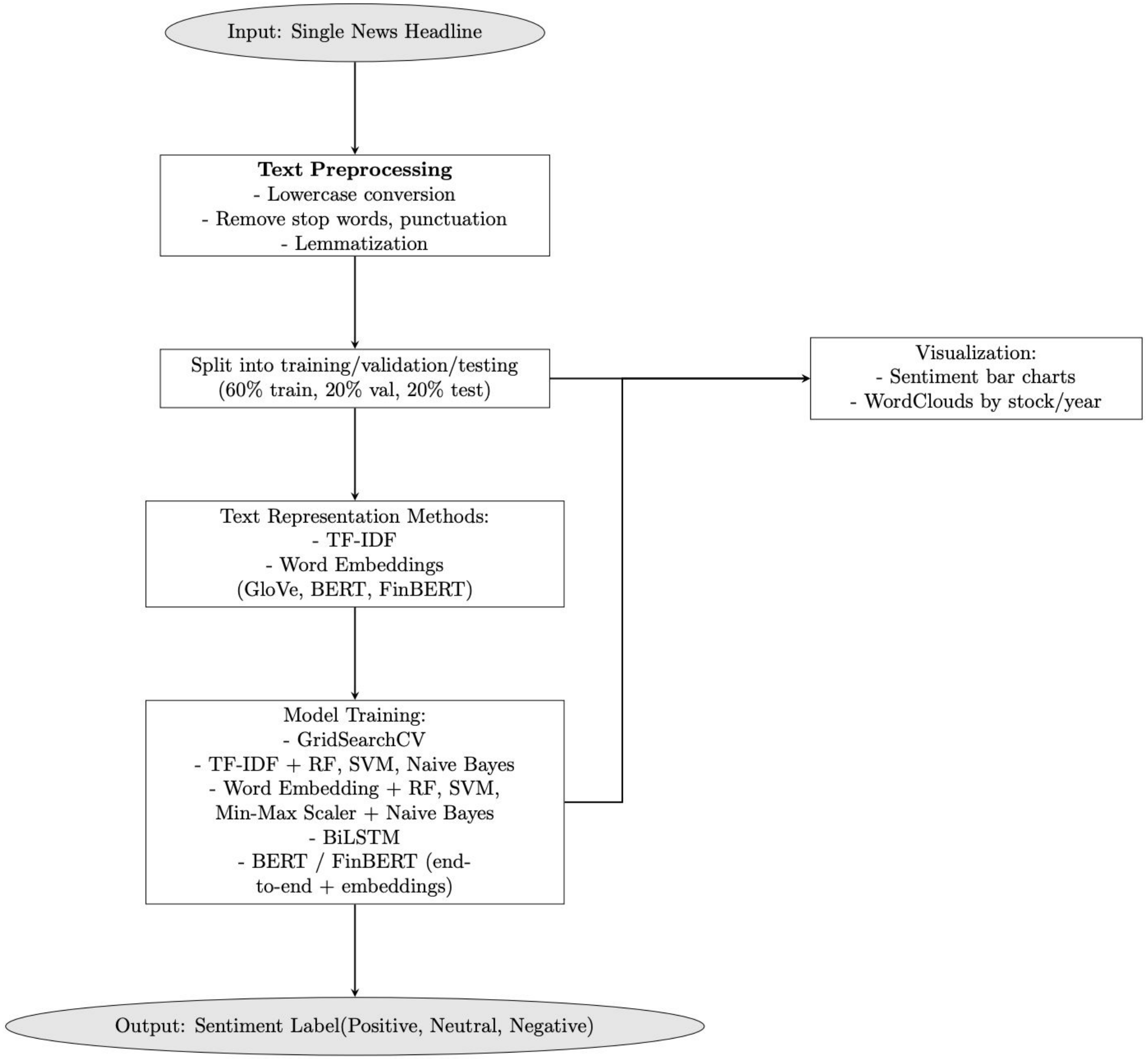




	date	symbol	close	rsi	26-Day EMA	NextDayClose	Diff	Titles	filename	Cleaned_titles
0	2021-01-05	aapl	131.009995	57.697283	3.371608	126.599998	0	Two former Apple insiders are building a laser...	headlinesIter0.csv	two former apple insider building laser techno...
2	2021-01-07	aapl	130.919998	55.965116	3.089118	132.050003	1	Dow Jones Futures Rise: Congress Certifies Bid...	headlinesIter2.csv	dow jones future rise congress certifies biden...
3	2021-01-08	aapl	132.050003	57.633668	2.969111	128.979996	0	The Best Mutual Funds Bet These Tech Stocks, I...	headlinesIter3.csv	best mutual fund bet tech stock ipo keep running
4	2021-01-11	aapl	128.979996	51.881886	2.812149	128.800003	0	Short Selling Legend Jim Chanos' Top 10 Stock ...	headlinesIter4.csv	short selling legend jim chanos top 10 stock p...
5	2021-01-12	aapl	128.800003	51.557013	2.630971	130.889999	1	The World's Top Maker of Mini Motors Bets It C...	headlinesIter5.csv	world top maker mini motor bet win tesla

Based on Models, I conduct the **sentiment analysis** for the cleaned single headlines in the dataset.

