TWITTER SENTIMENT ANALYSIS

CONTENT

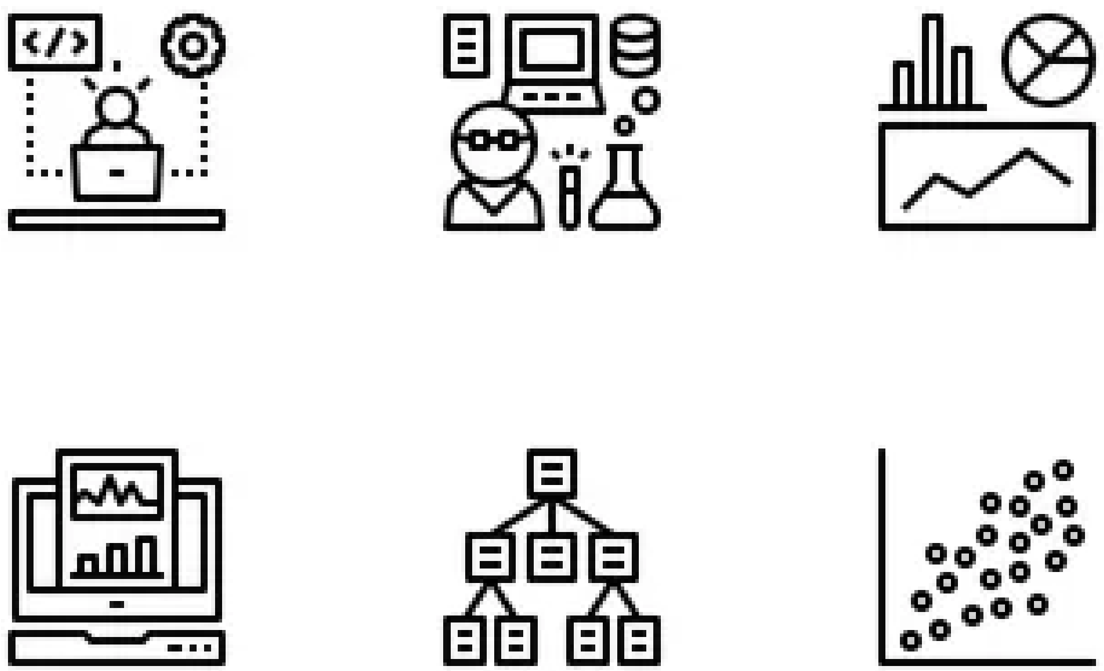
BUSINESS VENTURE

Independent Consultant:
Assist businesses in gauging the overall sentiment of their customer base, providing insights into whether the prevailing sentiment is positive or negative.

BUSINESS PROBLEM/OBJECTIVE

The business problem we aim to address is the need for an automated sentiment analysis solution that accurately classifies tweets as either positive or negative.

TWITTER DATA ANALYSIS



NLP Framework: Start to finish

- Source: CrowdFlower: Tweets over Apple and Google products
- Data frame included 9093 total tweets.
- 2978 positive tweets and 570 negative tweets. (Class imbalance)
- Split data: 80/20, stratify sentiment
- Tweet Cleaning:
  - removed the following
    - Lowercase
    - Punctuation. (., ?, !, etc.)
    - Hashtags
    - Mentions
    - RT
    - URLs
    - sxsw
- Lemmatization
- Text Vectorization and Normalization
- Handled Class imbalance with Random Oversampling of minority class.

Tweet 1 : .@wesley83 I have a 3G iPhone. After 3 hrs tweeting at #RISE_Austin, it was dead! I need to upgrade. Plugin stations at #SXSW. (Negative Emotion)

Tweet 2 : @jessedee Know about @fludapp ? Awesome iPad/iPhone app that you'll likely appreciate for its design. Also, they're giving free Ts at #SXSW (Positive Emotion)

```
[('#', 15875),
 ('sxsw', 9516),
 ('@', 7194),
 ('mention', 7124),
 ('.', 5506),
 ('the', 4424),
 ('link', 4313),
 ('}', 4298),
 ('{', 4296),
 ('to', 3586),
 (',', 3533),
 ('at', 3102),
 ('rt', 2962),
 (';', 2800),
 ('&', 2707),
 ('google', 2595),
 ('for', 2545),
 ('ipad', 2446),
 ('!', 2398),
 ('a', 2312)]
```

→

```
[('ipad', 941),
 ('apple', 775),
 ('google', 673),
 ('iphone', 531),
 ('store', 473),
 ('app', 361),
 ('new', 327),
 ('austin', 245),
 ('popup', 186),
 ('android', 181),
 ('ipad2', 178),
 ('launch', 159),
 ('get', 155),
 ('amp', 155),
 ('one', 138),
 ('time', 132),
 ('like', 129),
 ('line', 128),
 ('social', 125),
 ('circle', 120)]
```

