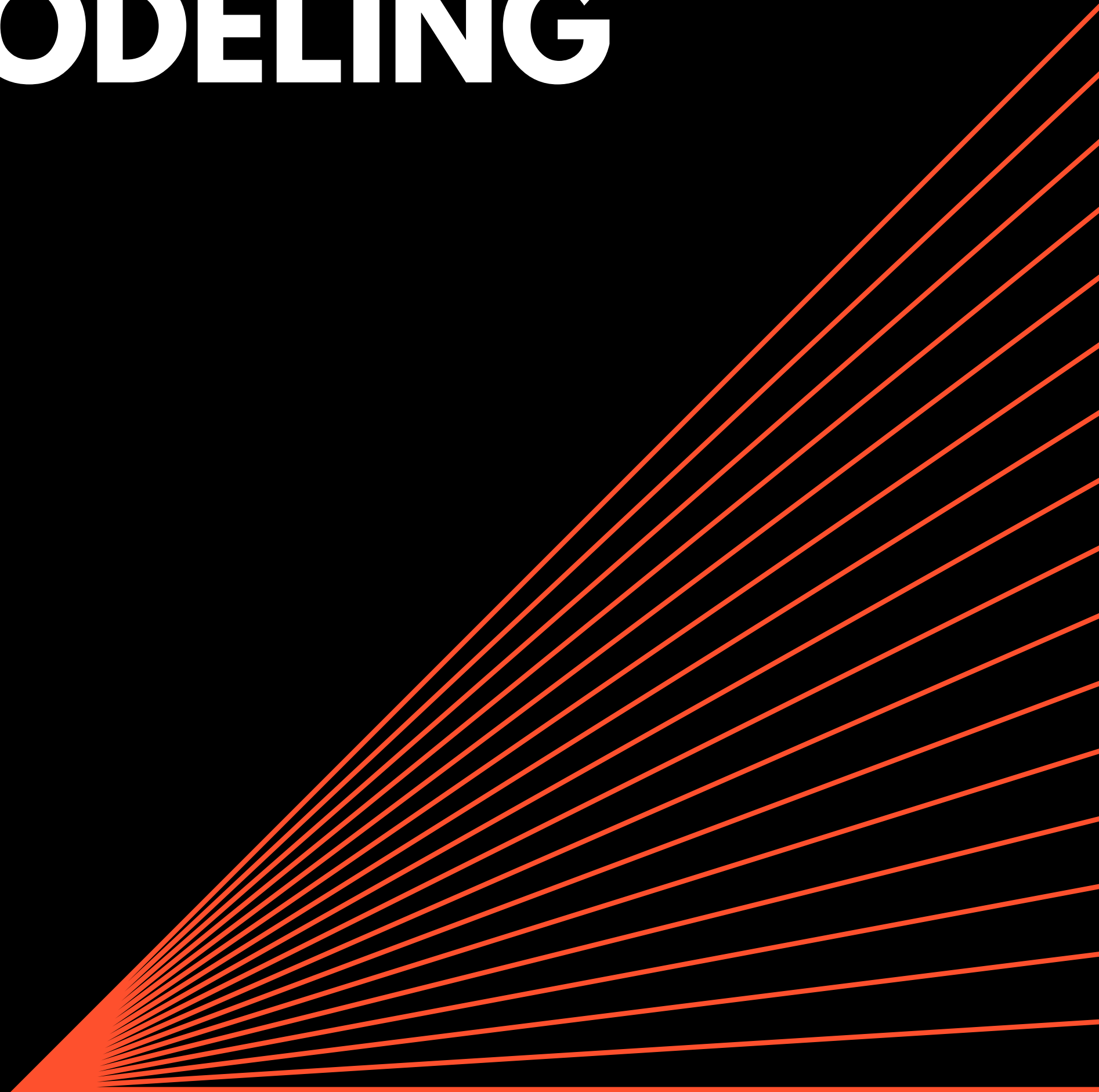


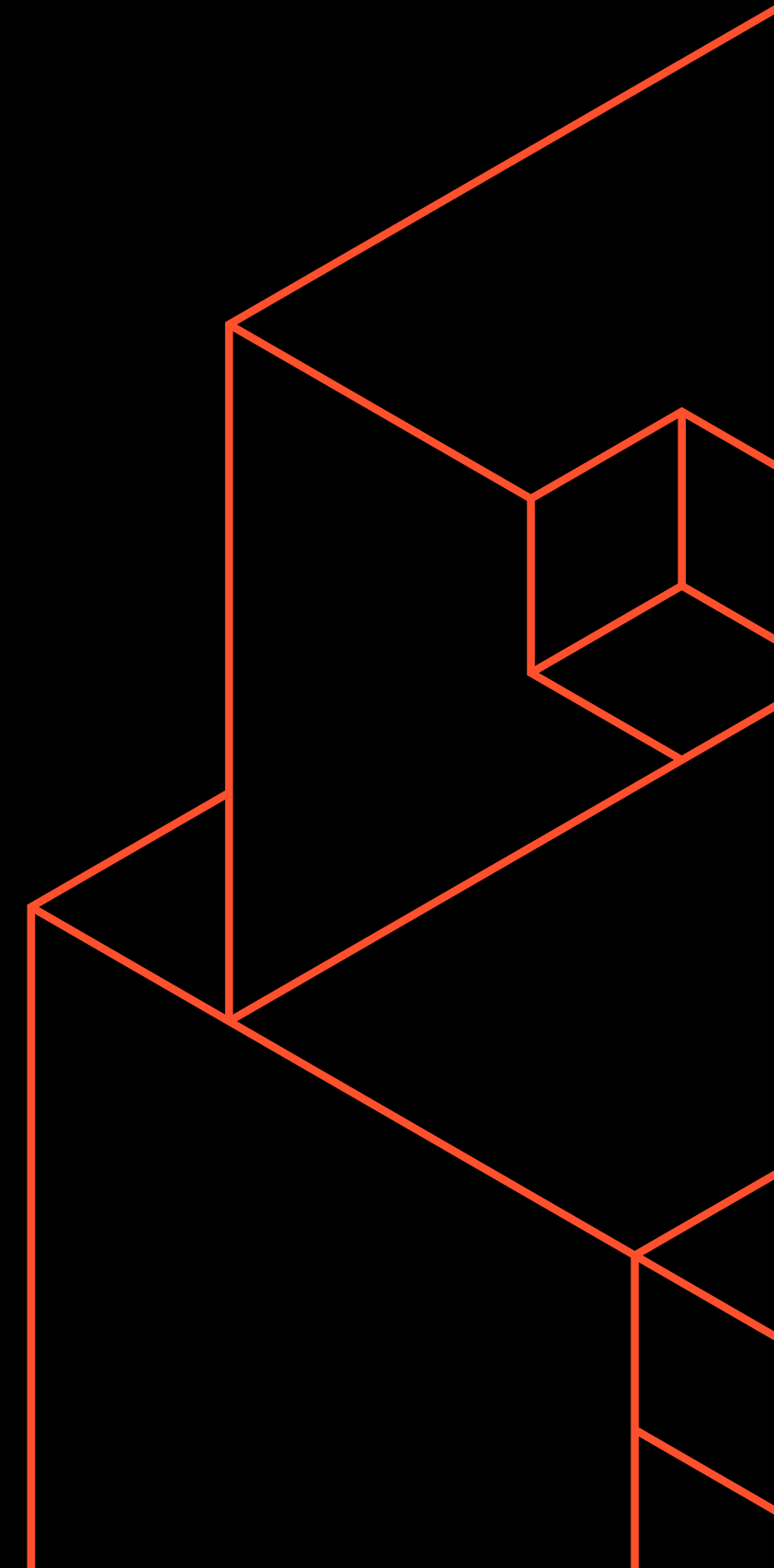
# SENTIMENT MODELING USING NLP

By Patrick Anastasio



# BUSINESS PROBLEM

An investment firm would like a model that can gauge sentiment in emerging technologies to better analyze where to make strategic investments





