FINANCIAL NEWS SENTIMENT ANALYSIS

PRESENTED BY:

JUNJIE CHEN

PABLO MOLLA

PAVLO POLIUHA

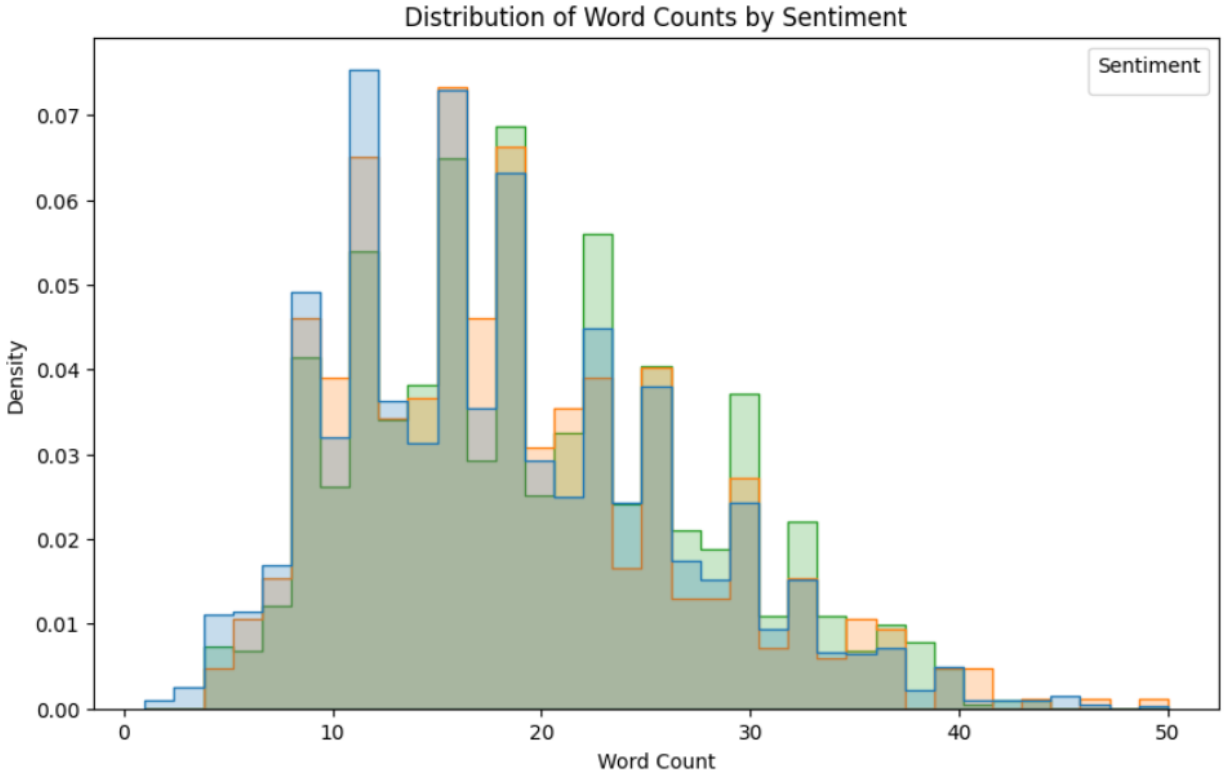
AGENDA

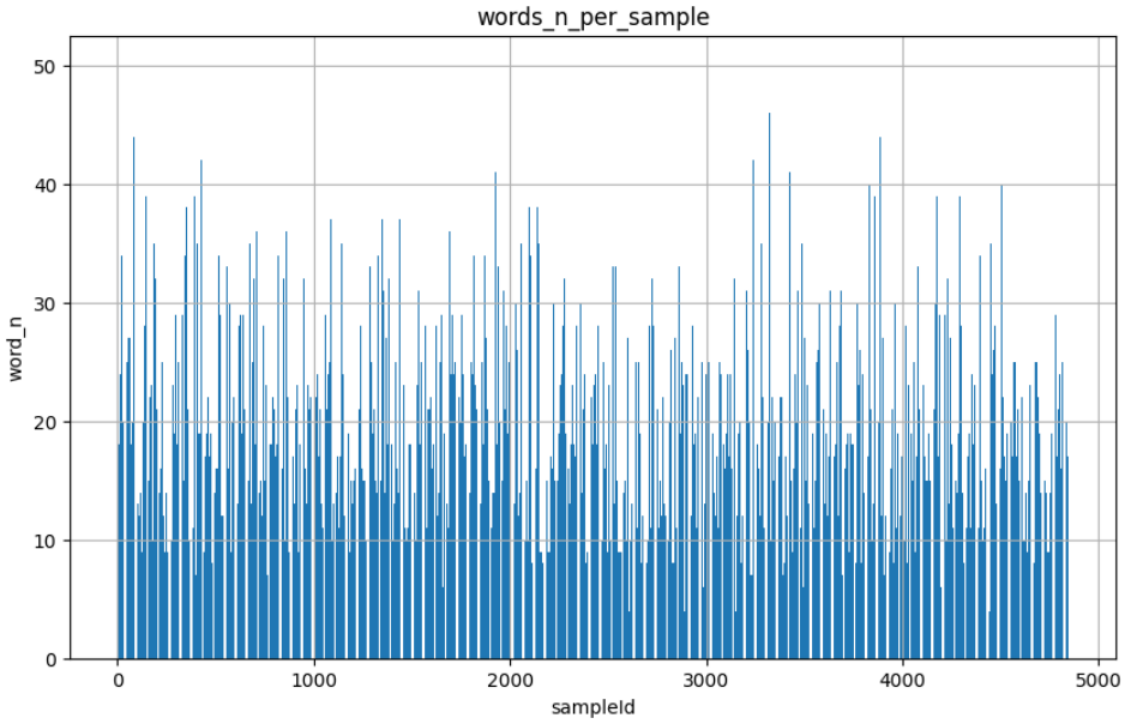
1) Dataset Description
2) Data Preprocessing & Analysis
3) Base Models
4) Transformer
5) Fine-Tuning
6) Results
7) Conclusion

DATASET DESCRIPTION

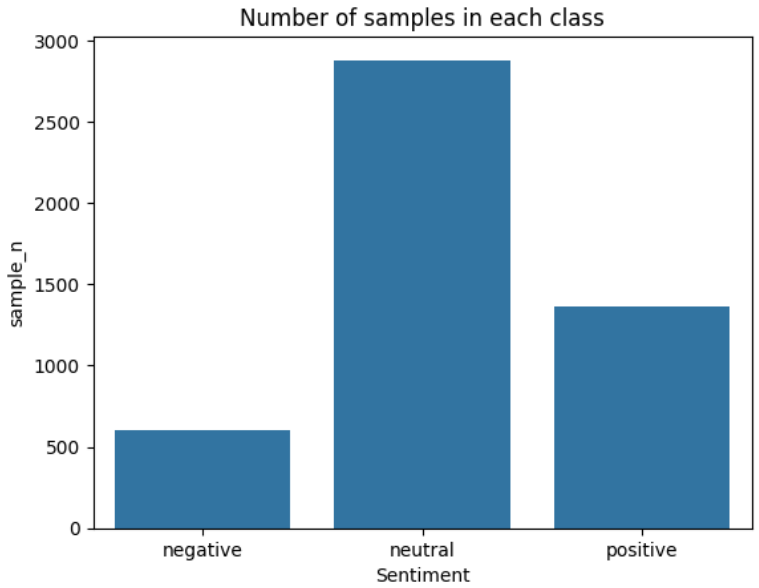
Adjusted for changes in the Group structure , the Division 's net sales increased by 1.7 %.	→	positive
No blind-spots coming from 1 vantage point allows investigators to see the whole story in high resolution with no unbroken video.	→	neutral
Cash Flow from Operations for the most recent quarter also reached a eight year low.	→	negative