Baseline Models

	Model	CV Train Mean	CV Test Mean	CV Test Std
2	Random Forest	0.9988	0.9740	0.0040
1	Logistic Regression	0.9837	0.9509	0.0076
0	Multinomial Naive Bayes	0.9727	0.9215	0.0095
4	XGBoost	0.9622	0.8969	0.0058
5	K-Nearest Neighbors	0.9102	0.8587	0.0139
3	AdaBoost	0.7803	0.7550	0.0061

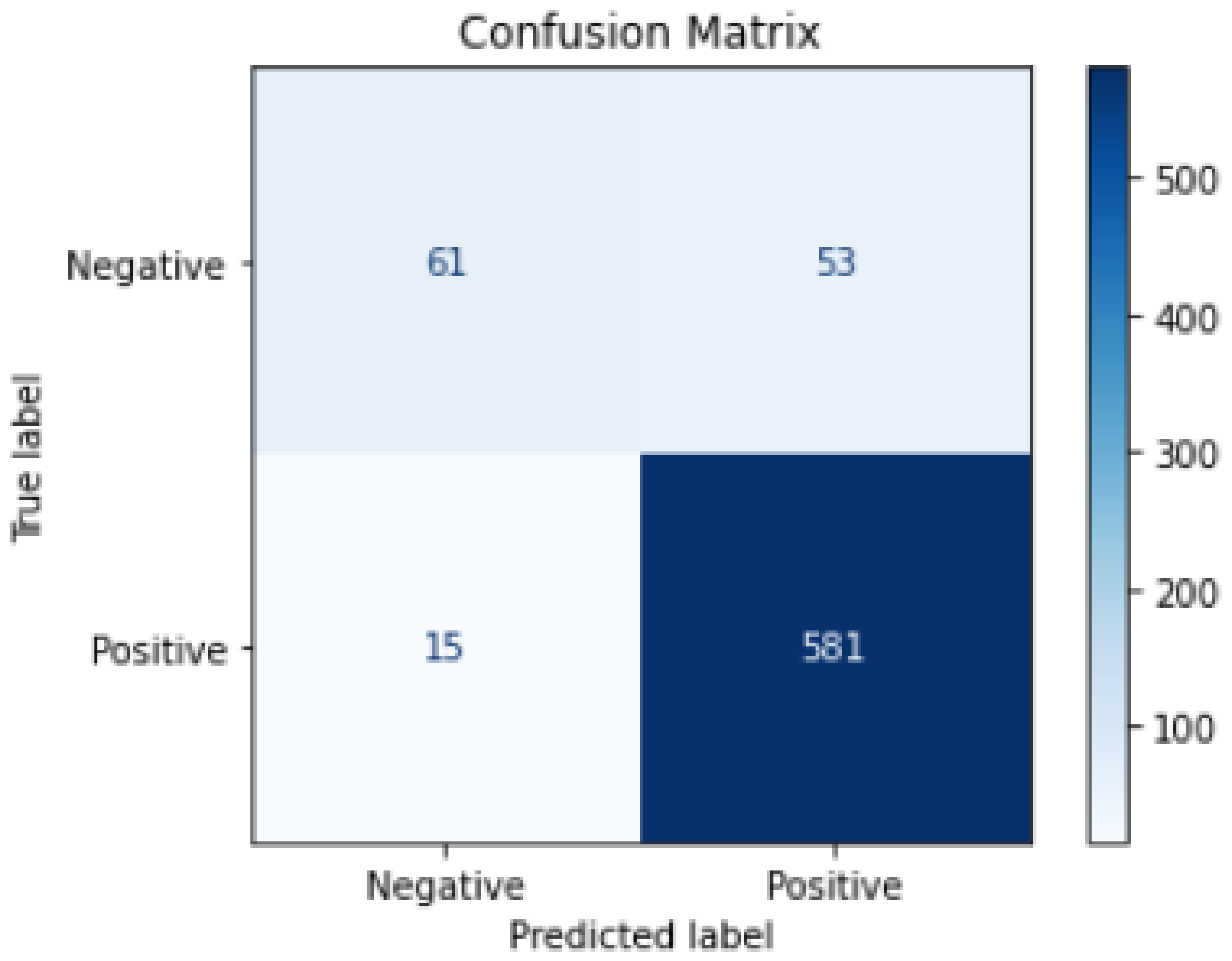
Model Results with Hyperparameter Tuning

	Model	Baseline	GSCV Score	Improvement	Test Accuracy
0	Random Forest	0.9740	0.9866	0.0126	0.8958
1	Logistic Regression	0.9509	0.9700	0.0191	0.8944
2	XGBoost	0.8969	0.9500	0.0531	0.8775
3	Multinomial Naive Bayes	0.9215	0.9421	0.0206	0.8634
4	K-Nearest Neighbors	0.8587	0.9190	0.0603	0.8310
5	AdaBoost	0.7550	0.8772	0.1222	0.8254