Methodology Approach



Tools Used

pandas

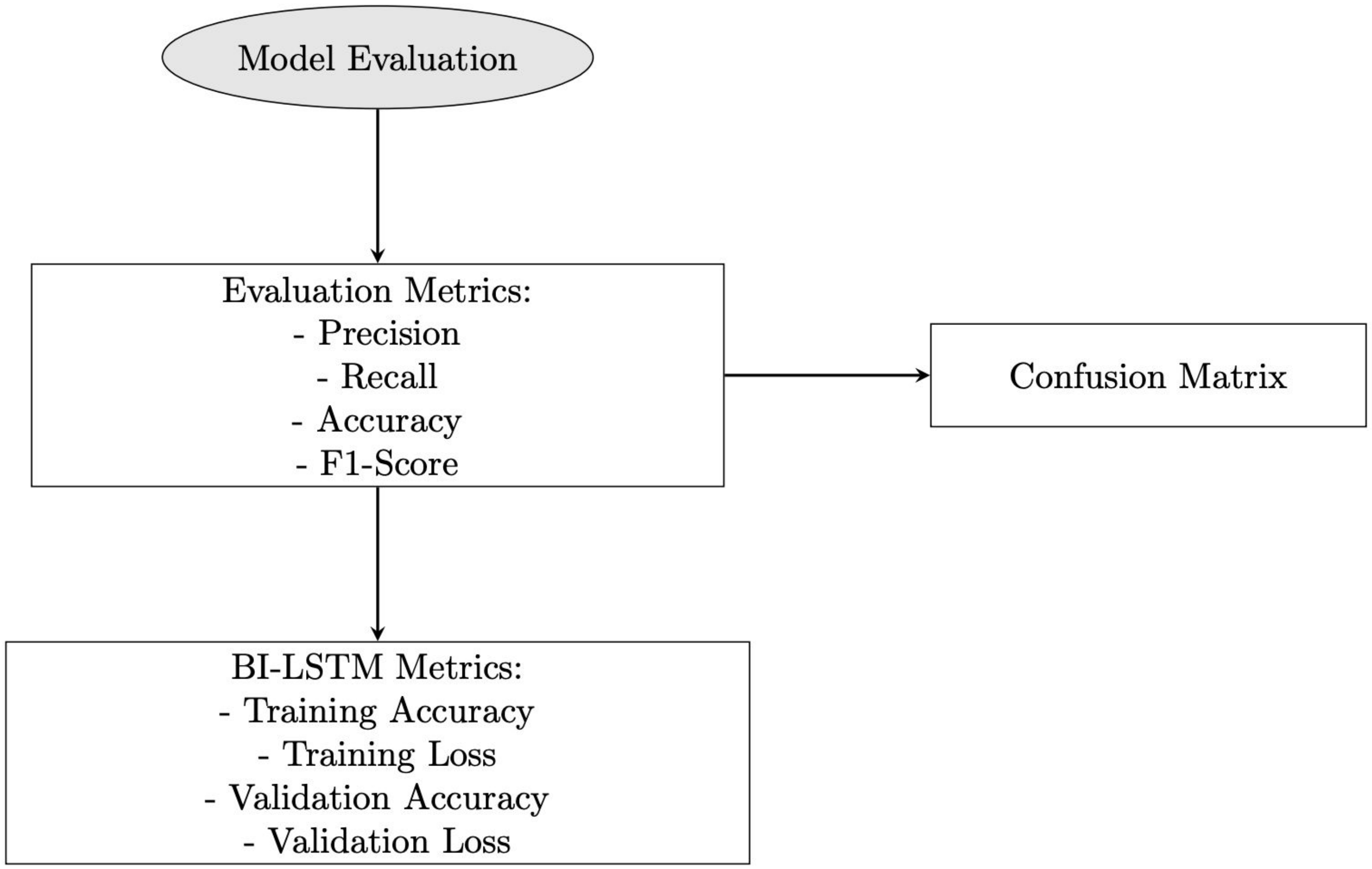
NumPy

scikit learn

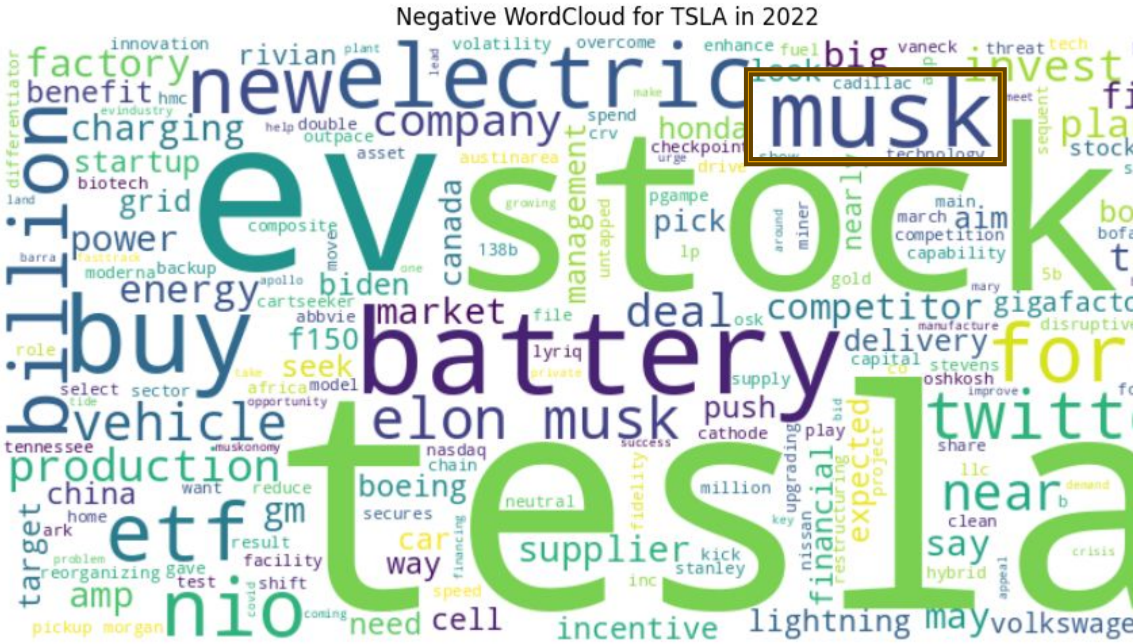
matplotlib

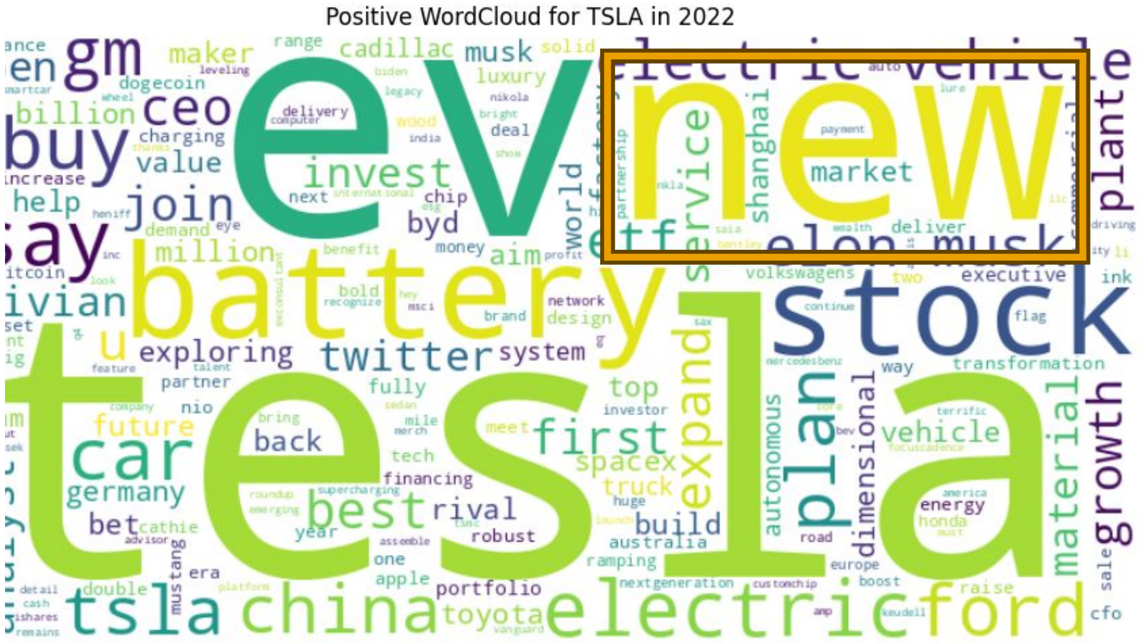
TensorFlow

Metrics Adopted

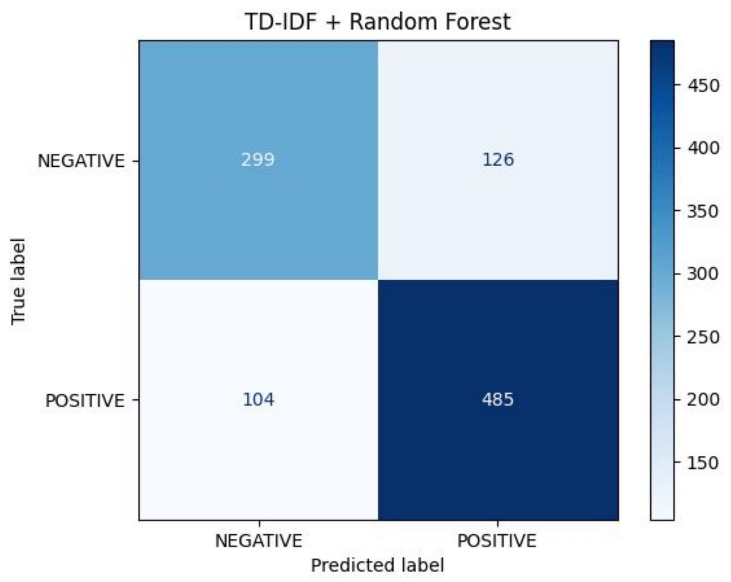


Experimental results: Word Cloud

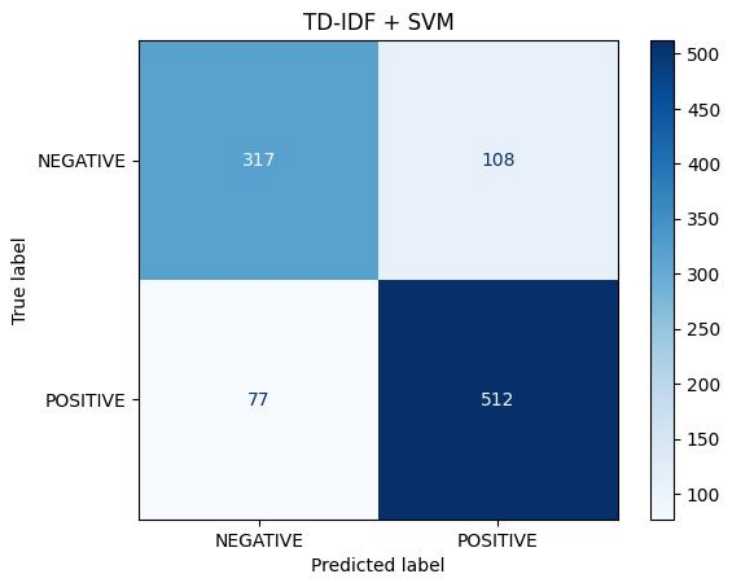