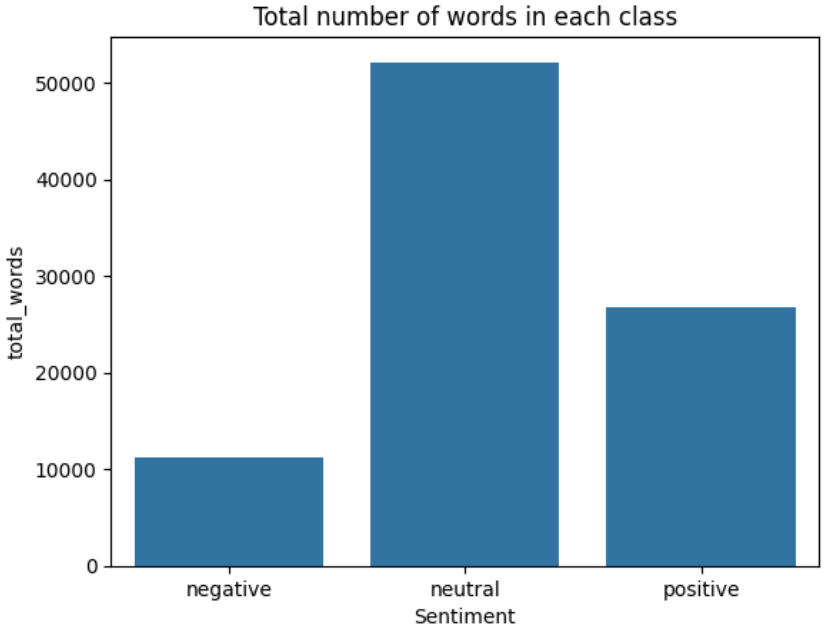
DATA PREPROCESSING & ANALYSIS(1)

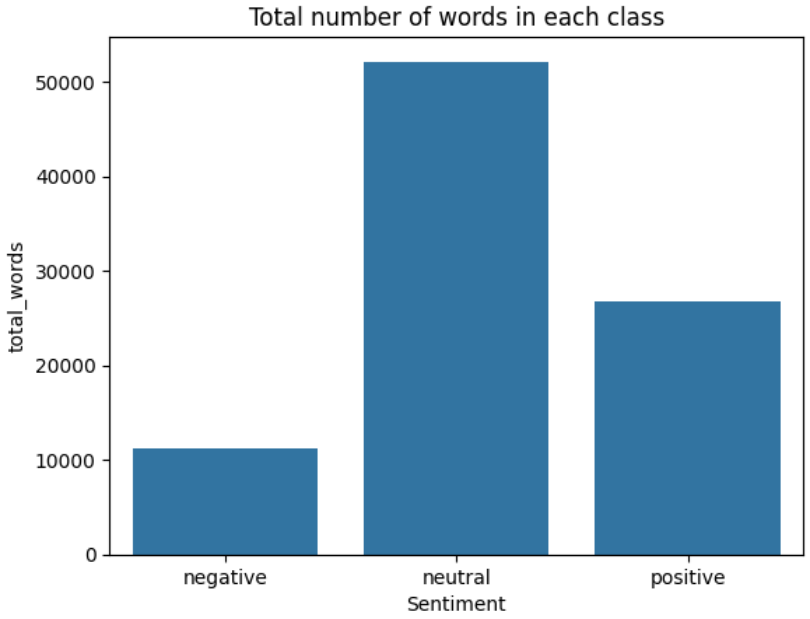




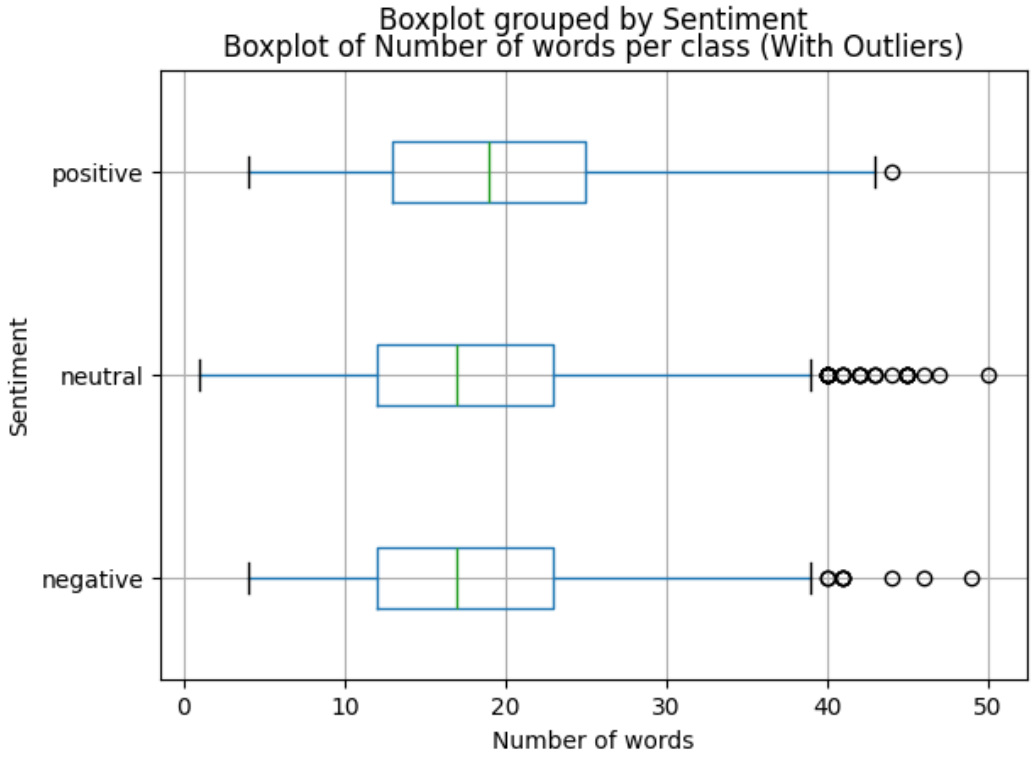
DATA PREPROCESSING & ANALYSIS(2)

