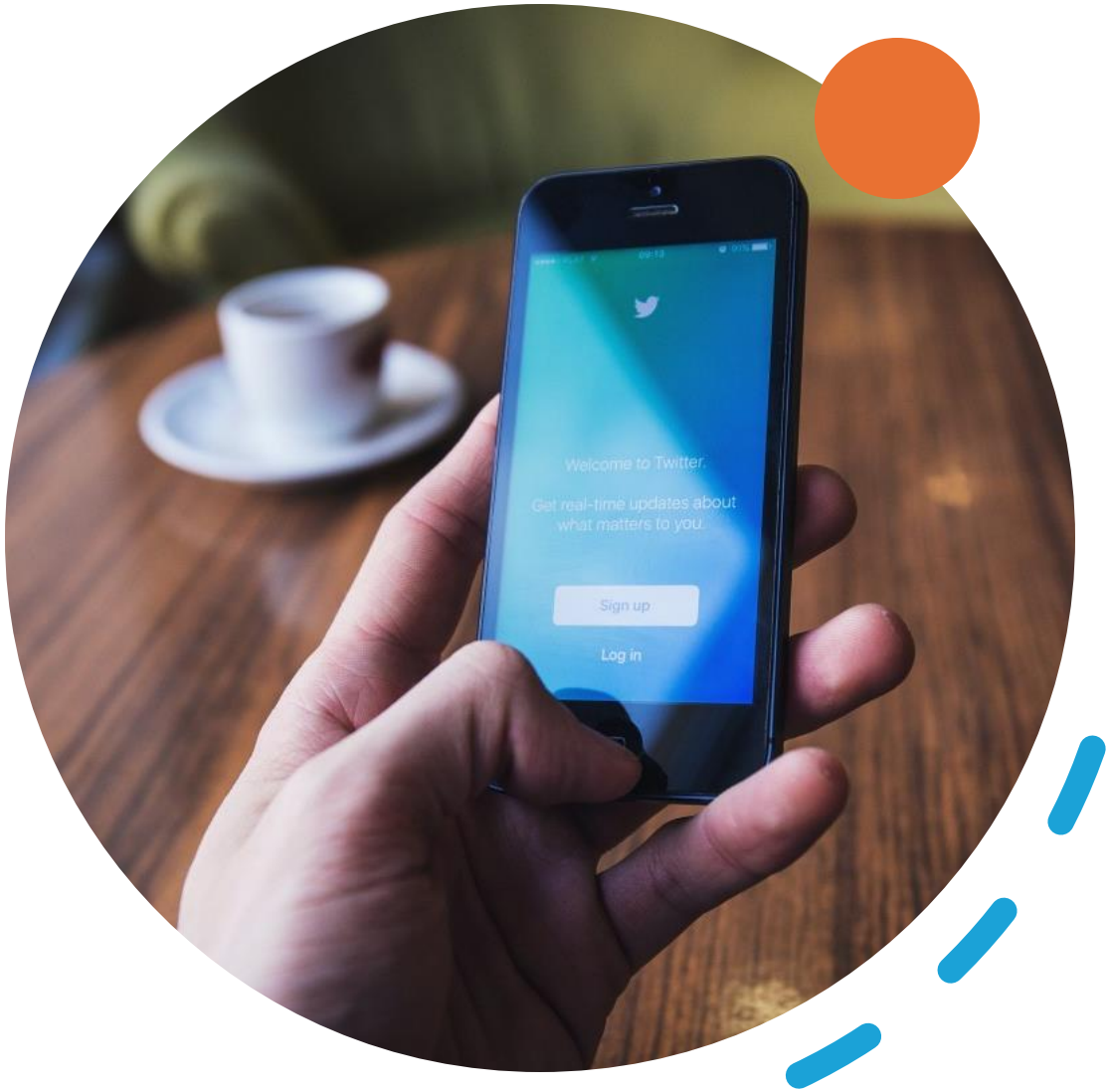
Sentiment Classification Via NLP and Machine Learning

By: Andrew Reusche

Business Problem

Area of concern: Tweet sentiment classification is a time-consuming and an expensive process.

Goal of model: Automate tweet classification to improve research efficiency and and reduce human labor time.