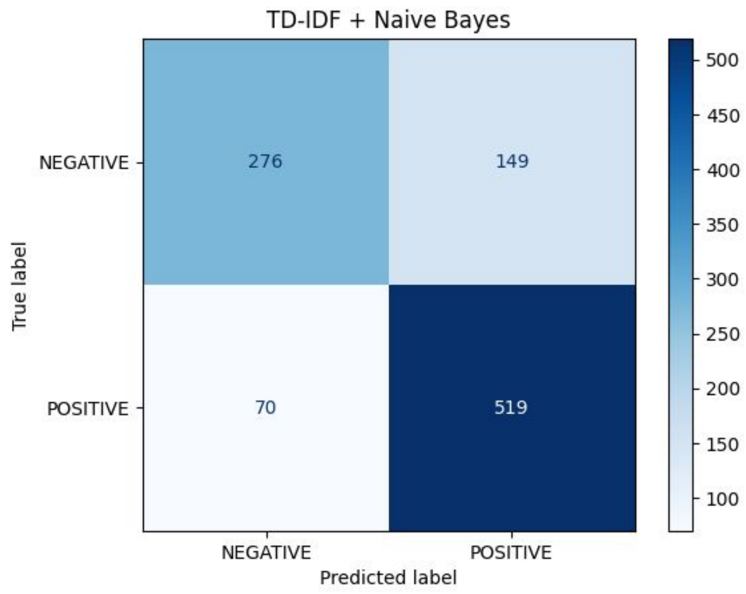
Experimental results: TD-IDF



Best Parameters: {'bootstrap': False, 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}
Test Accuracy: 77.317%



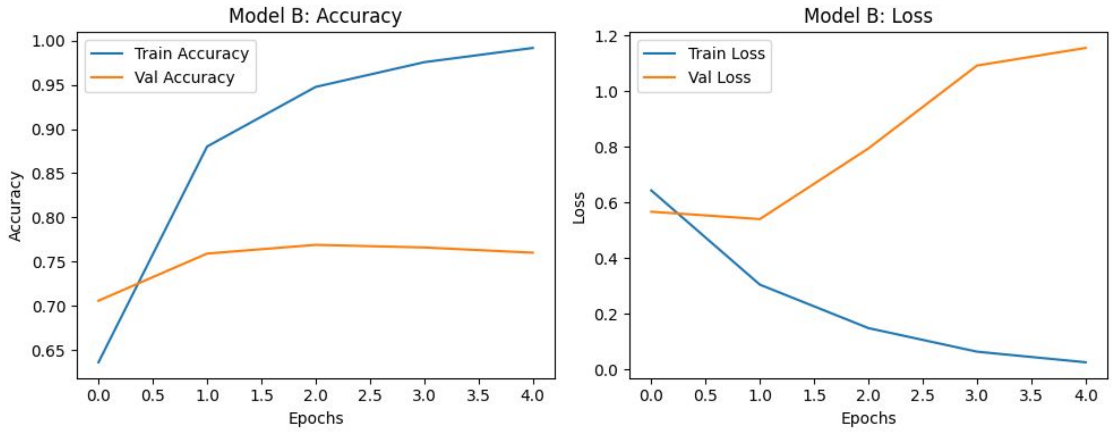
Best Parameters: {'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}
Test Accuracy: 81.755%