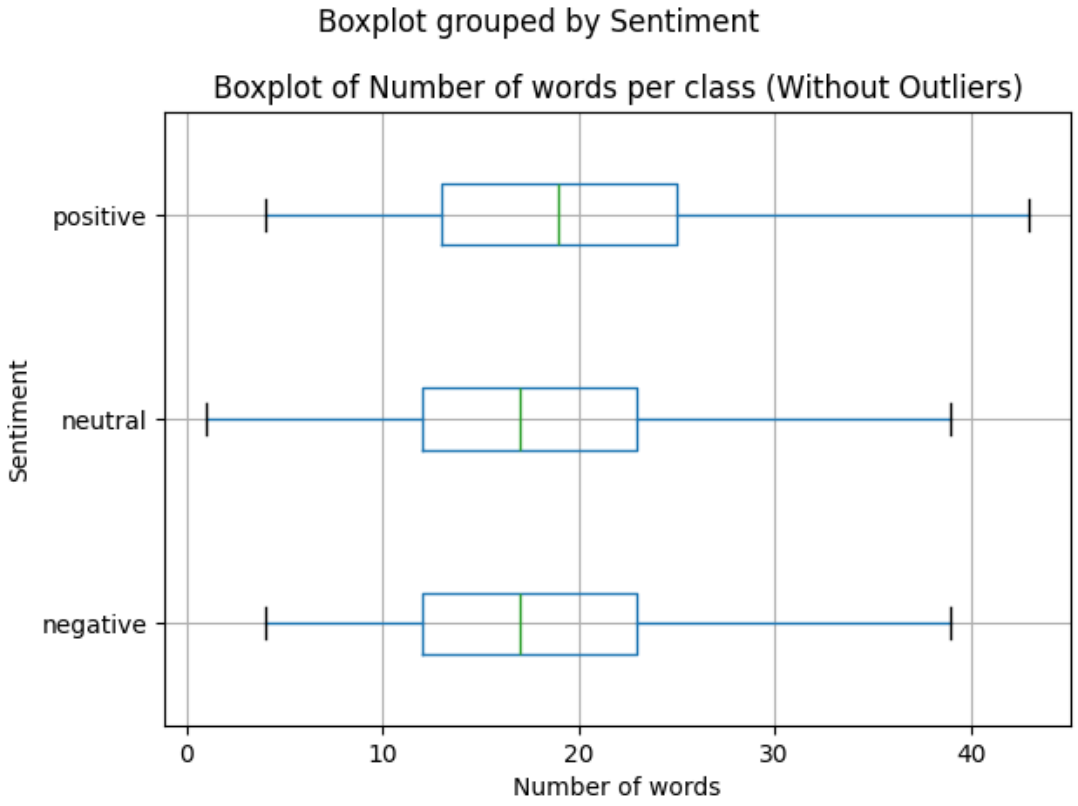




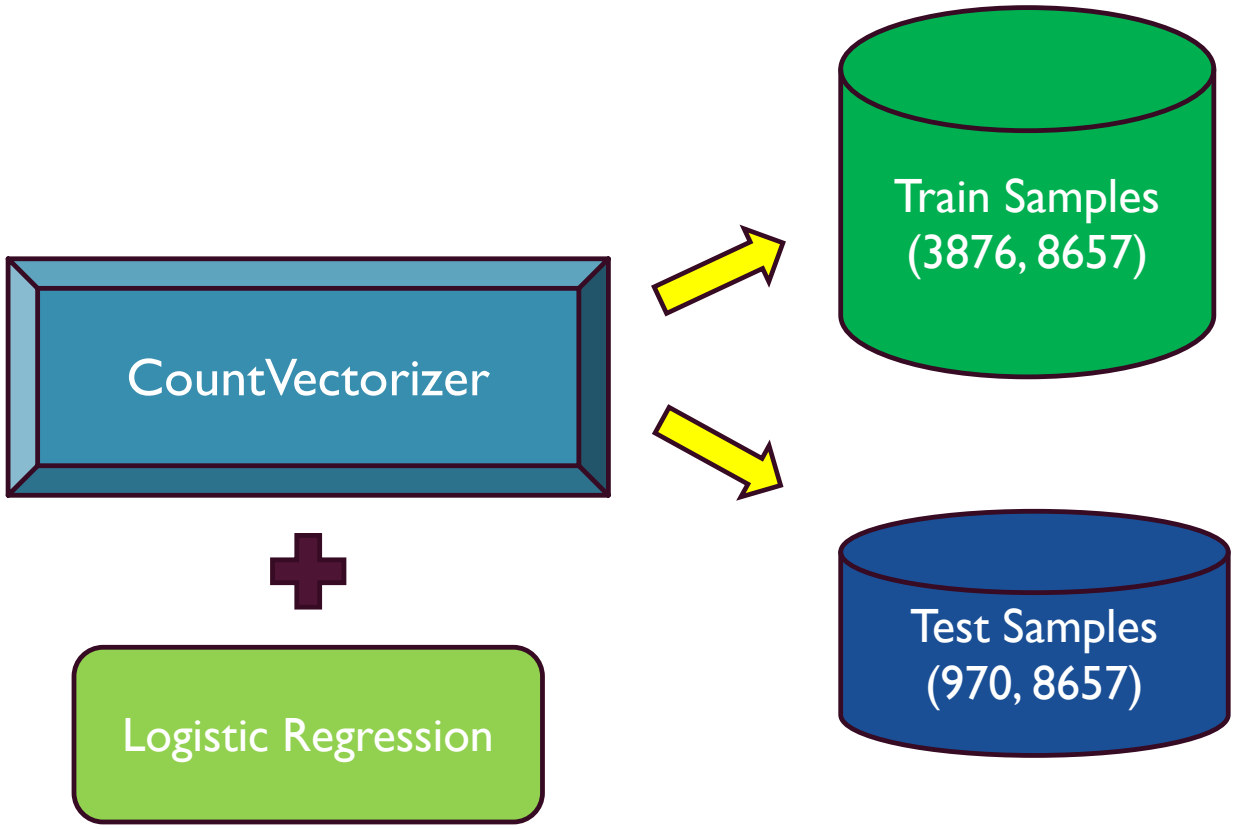


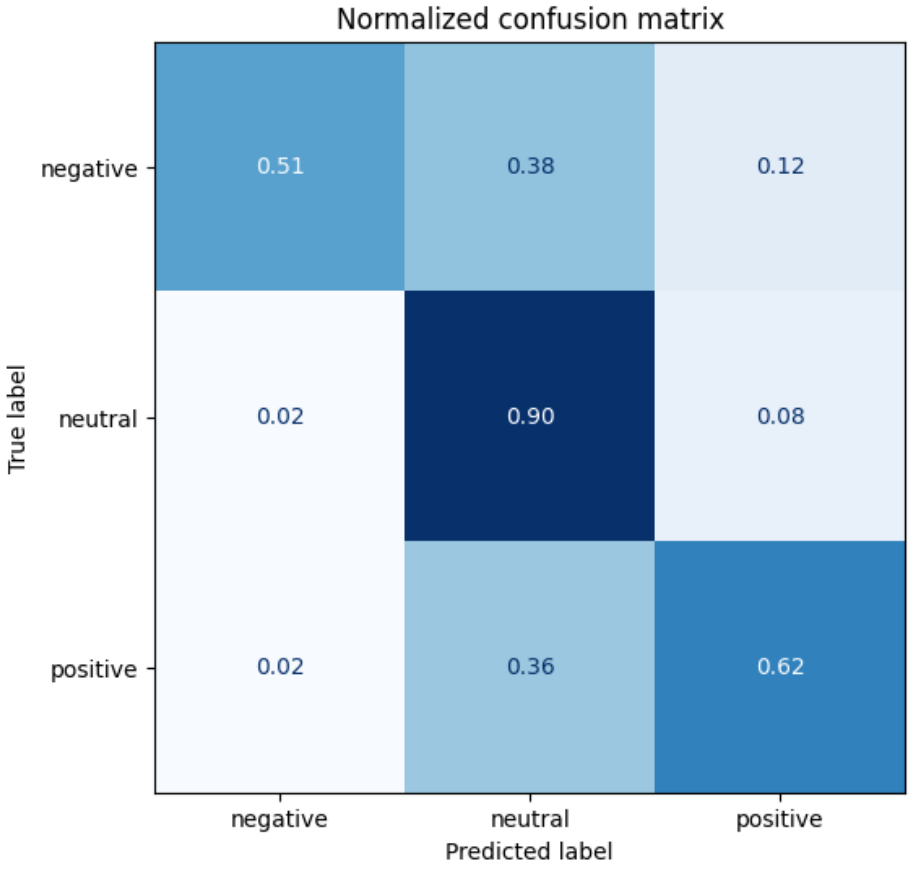
DATA PREPROCESSING & ANALYSIS(3)