Ensemble Modeling

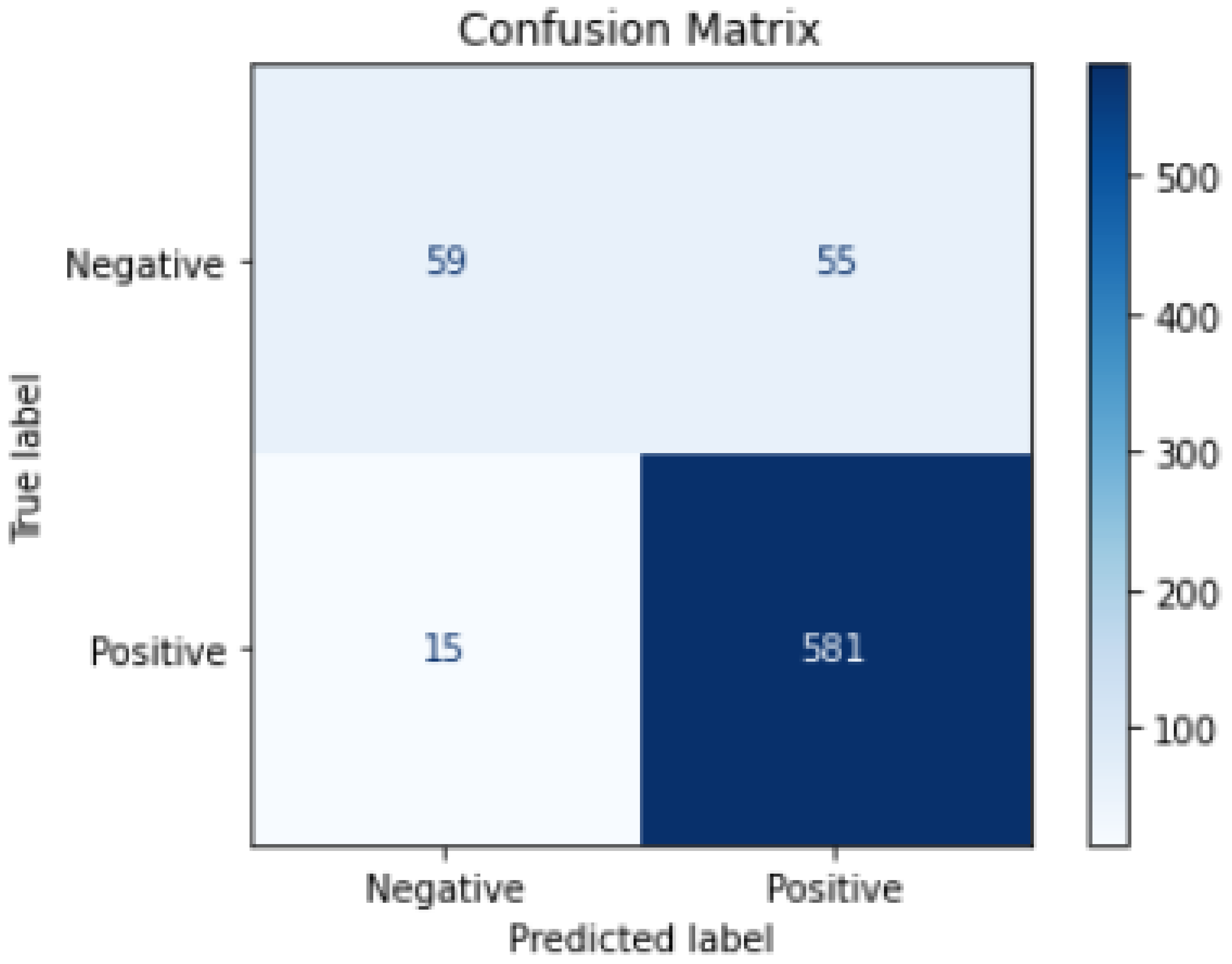
HARD VOTING

```
Best Combination: ['lr', 'nb', 'rf']
Test Accuracy: 0.9042253521126761
```



SOFT VOTING

```
Best Combination: ['nb', 'rf', 'knn', 'xgb']
Test Accuracy: 0.9014084507042254
```



CONCLUSION/NEXT STEPS

- Summary:
  - The hard voting classifier emerged as the optimal choice, but still room for improvement.

- Next Steps:
  - We still have a hard time correctly classifying the negative classes.
    - Recall score of 0.5175 for class 0.
    - Need more Negative sentiment data points.
  - Test Different Data
  - Experiment with a broader range of hyperparameter tuning.
  - Test other models like a neural network.

Contact

- David Johnson : Johnsondavidbjr@gmail.com