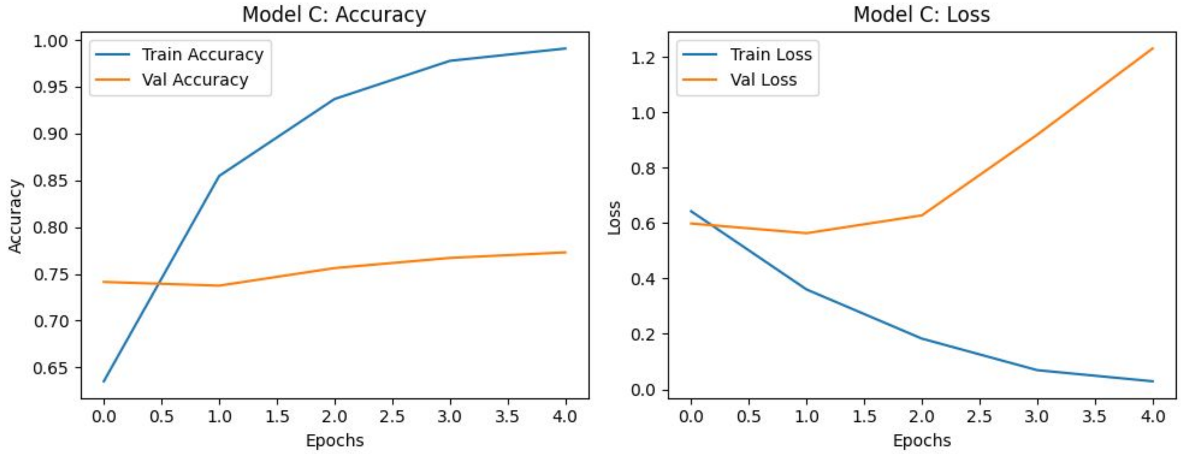
Best Parameters: {'alpha': 0.5, 'fit_prior': True}
Test Accuracy: 78.402%

Experimental results: BiLSTM

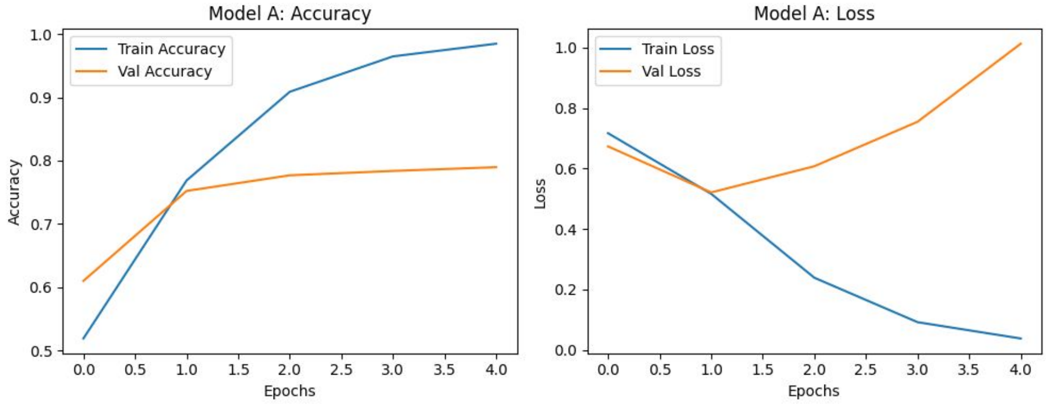
Model B – Medium Bi-LSTM (No Dropout)
Embedding: 64 | LSTM: 64→32 | Dropout: None | LR: 0.01



Model C – Large Bi-LSTM + Dropout
Embedding: 128 | LSTM: 64→32 | Dropout: 0.2 | LR: 0.01

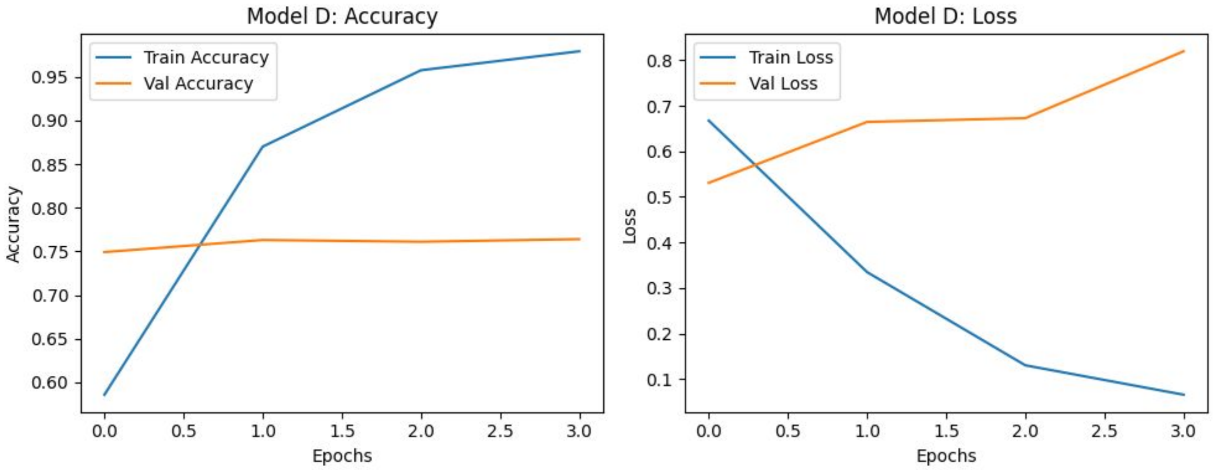


Model A – Small Bi-LSTM + Dropout
Embedding: 32 | LSTM: 32→16 | Dropout: 0.2 | LR: 0.01