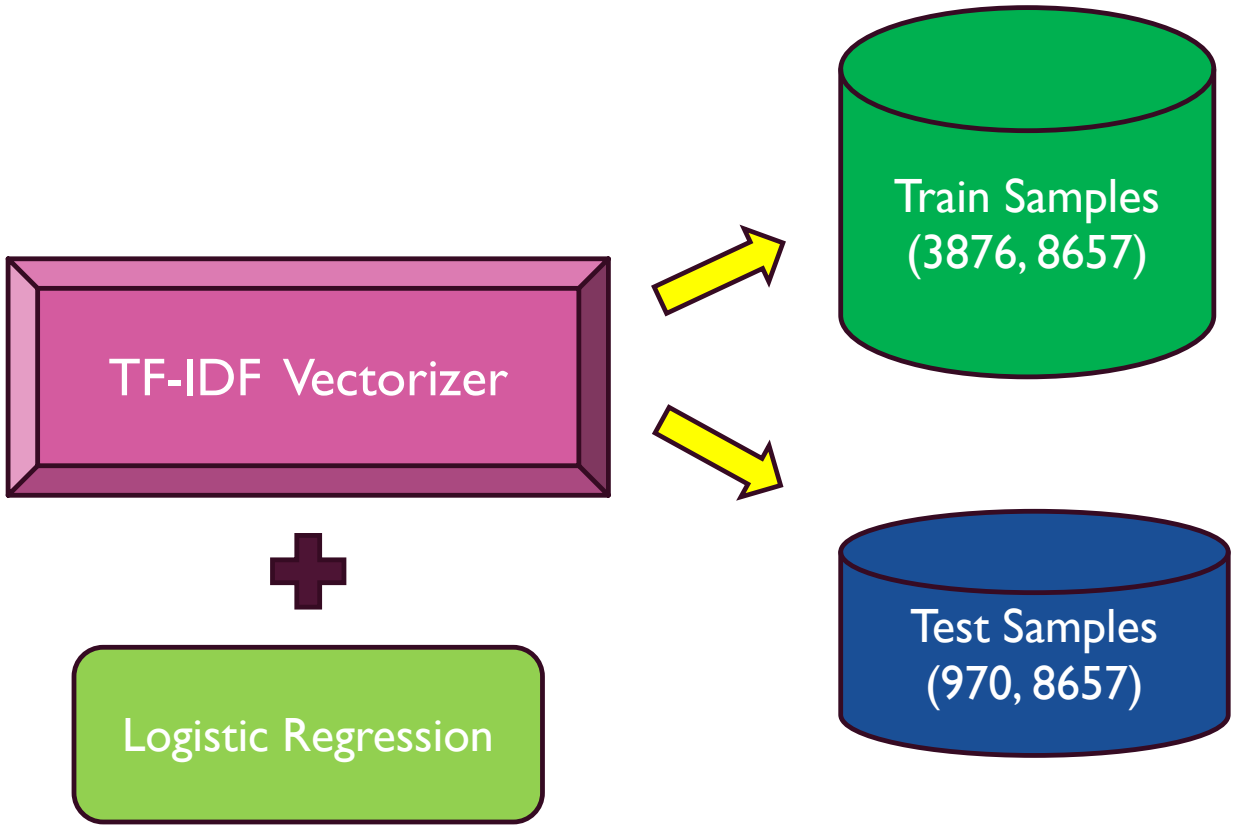


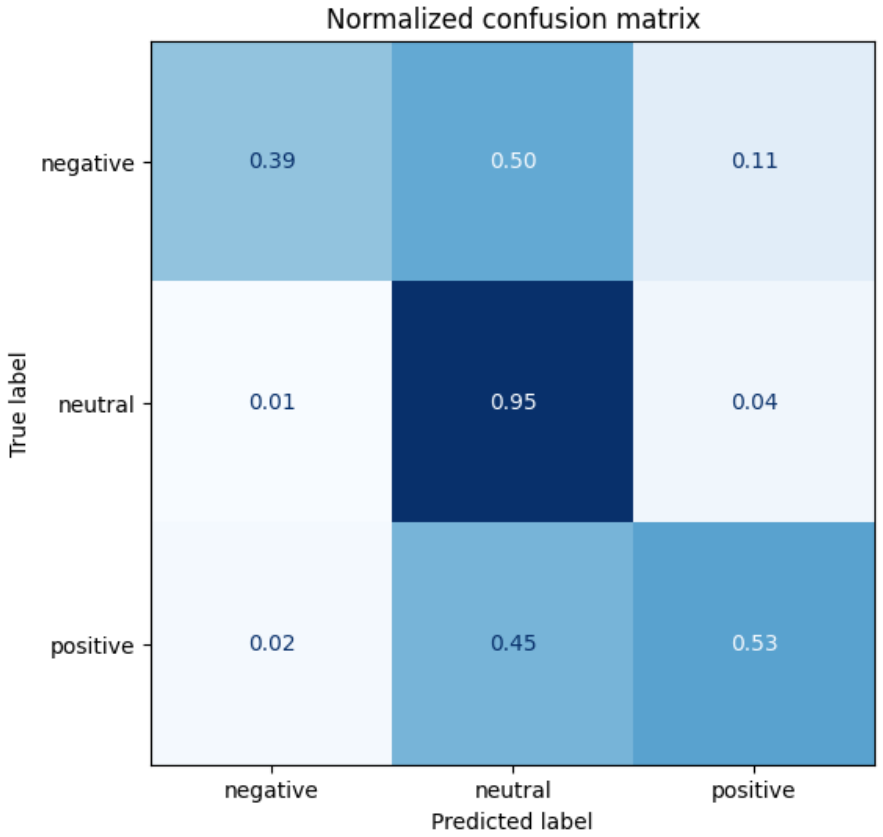
BASE MODEL (1)


CountVectorizer
Logistic Regression
Train Samples
(3876, 8657)
Test Samples
(970, 8657)

