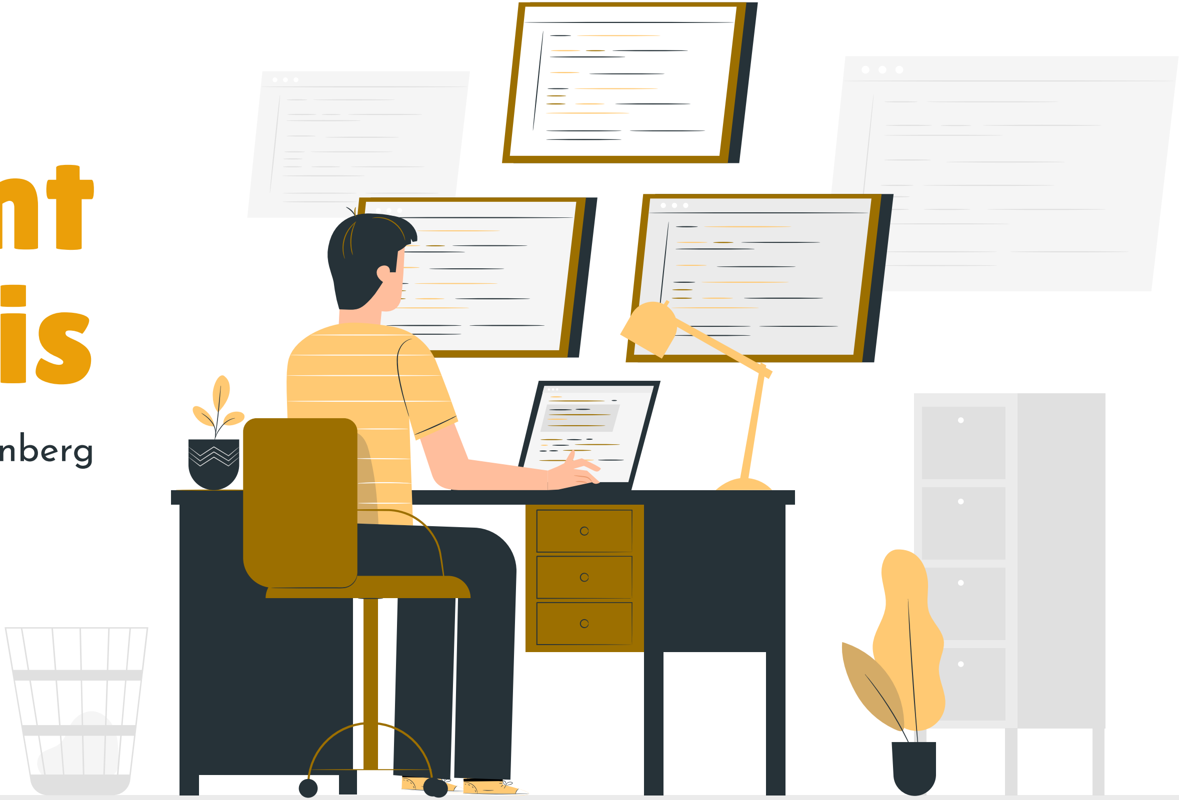
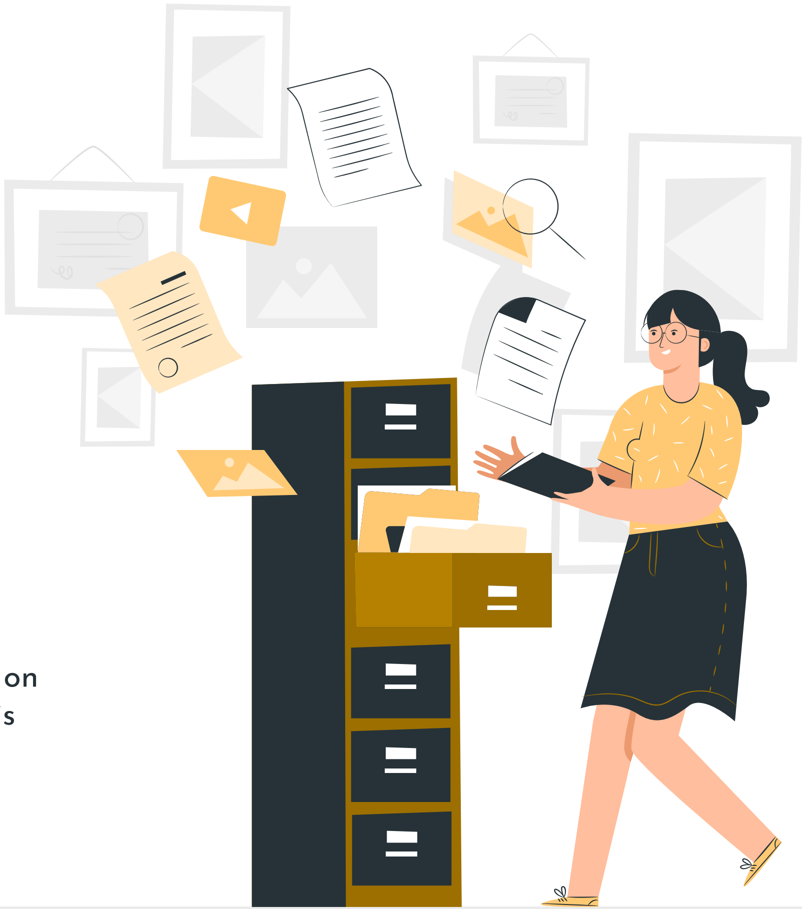
Sentiment Analysis