## **DATA:**

- Tweets regarding new Google and Apple products @ SXSW
- Exhibit either a positive or negative opinion

## **METHOD:**

- Used Natural Language Processing methods to pre-process data and evaluate models
- Analyze & score several models to predict the sentiment of emerging technology



# MODEL SELECTION & SCORING

- Heavily Imbalanced Dataset
  - Trained on a synthetically balanced dataset using the SMOTE process to balance the classes
- Trained models using different text processing and vectorization methods
  - stem words, lemmatization, removing mutual words
  - Count and TF-IDF vectorization
- False Positive - Negative sentiment predicted as Positive
- False Negative - Positive Sentiment predicted as Negative

# MUTUALLY EXCLUSIVE WORDS

- Eliminated words that appear in both classes
- Random Forest scored best - Cross Validation Score = 0.826

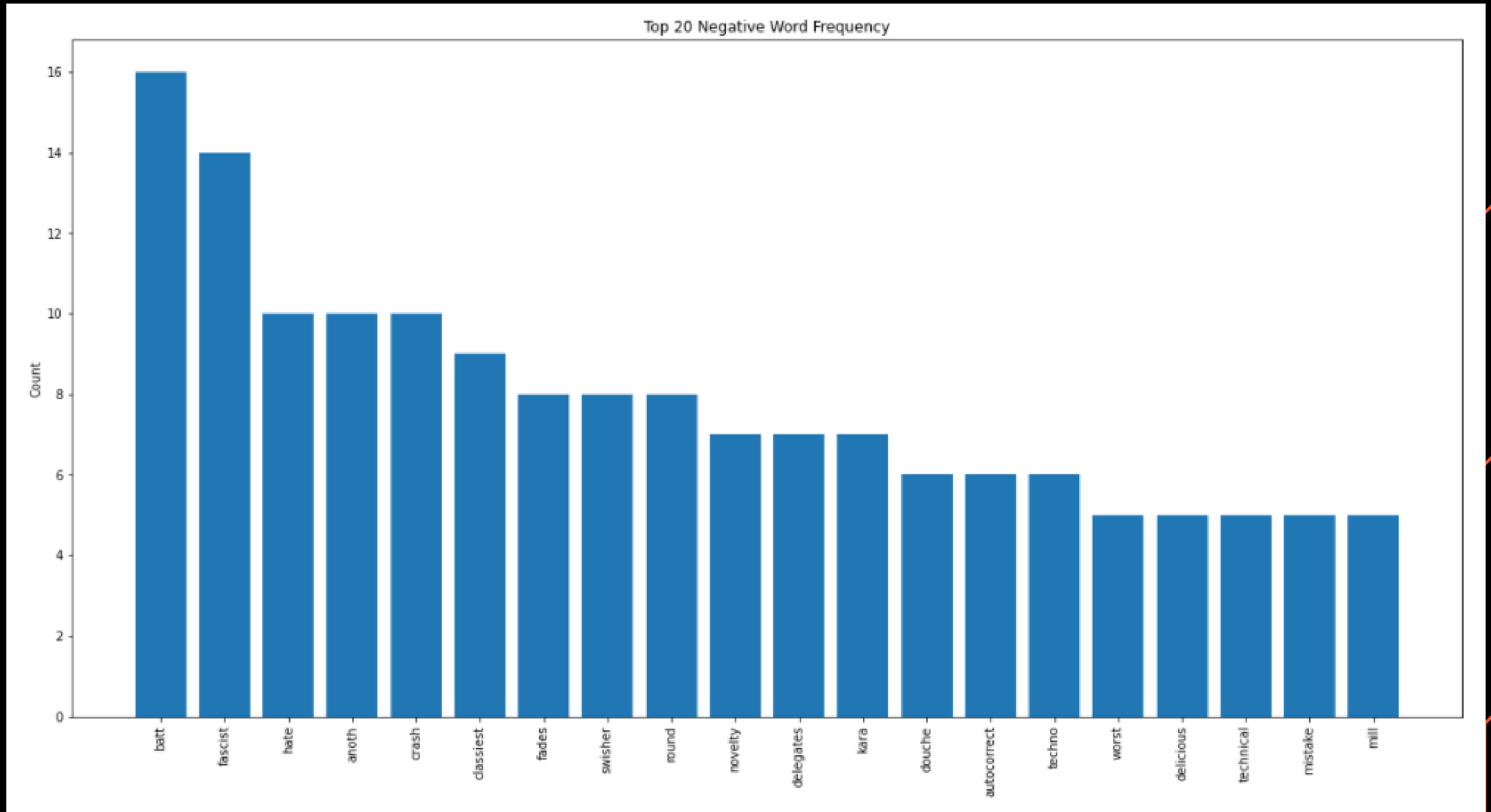
True Class	Predicted Class	
	Negative	Positive
Negative	79	42
Positive	171	391