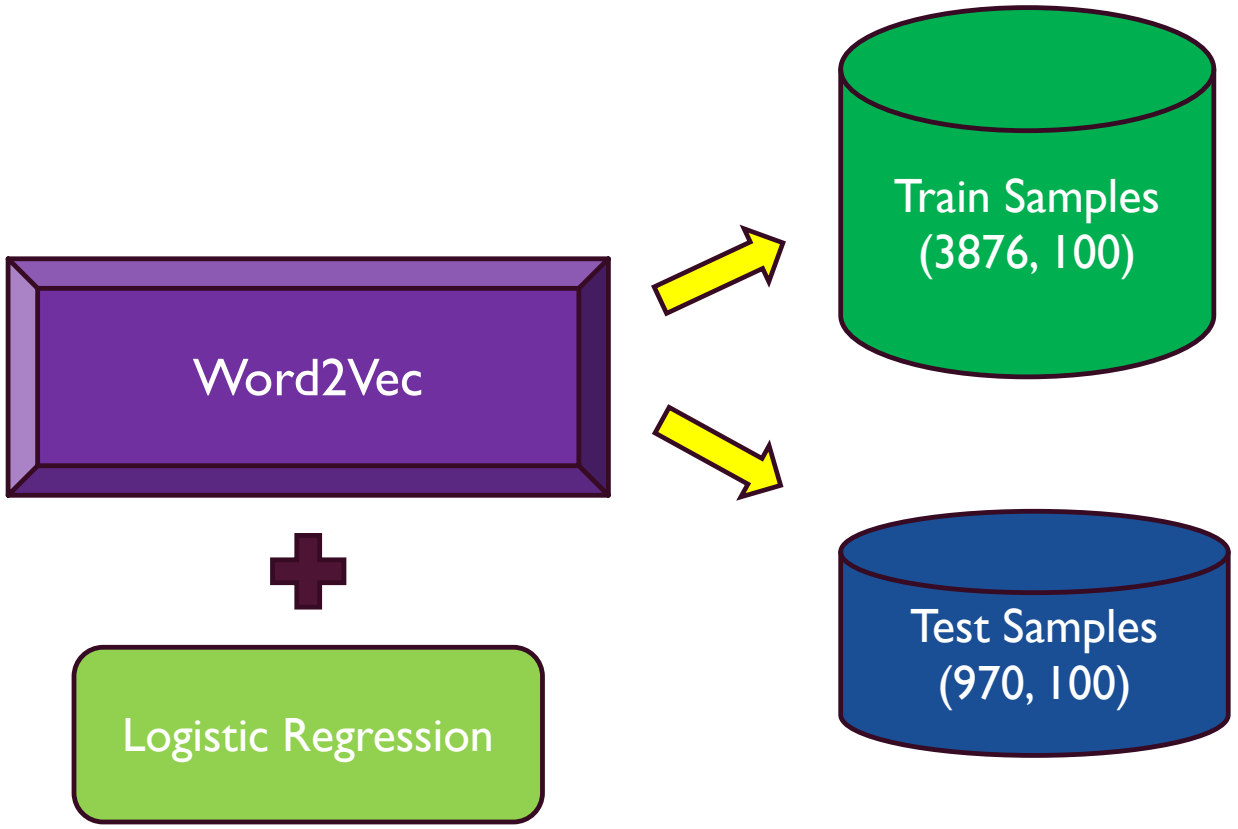

Normalized confusion matrix
negative
neutral
positive
True label
0.51
0.38
0.12
0.02
0.90
0.08
0.02
0.36
0.62
negative
neutral
positive
Predicted label

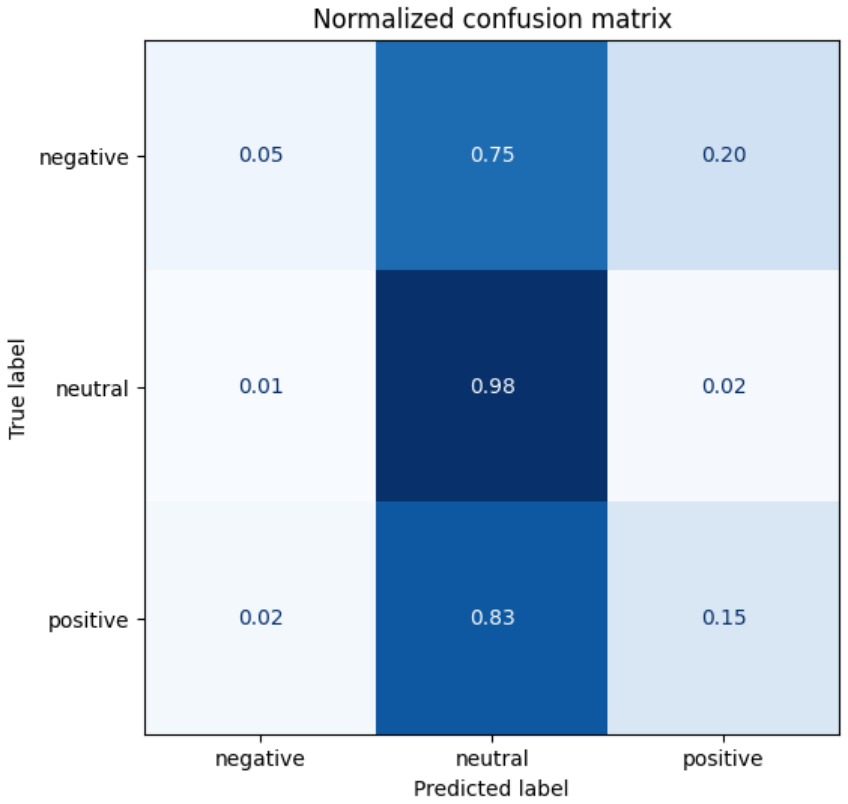
BASE MODEL (2)


TF-IDF Vectorizer
Logistic Regression
Train Samples
(3876, 8657)
Test Samples
(970, 8657)