By Sam Stoltenberg



ABOUT THE DATA

The data used was from data.world, created by crowdflower. It is a crowdsourced dataset based on the sentiment of a tweet, and what the emotion is directed at.



Questions:

01 Which model will perform best?

02 How well will a model perform?

03 What can androids or iphones improve in their products to reduce negative feedback

OSEMN PROCESS

1. Obtain

Obtain the data by either scraping it, pulling from an API or downloading.

2. Scrub

Scrub the data, and transform it to a usable format.

3.Explore

Search for null values, and ways you can interpret the data.

4.Model

Build a model that can predict and interpret unseen data.

5. Interpret

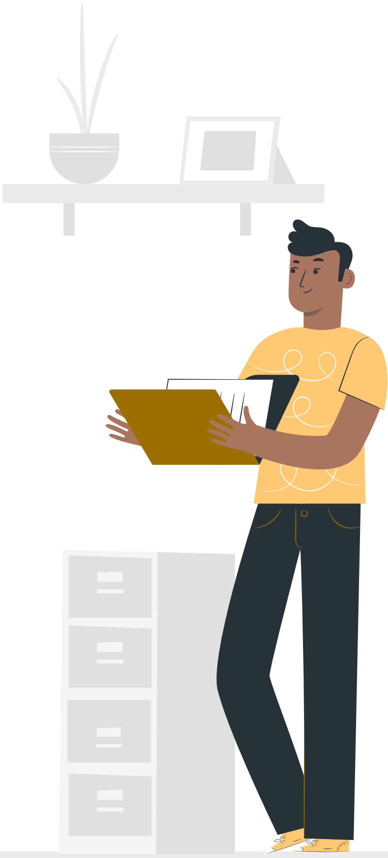
Put the model to use on new data.

Scrubbing the Data:

- There were some null values.
- Some yes/nos needed to be encoded to 0/1
- Imbalanced classes needed to be fixed as only 16% of sentiment was negative.
- The text needed to be vectorized i.e