Overall Prediction Accuracy of 69%

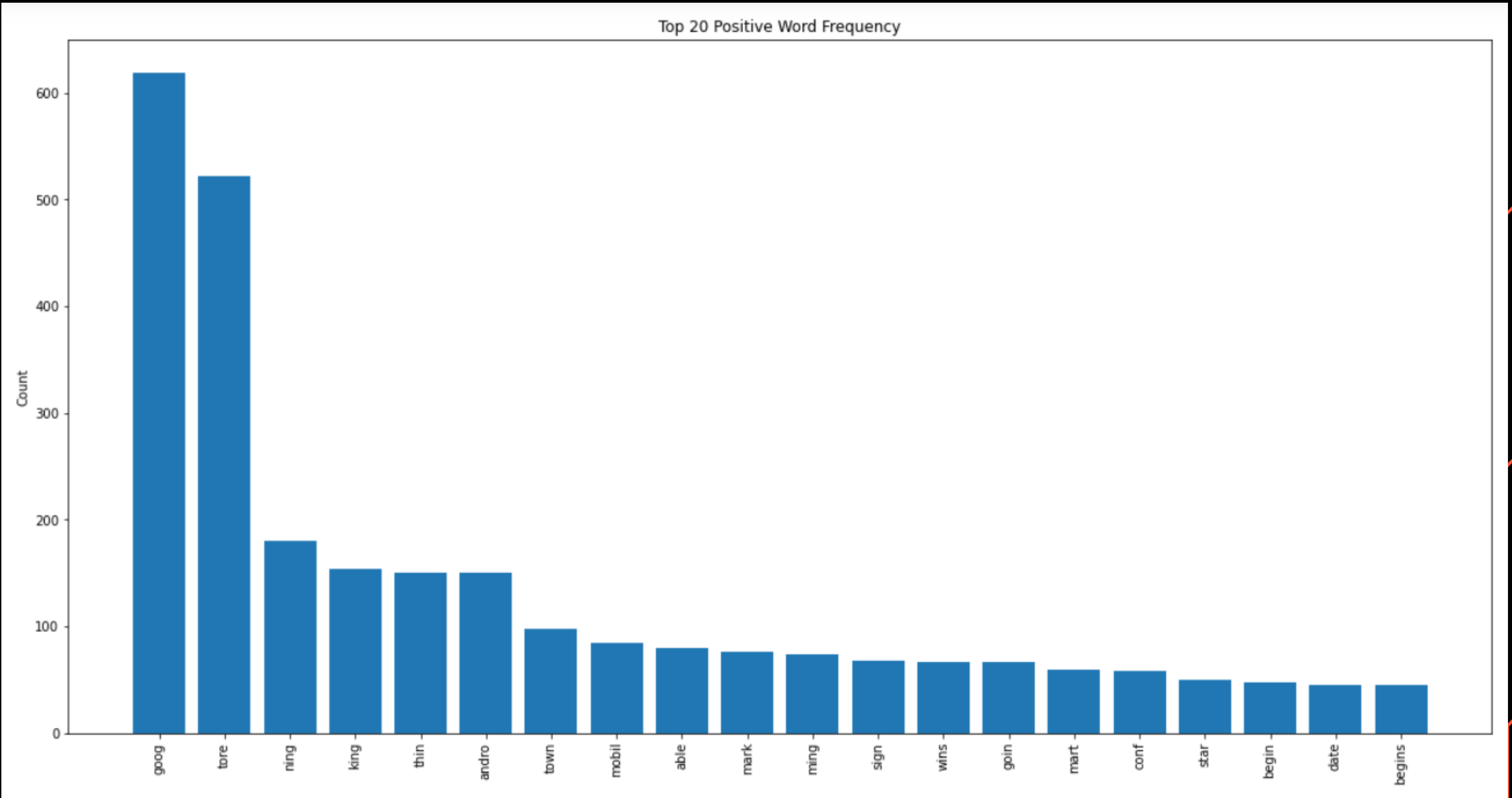
- Positive prediction accuracy of 70%
- Negative prediction accuracy of 65%