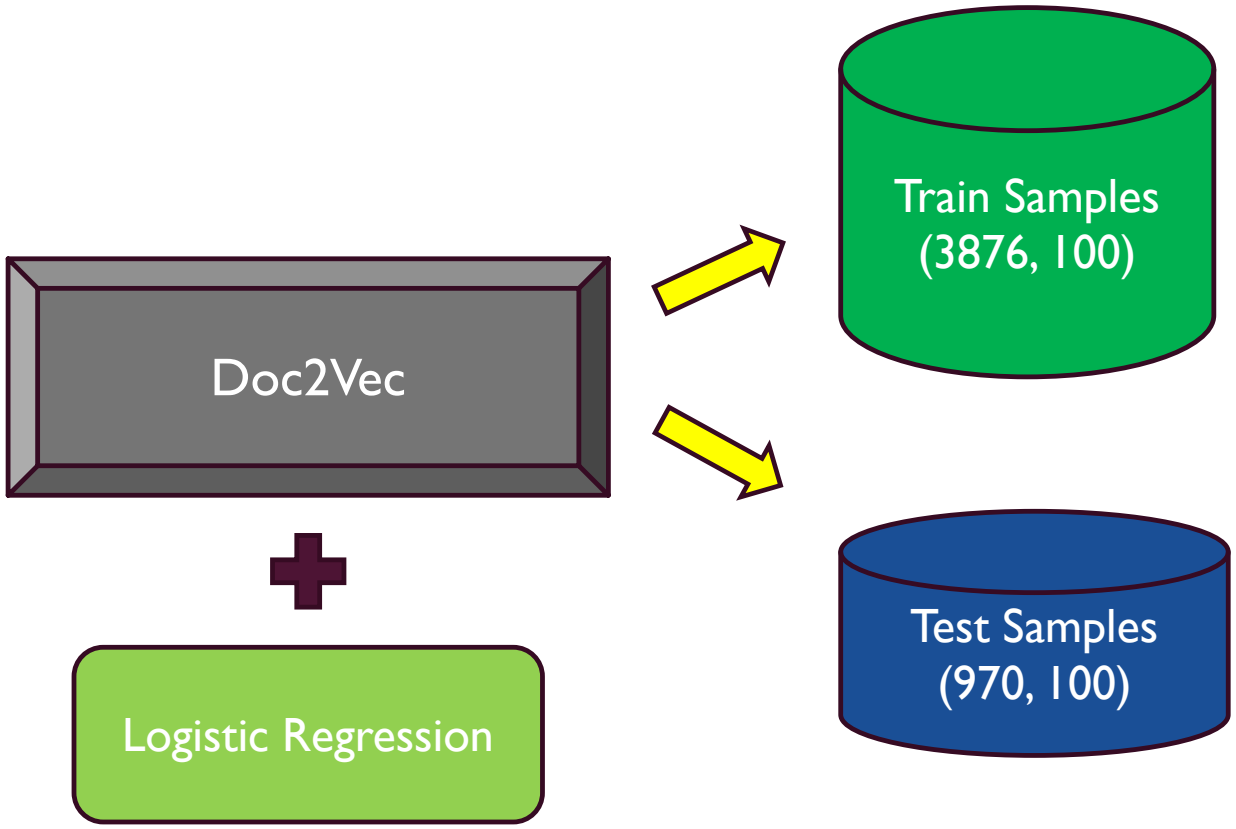

Normalized confusion matrix
negative
neutral
positive
True label
0.39
0.50
0.11
0.01
0.95
0.04
0.02
0.45
0.53
negative
neutral
positive
Predicted label

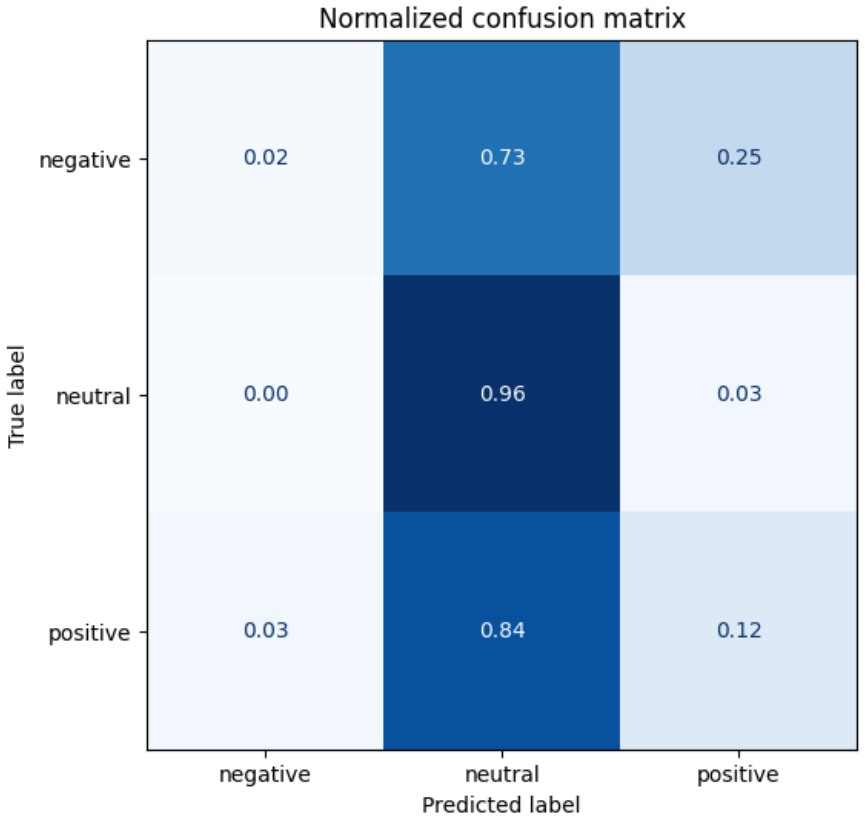
BASE MODEL (3)


Word2Vec
Logistic Regression
Train Samples
(3876, 100)
Test Samples
(970, 100)


Normalized confusion matrix
negative
neutral
positive
True label
0.05
0.75
0.20
0.01
0.98
0.02
0.02
0.83
0.15
negative
neutral
positive
Predicted label

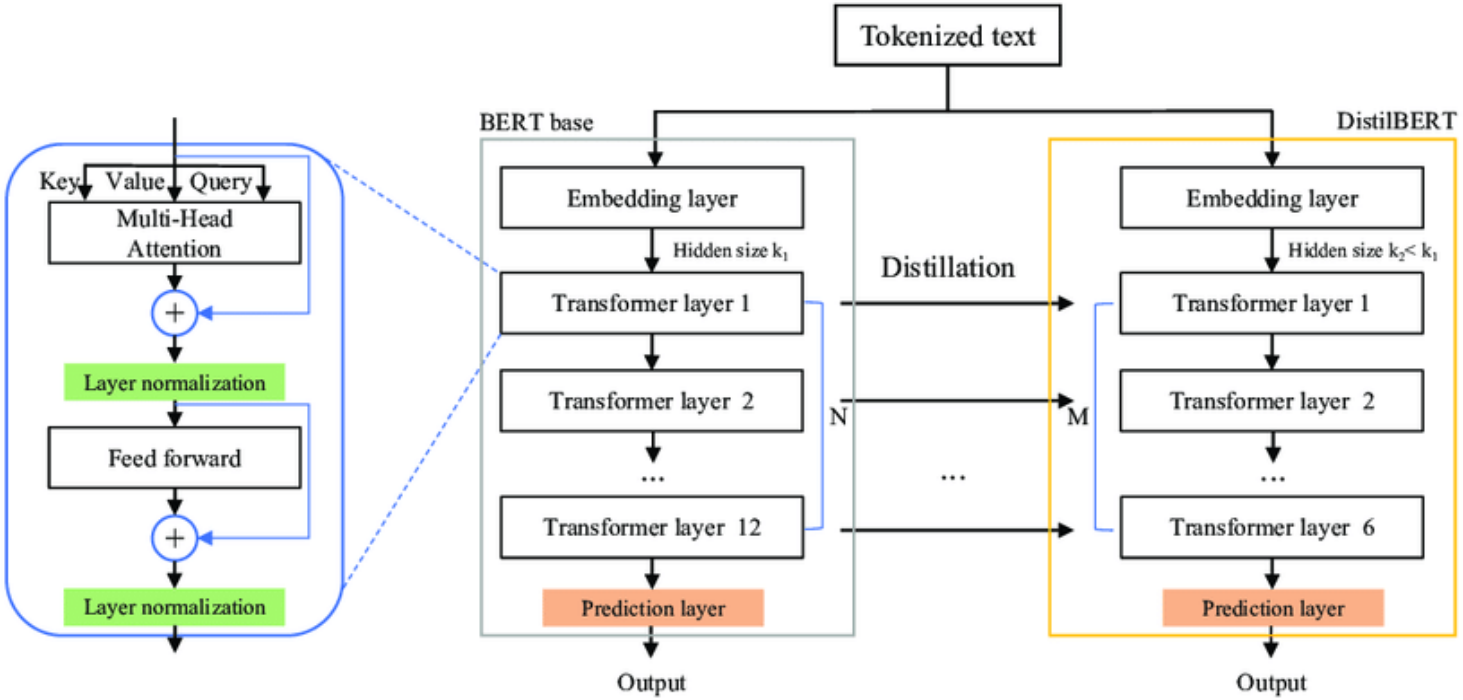
BASE MODEL (4)