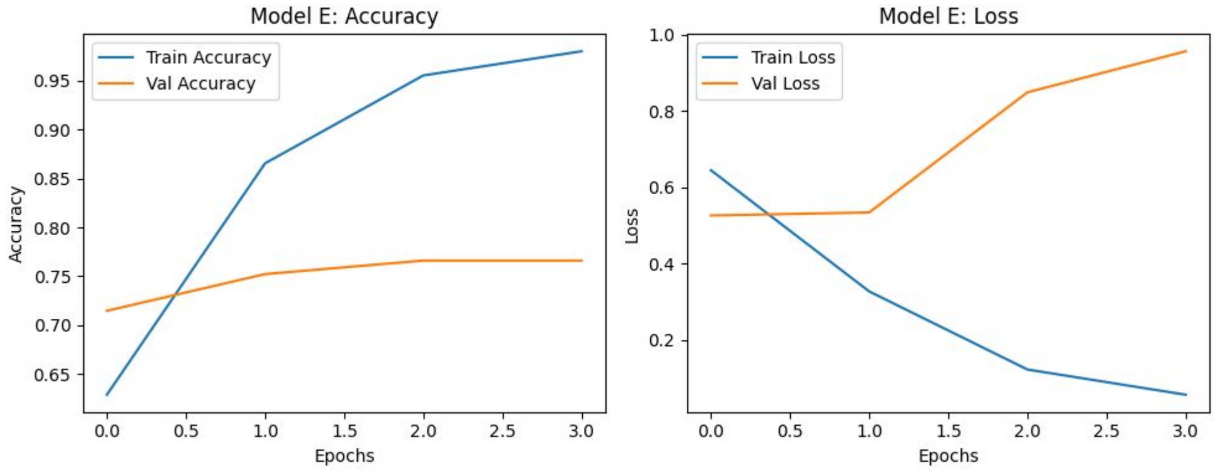


Experimental results: BiLSTM

Model D – Medium Balanced Bi-LSTM + Dropout
Embedding: 64 | LSTM: 32→16 | Dropout: 0.2 | LR: 0.01



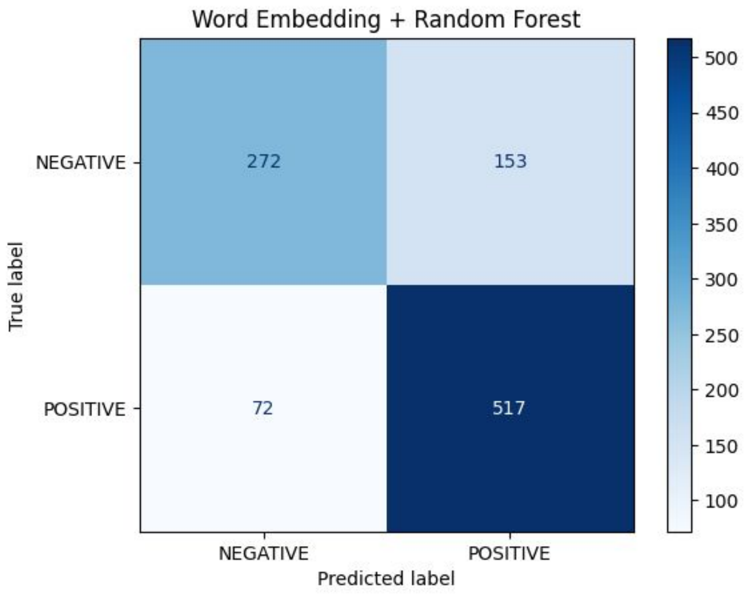
Model E – Deep Bi-LSTM (No Dropout)
Embedding: 64 | LSTM: 64→64 | Dropout: None | LR: 0.01

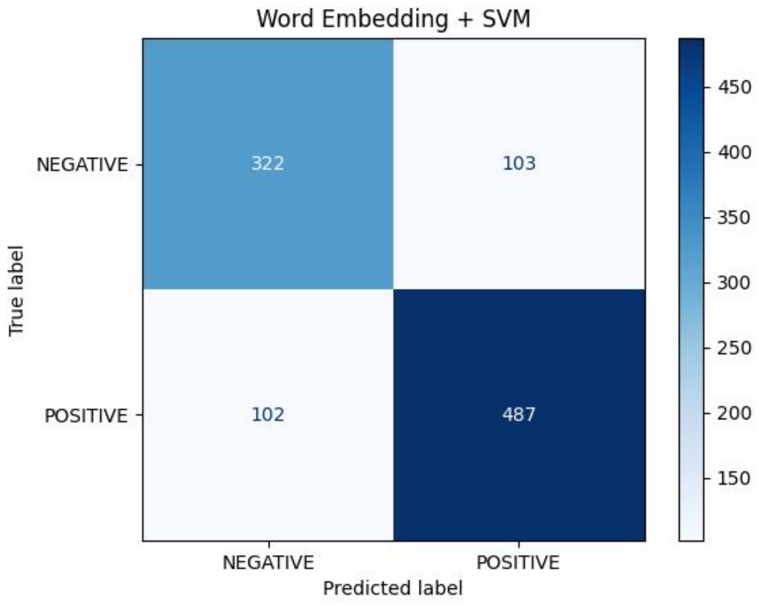


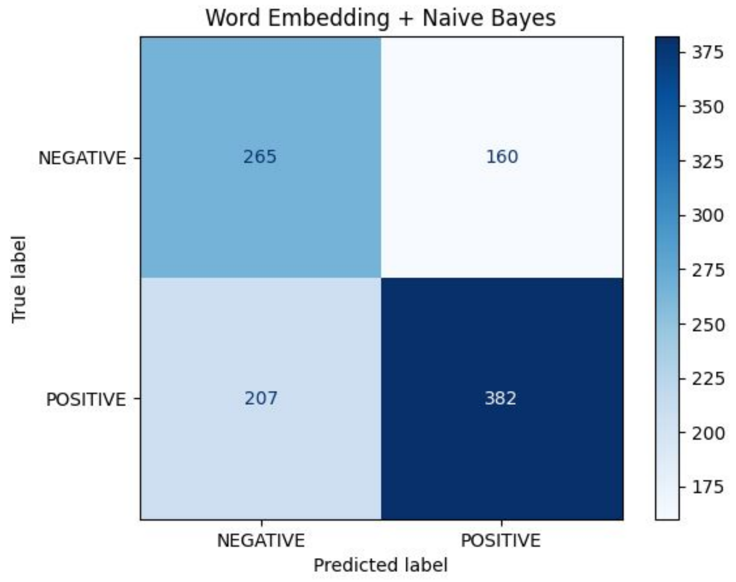
Experiment results: BiLSTM (Explanation)