	precision	recall	f1-score	support
0	0.32	0.65	0.43	121
1	0.90	0.70	0.79	562
accuracy			0.69	683
macro avg	0.61	0.67	0.61	683
weighted avg	0.80	0.69	0.72	683

# MUTUALLY EXCLUSIVE NEGATIVE WORD FREQUENCY



# MUTUALLY EXCLUSIVE POSITIVE WORD FREQUENCY



# ULTIMATELY WE DID NOT EXCLUDE MUTUAL WORDS

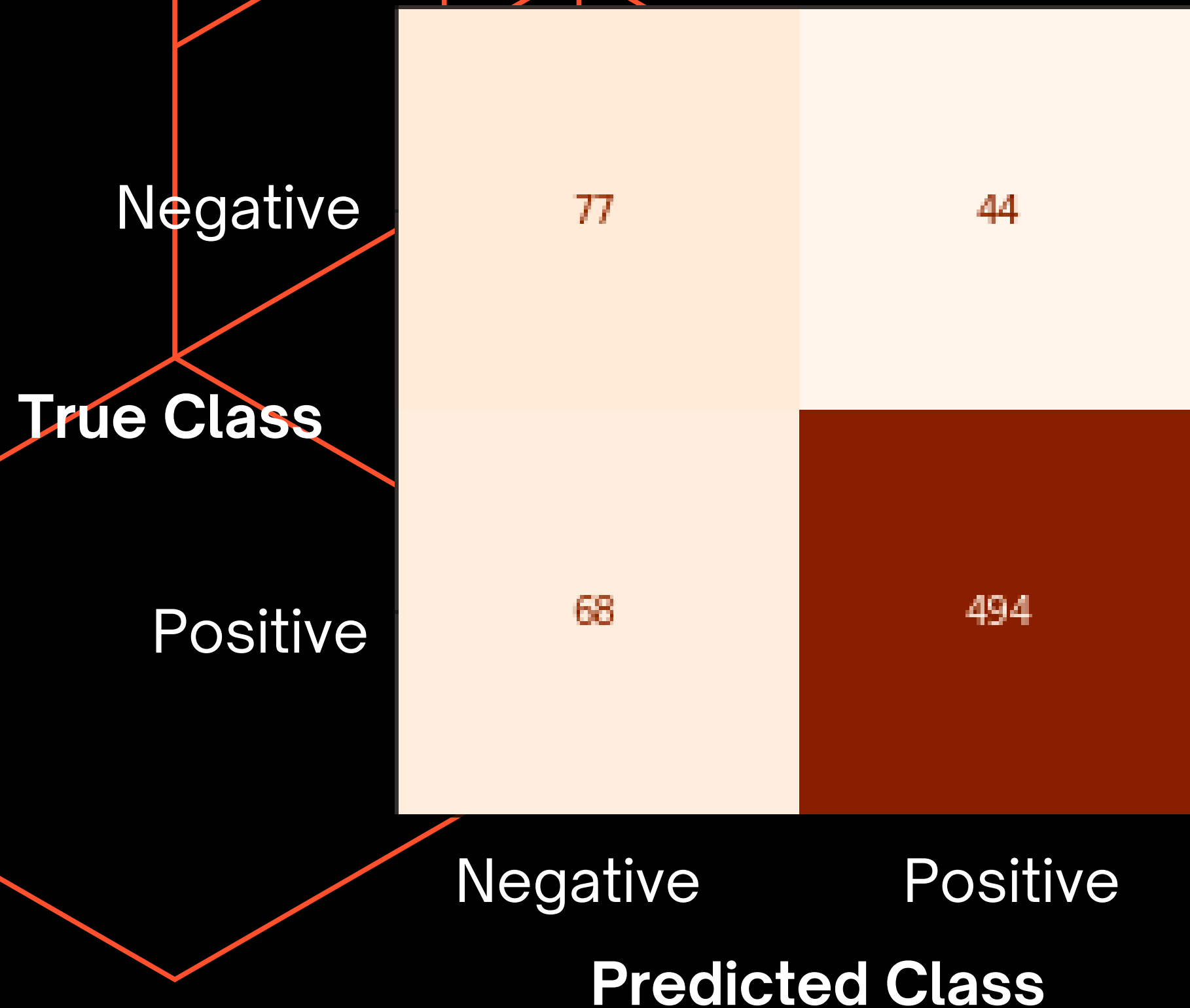
- The models using the lemmatized text before removing mutual words scored a little better on test data.
- Surprisingly the model that predicted the best was our baseline using a default **Multinomial Naive Bayes** model