Doc2Vec
Logistic Regression
Train Samples
(3876, 100)
Test Samples
(970, 100)


Normalized confusion matrix
negative
neutral
positive
True label
0.02
0.73
0.25
0.00
0.96
0.03
0.03
0.84
0.12
negative
neutral
positive
Predicted label

TRANSFORMER MODEL

DistilBERT



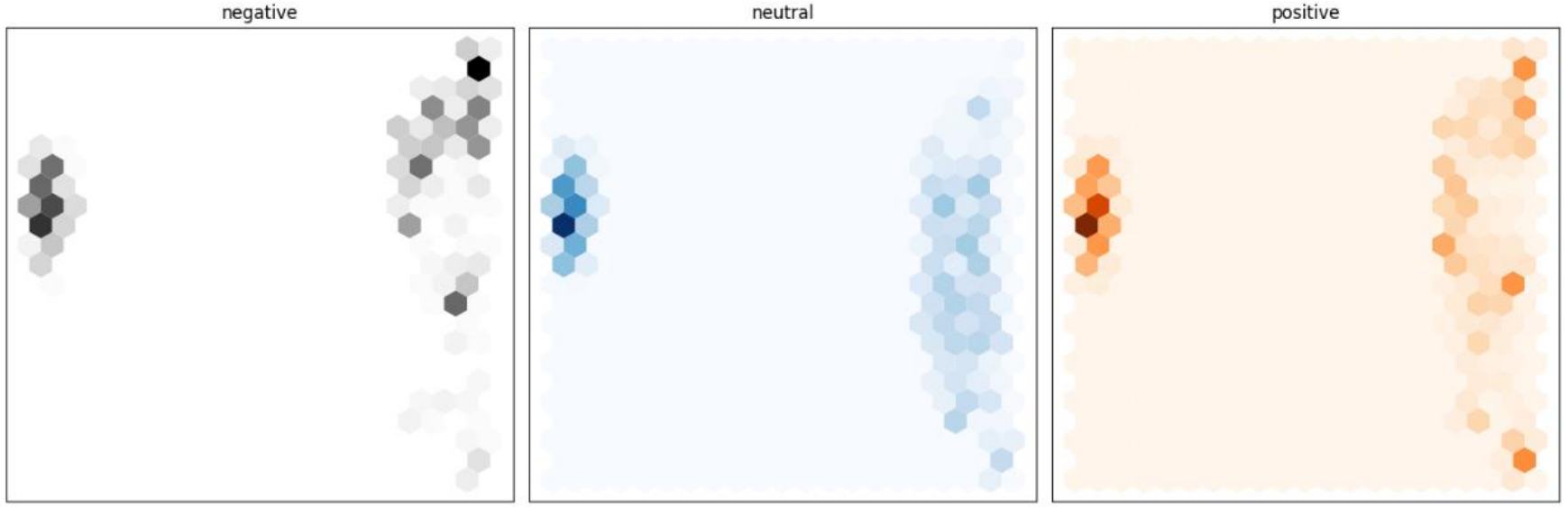
SIZE & SPEED

TRAINING METHOD

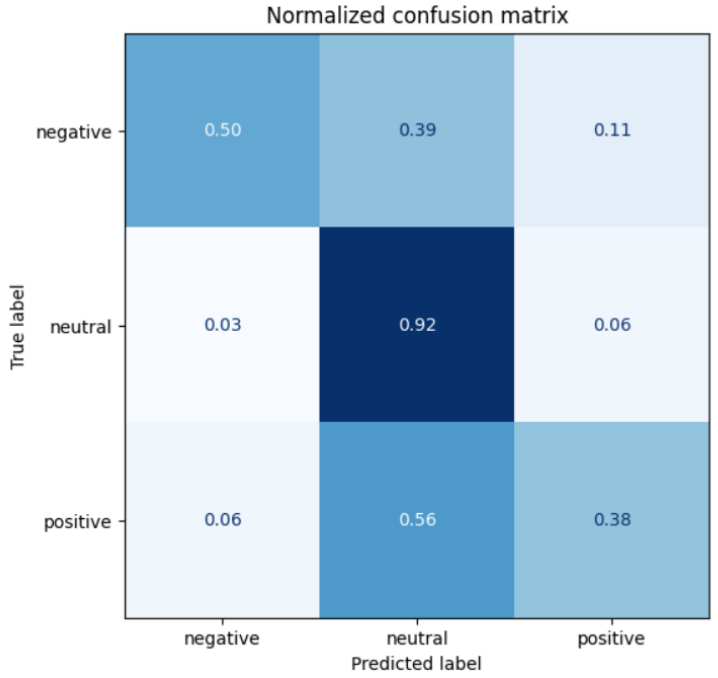
PERFORMANCE

PARAMETERS

SENTIMENT REGIONS

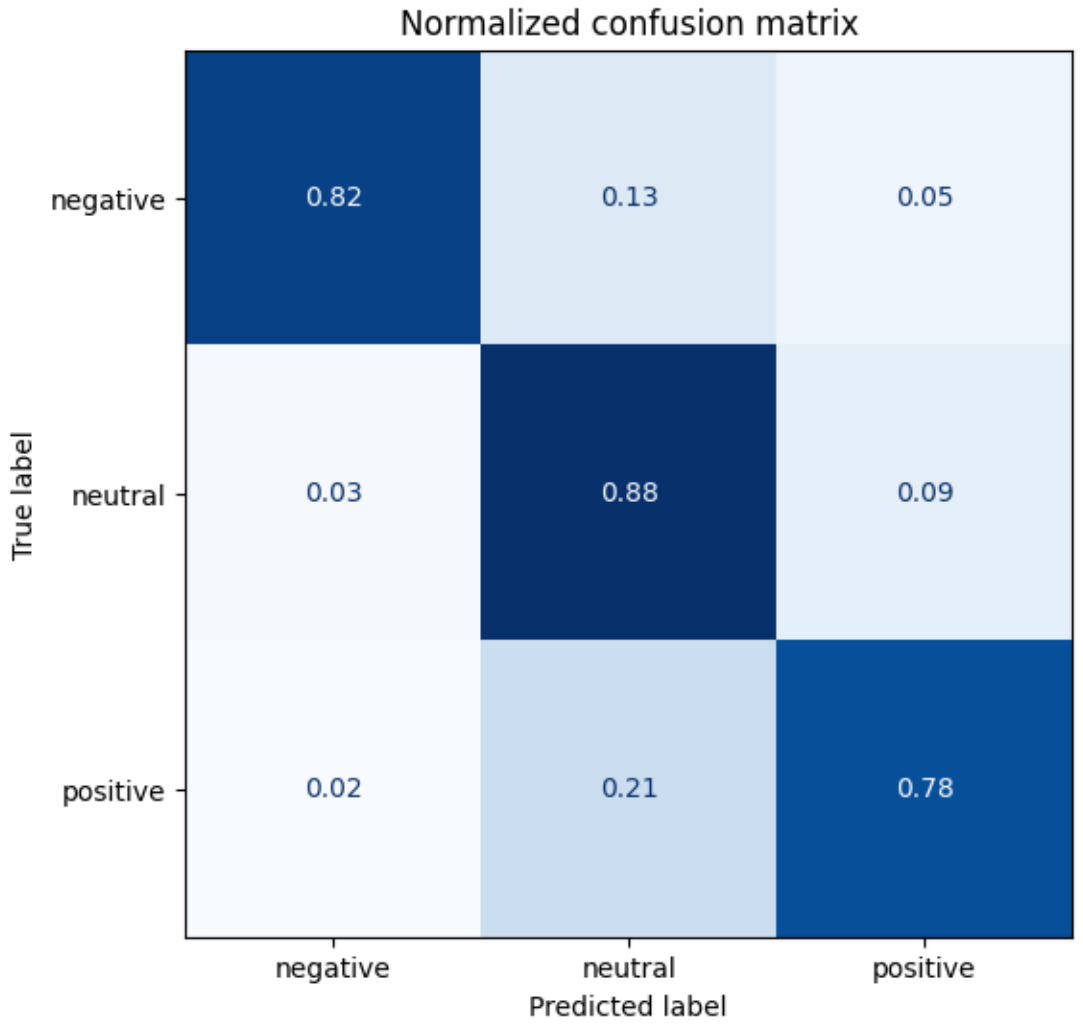


CONFUSION MATRIX OF DISTILBERT+ LOGISTIC REGRESSION

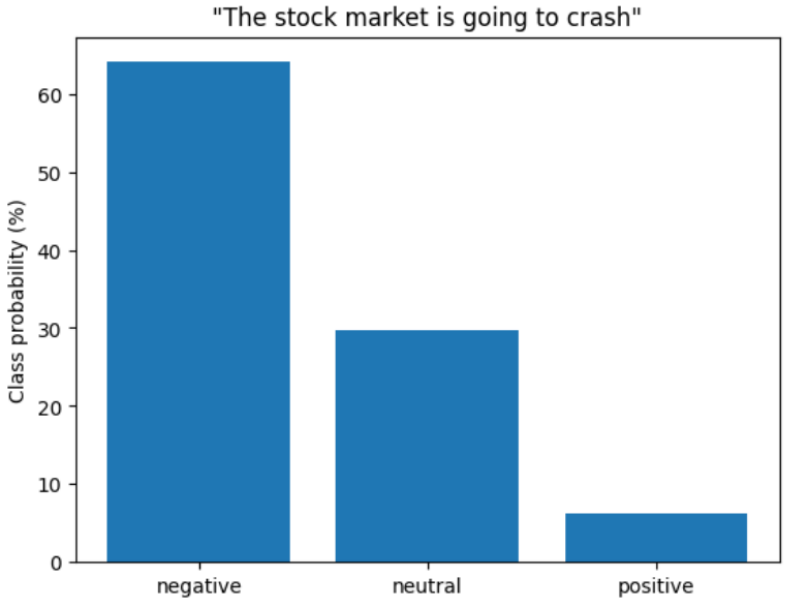


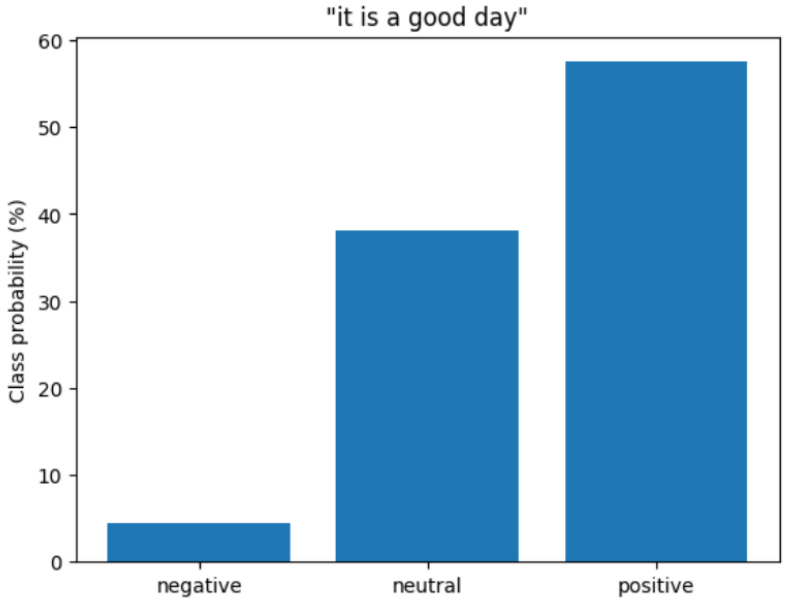
FINE-TUNED TRANSFORMER

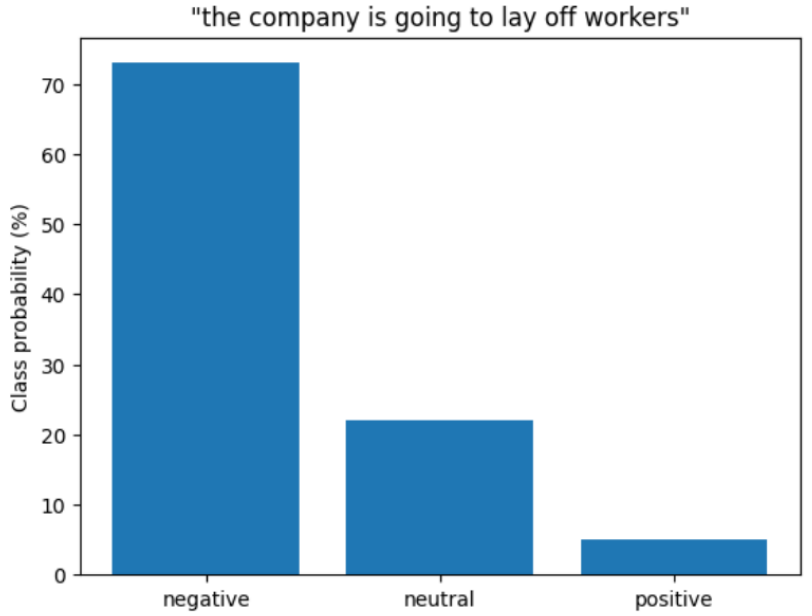
CONFUSION MATRIX OF DISTILBERT FINE-TUNED



RESULTS


"The stock market is going to crash"
Class probability (%)
0
10
20
30
40
50
60
negative
neutral
positive


"it is a good day"
Class probability (%)
0
10
20
30
40
50
60
negative
neutral
positive


"the company is going to lay off workers"
Class probability (%)
0
10
20
30
40
50
60
70
negative
neutral
positive

CONCLUSION

Current Results

- Fine-Tuned DistilBERT accuracy takes the lead on our dataset (85%)
- CountVectorizer and TF-IDF Vectorizer with Logistic Regression show slightly worse results (77%, 76% respectively)
- DistilBERT(as feature extraction) using Logistic Regression is on the fourth place with 73% accuracy
- Word2Vec, Doc2Vec with Logistic Regression provide the worst results (64% and 60% respectively)

Plans for the future

- Tackle imbalanced data
- Focus on more training data
- Make tests with other datasets
- Compare the results with other Pre-Trained models from Hugging Face

THANK YOU FOR YOUR ATTENTION