- **Model A:**
  - **Accuracy Plot:**Train accuracy steadily increases to nearly 100%. Validation accuracy increases but plateaus early around 78%.
  - **Loss Plot:** Training loss drops consistently. Validation loss starts low but rises after about 2 epochs → signs of overfitting.
- **Model B:**
  - **Accuracy Plot:** Train accuracy rises to 99%. Validation accuracy is lower (~75%) and plateaus early.
  - **Loss Plot:** Train loss declines well. Validation loss increases heavily, especially after 2 epochs. Worse overfitting than Model A → model memorizes training data, but fails to generalize to validation.
- **Model C:**
  - **Accuracy Plot:** Train accuracy improves toward ~100%. Validation accuracy slightly improves (~76–78%) but stays relatively flat.
  - **Loss Plot:** Train loss steadily decreases. Validation loss first decreases but then increases sharply → overfitting.
- **Model D:**
  - **Accuracy Plot:** Train accuracy rises well (~95%). Validation accuracy modestly improves (~76%).
  - **Loss Plot:** Train loss drops nicely. Validation loss rises slowly — less overfitting compared to earlier models.
- **Model E:**
  - **Accuracy Plot:** Train accuracy quickly reaches ~98%. Validation accuracy is highest (~80%) and most stable across epochs.
  - **Loss Plot:** Train loss decreases sharply. Validation loss increases moderately but not as sharply as others.

Experimental results: Word Embedding





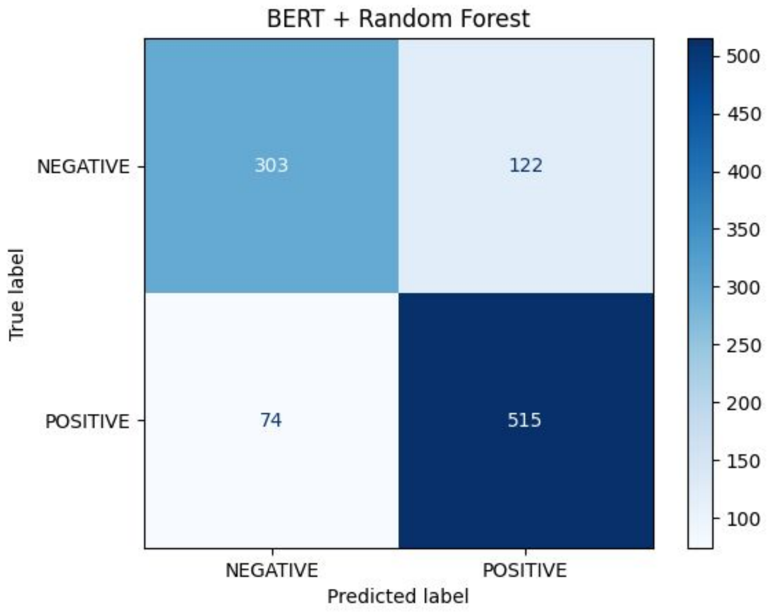


Best Parameters: {'bootstrap': True, 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200}
Test Accuracy: 77.810%

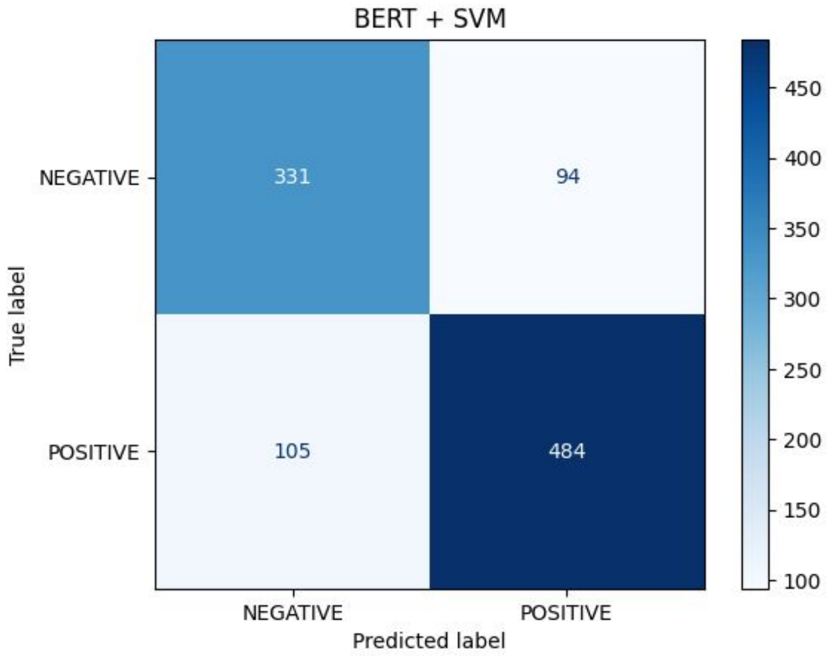
Best Parameters: {'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}
Test Accuracy: 79.783%

Best Parameters: {'alpha': 0.1, 'fit_prior': False}
Test Accuracy: 63.806%

Experimental results: BERT

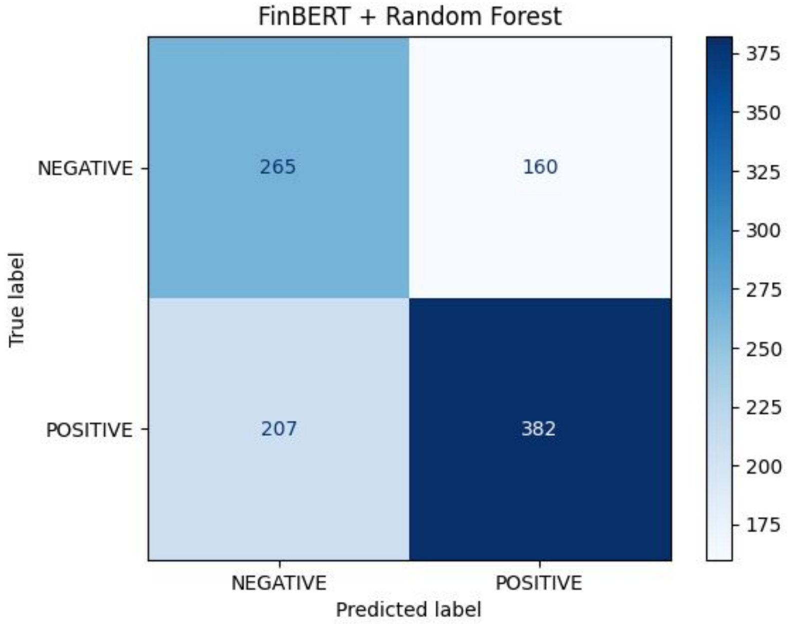


Best Parameters: {'bootstrap': False, 'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 200}
Test Accuracy: 63.806%