	precision	recall	f1-score	support
0	0.53	0.64	0.58	121
1	0.92	0.88	0.90	562
accuracy			0.84	683
macro avg	0.72	0.76	0.74	683
weighted avg	0.85	0.84	0.84	683



# MULTINOMIAL NAIVE BAYES

Lemmatized text, Count Vectorization



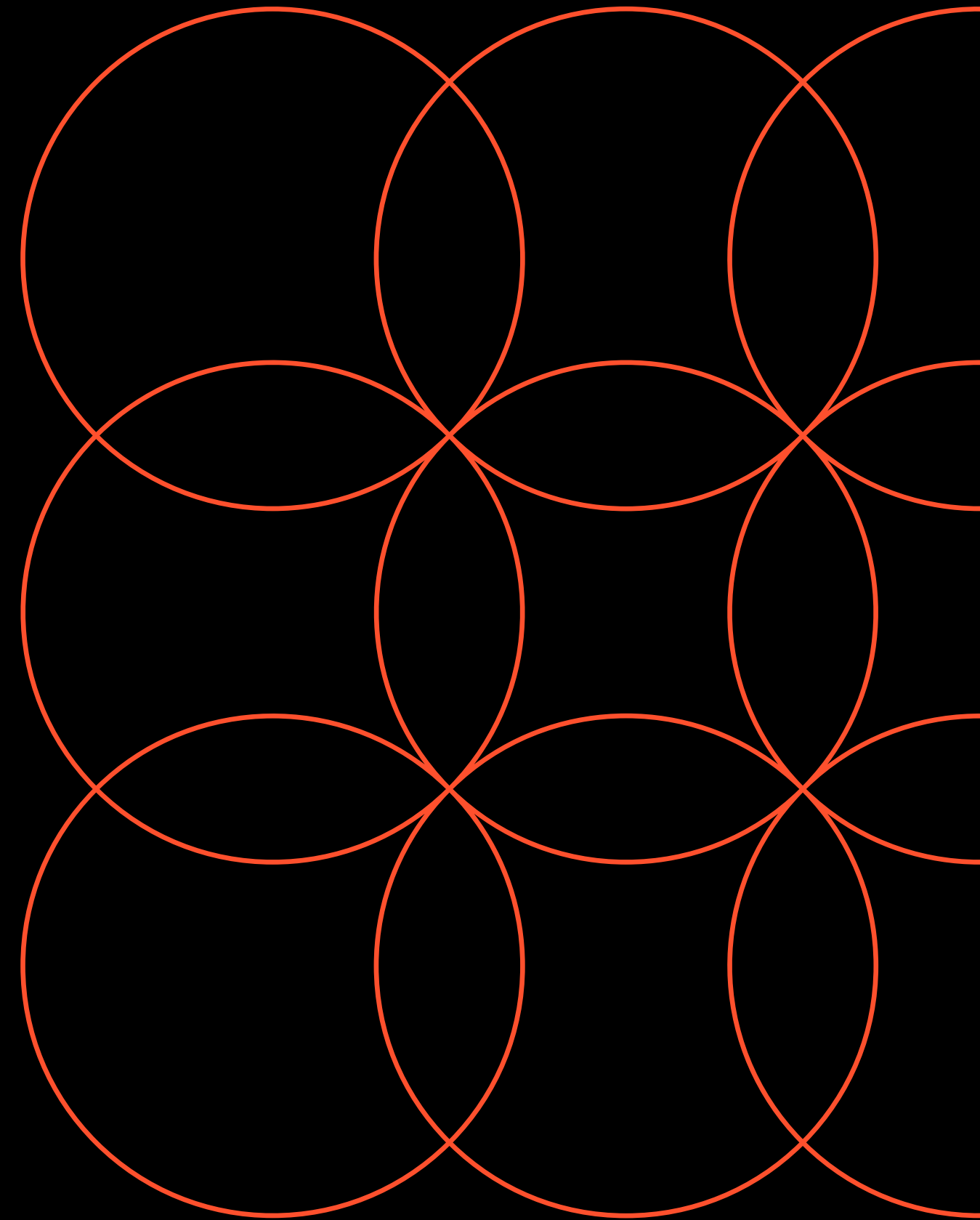
Cross Validation Score: 0.80

Overall Test Accuracy: 0.84

- Positive prediction accuracy of 88%
- Negative prediction accuracy of 64%

# INTERPRETATION

- Imbalanced data posed problems when testing the model
  - Predicted the true positive class much better than the true negative class leading to a higher false positive rate
- Words expressing positive sentiment like "great", and "awesome" appeared in both classes
  - Could lead to false negatives
  - Did not appear in the mutually exclusive word models




# RESULTS

- Based on a general sense of the dataset the overwhelming sentiment was positive.
- This positive sentiment can be used in an overall investment strategy of emerging technologies
- This was a small dataset and more data is needed to improve the model
  - The use of synthetically produced data was used to balance the dataset
  - With more data a downsampling technique could be employed instead



# FURTHER RESEARCH

- Gather more data to provide more balance for testing
  - Explore Neural Networking techniques and Word2Vec for modeling
  - Explore techniques for weighting specific words
  - Look at sentiment on an individual product level
- 
- A series of approximately ten parallel orange lines of varying lengths, originating from the bottom right corner and extending diagonally upwards and to the left across the right side of the slide.

# THANK YOU



PATRICK ANASTASIO

- [SUDOMAKECOFFEE1@GMAIL.COM](mailto:SUDOMAKECOFFEE1@GMAIL.COM)
- [LINKEDIN: /PATRICKANASTASIO](https://www.linkedin.com/company/patrickanastasio/)
- [GITHUB: PATRICK-ANASTASIO](https://github.com/PATRICK-ANASTASIO)
- [MEDIUM: @PATRICKANASTASIO](https://medium.com/@PATRICKANASTASIO)