The Data

Contains

- 9,093 Tweets
- 4 Emotional sentiments, converted to a “Positive” binary classifier
- 9 Tech product-related subject matters

Source: "Brands and Product Emotions" Data World, CrowdFlower, 2016, https://data.world/crowdflower/brands-and-product-emotions

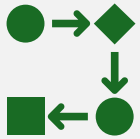
The Data

Limitations

- Unconfirmed classifications
- Short text and limited vocabulary
- Unbalanced “Positive” sentiment
- Narrow subject matter

Data and Model Manipulation Methodology

Step 1: Data Preprocessing



Step 2: Hyperparameter Tuning

Step 3: Run Model and Evaluate

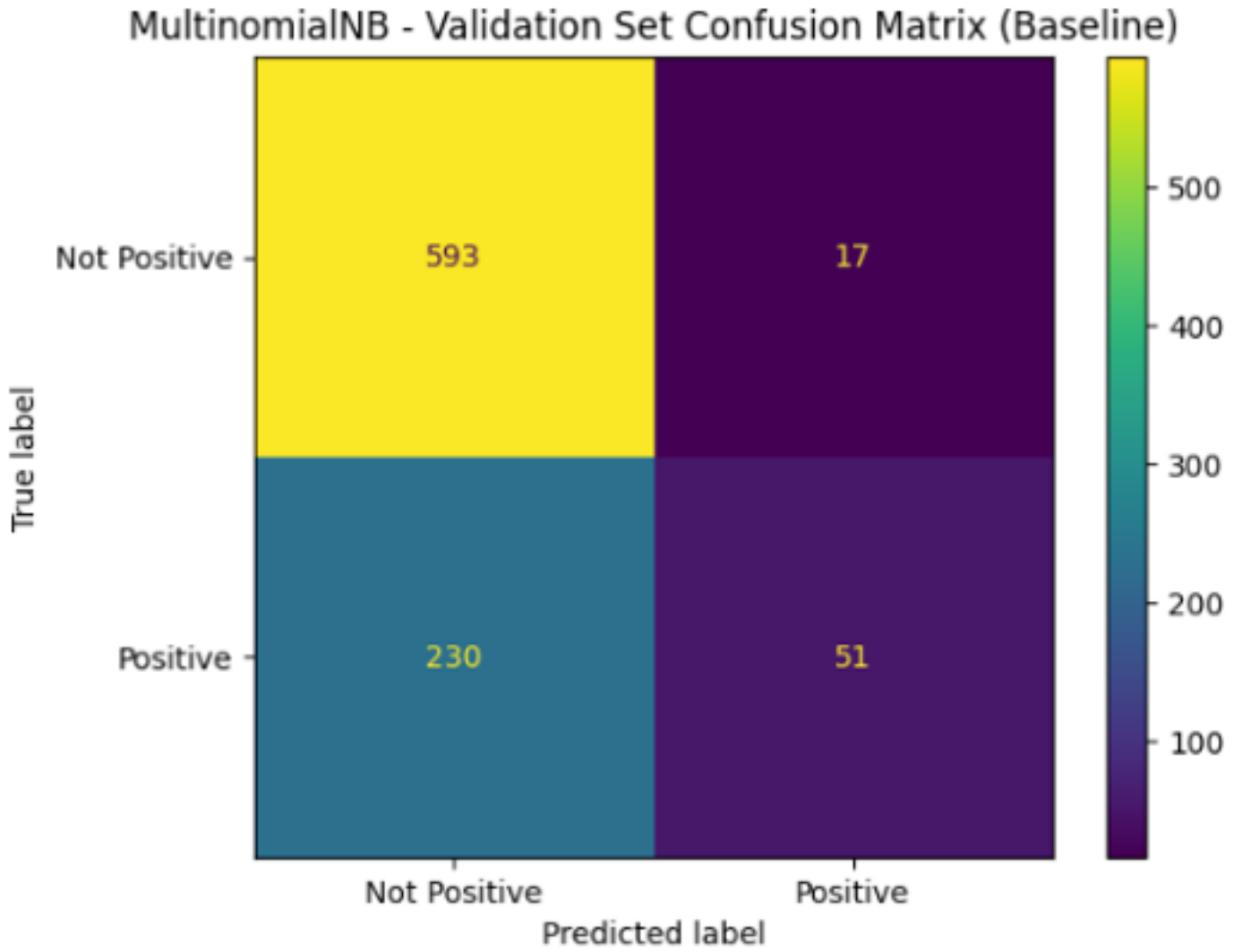
Baseline Model:
Multinomial Naive Bayes

Results on Validation set

Accuracy: 72%

Positive Recall: 18%

Positive Precision: 75%



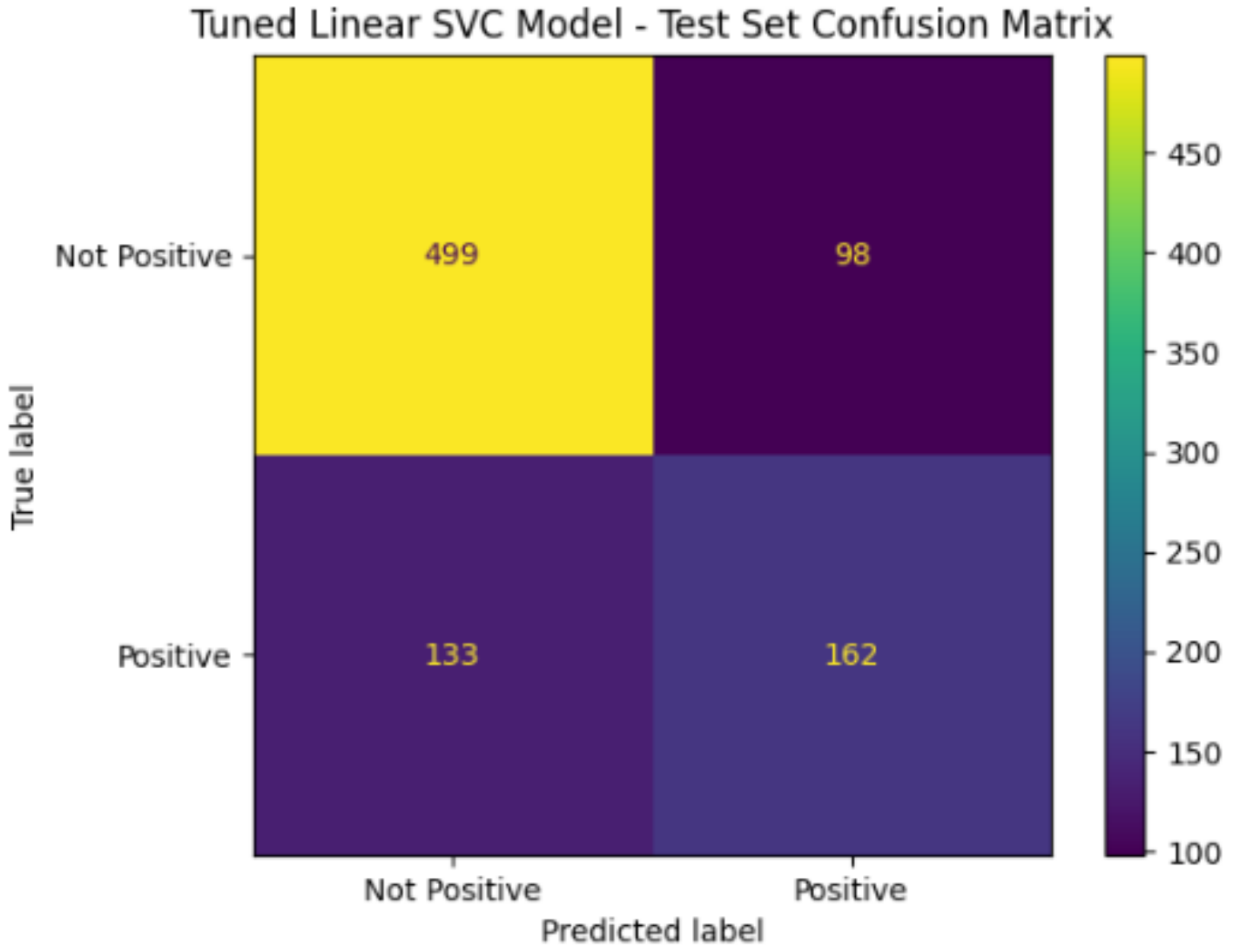
Final Tuned Model:
Linear Support Vector Classification