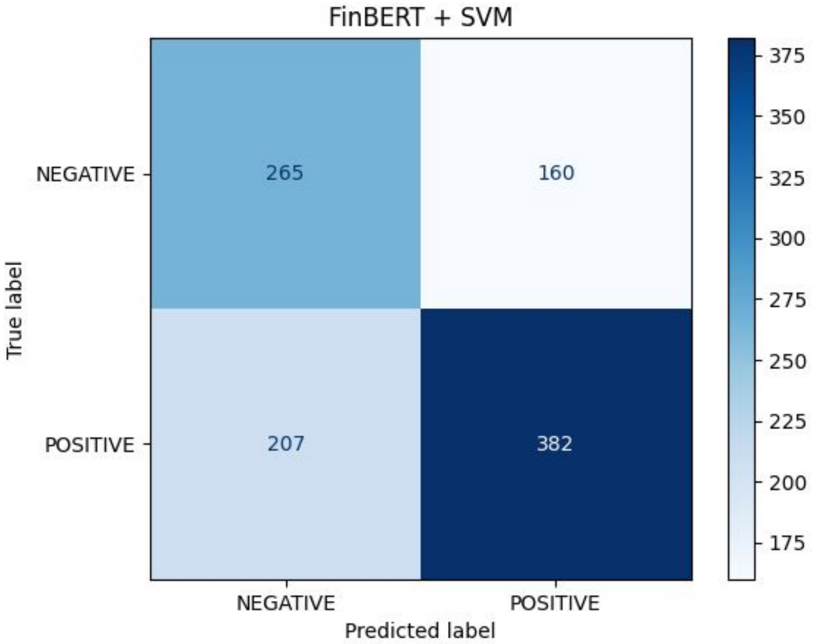


Best Parameters: {'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}
Test Accuracy: 63.806%

Experimental results: FinBERT

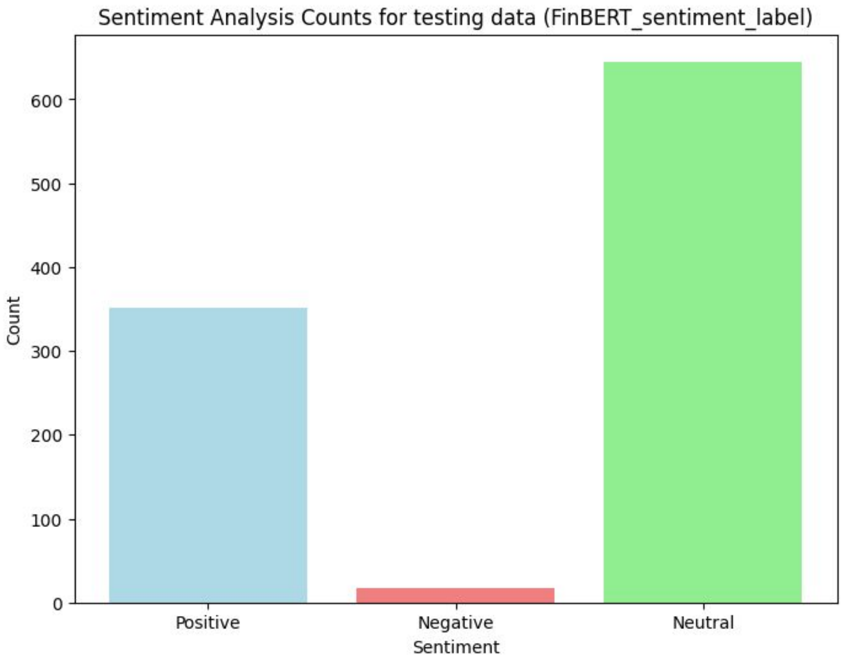


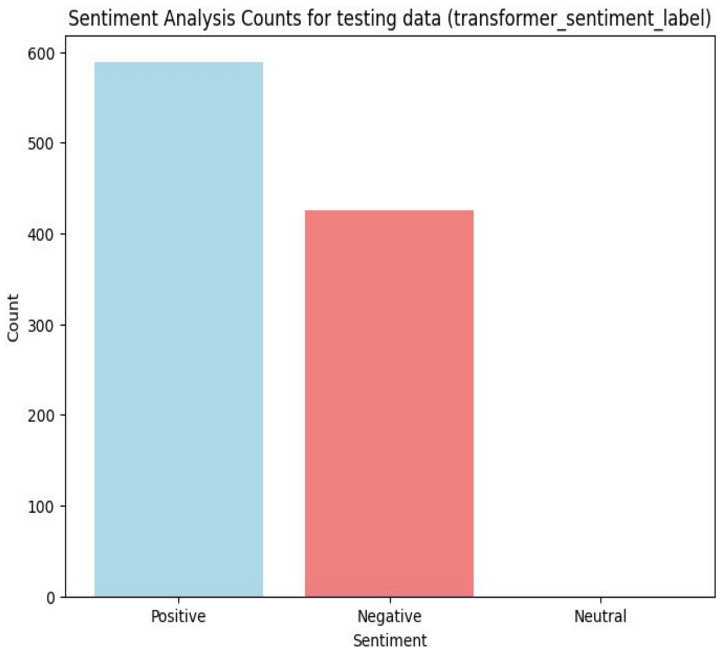
Best Parameters: {'bootstrap': False, 'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 50}
Test Accuracy: 63.806%

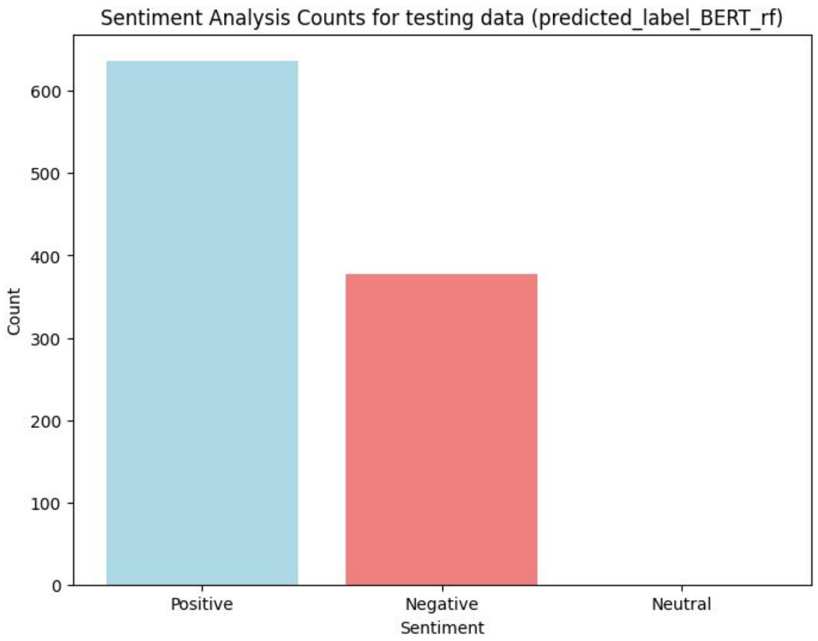


Best Parameters: {'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}
Test Accuracy: 63.806%

Experimental results: Sentiment Label Counts



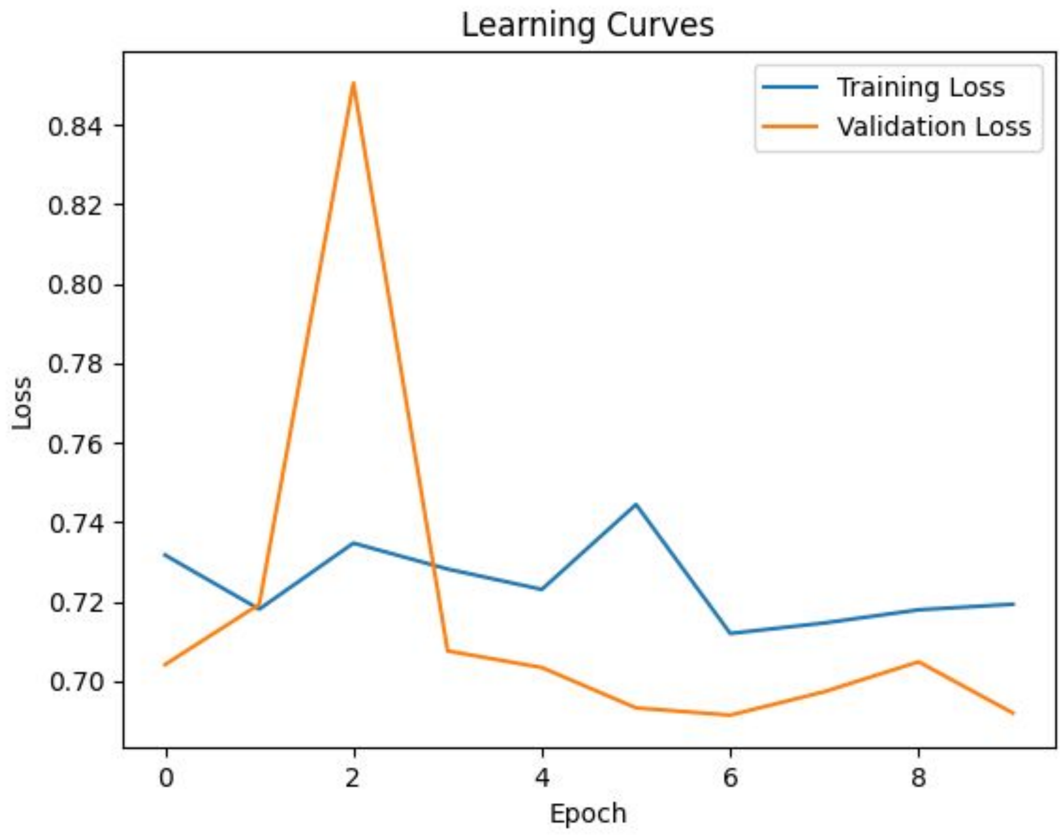




Overall Knowledge Extracted

- TF-IDF + SVM achieved the highest test accuracy (81.76%), showing strong generalization and robustness using basic textual features.
- Among the Bi-LSTM models, Model E (Deep Bi-LSTM, No Dropout) showed the highest validation accuracy (~80%) and stable convergence, making it the top-performing neural architecture.
- Despite their power, FinBERT and BERT models paired with Random Forest or SVM consistently plateaued at ~63.81% accuracy—likely due to lack of fine-tuning and smaller training sizes.
- Word Embedding + Naive Bayes had the lowest performance (63.8%), highlighting the limitations of simple classifiers with dense features.
- Classic TF-IDF features, especially when combined with SVM and Naive Bayes, consistently outperformed more complex word embeddings in non-deep settings.
- Confusion matrices revealed a heavier bias toward positive predictions, suggesting a skew in sentiment distribution within the data.
- Dropout helped reduce overfitting in shallow Bi-LSTMs, but deeper models like Model E performed better even without dropout, indicating the importance of capacity over regularization.
- FinBERT heavily classifies sentiments as Neutral, while the Transformer model and BERT-Random Forest model predict mostly Positive and Negative sentiments, with no Neutral labels.

Directions for Future Research



- Currently, the dataset is limited to pre-scraped headlines. A key future step is to scrape fresh data from 2023 to the end of 2024 to enhance model generalization. Instead of a single news headline per day - use all (merge all news into a single file).
- Instead of a single headline per day, aggregate all daily headlines into a unified text block to better capture sentiment context.
- Fine-tune BERT & FinBERT models by experimenting with optimizers, loss functions, and evaluation metrics, and evaluating test loss and accuracy across 5–15 epochs for stability insights.
- Leverage TensorFlow Data Pipeline by using tf.data to efficiently load, batch, and feed data into BERT/FinBERT for more scalable training and smoother learning curve analysis.

This is what I could do so far with epochs = 10 and got accuracy score of 58% for the Fin-BERT model.

References

- **Data: Stock Fundamental Data API**
- **Bidirectional LSTM in NLP: https://www.geeksforgeeks.org/bidirectional-lstm-in-nlp/**
- **Sentiment Classification Using BERT: https://www.geeksforgeeks.org/sentiment-classification-using-bert/**
- **FinBERT: https://huggingface.co/yiyanghkust/finbert-tone**
- **Sentiment Analysis with LSTM: https://www.analyticsvidhya.com/blog/2022/01/sentiment-analysis-with-lstm/**
- **GitHub (the data and the code): https://github.com/erica-prog/stock_news_headlines_sentiment_analysis**