Results on Test holdout set

Accuracy: 74%

Positive Recall: 55%

Positive Precision: 62%



Conclusions

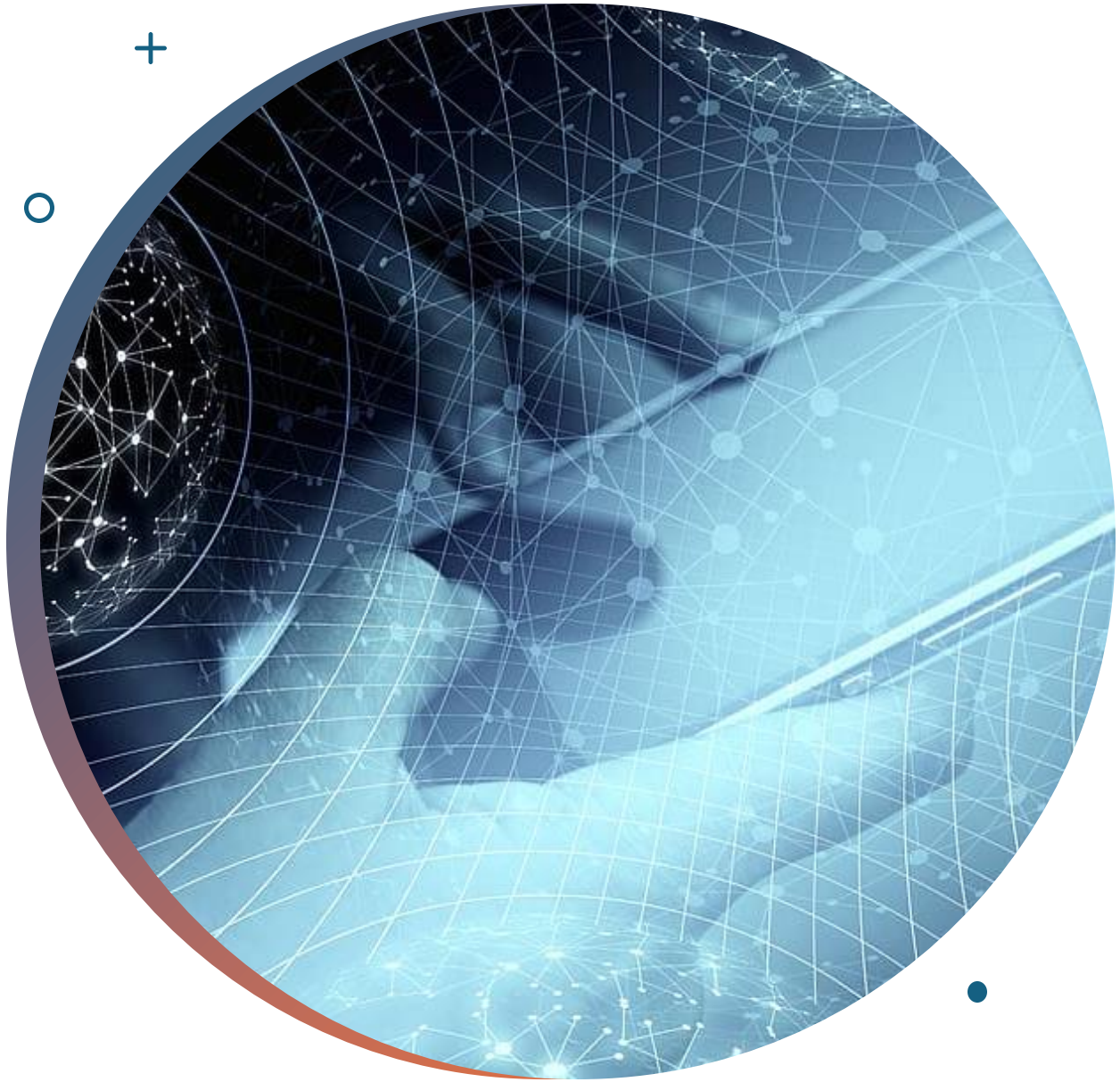
74% of all Tweets are correctly classified

Only 38% of all tweets predicted ”Positive” are falsely classified

55% of the total “Positive” tweets are identified

Business Implications

Tweet sentiment classification can be partially automated, reducing, but not eliminating, the need for manual oversite.



Next Steps?

1. More emotions
2. Live Tweet Predictions
3. Topic Clustering

Thank you

Email:

A.J.Reusche@gmail.com

Github:
https://github.com/AndrewReusche

LinkedIn:
https://www.linkedin.com/in/andrew-reusche-1397bb311/