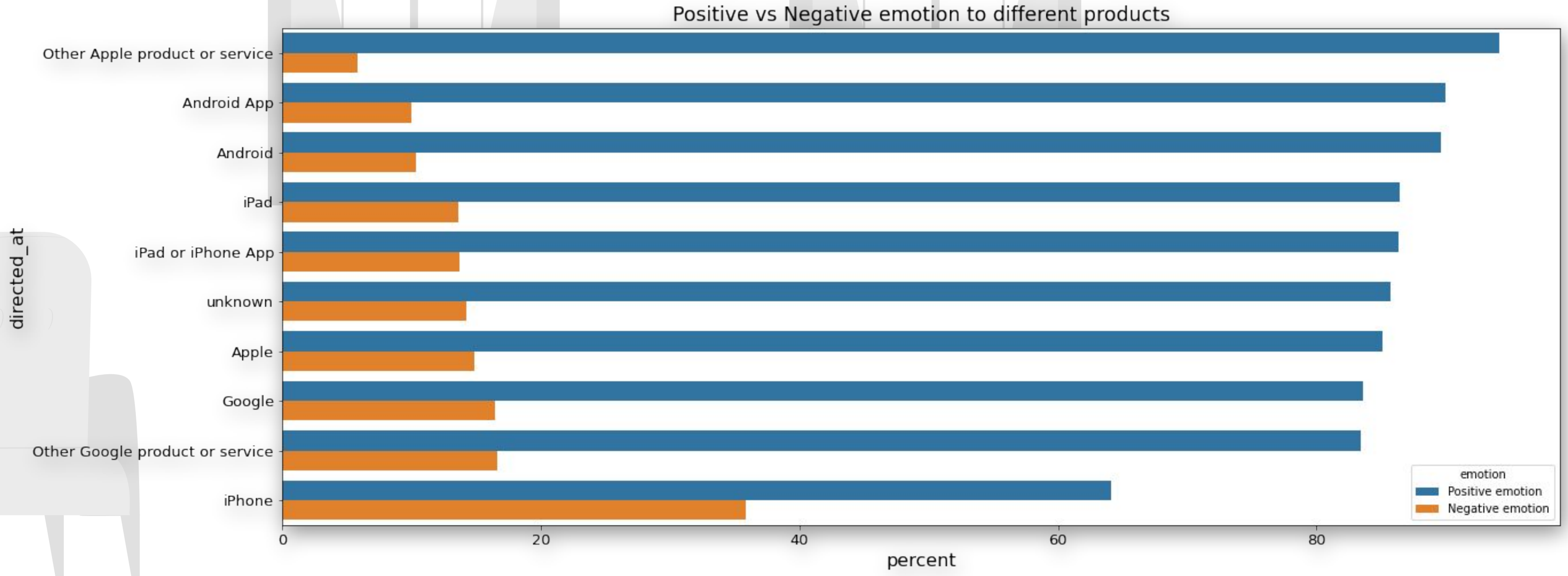
[text, needed, vectorized, i.e]

[4, 18, 32, 13]



Explore

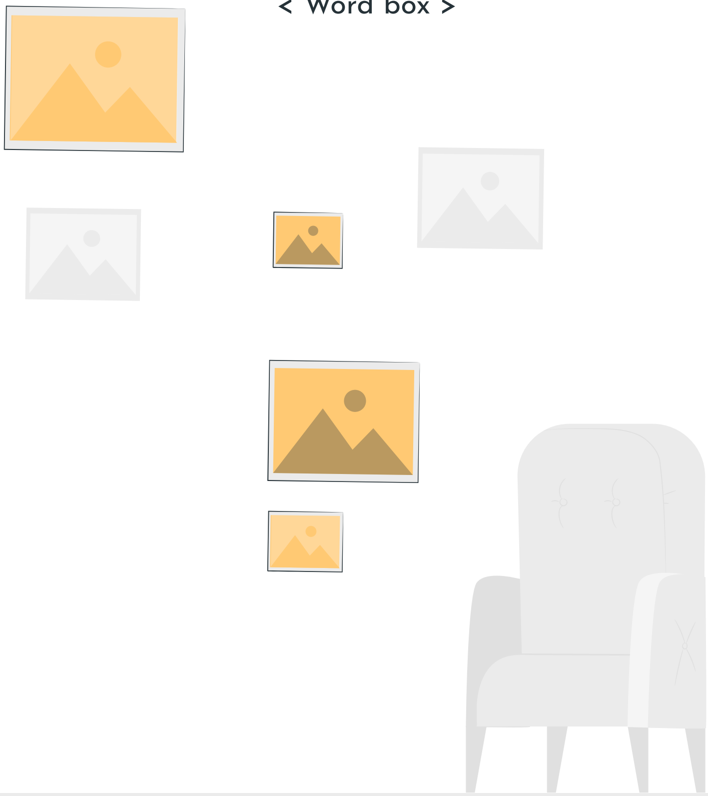
Overall the Iphone has the most negative emotion attached to it, and other Apple product or service has the most positive emotion.



Positive and Negative Word Clouds

Positive

Negative



There should be words here?
< Word box >

Models Used

- True negatives are how often the model successfully predicts negative sentiment.
- True positives are how often the model successfully predicts positive sentiment.
- Accuracy is how well a model performed on test data.

01 Random Forest

- 69% True Negatives
- 79% True Positives
- 77% Accuracy

02 Gradient Boosting

- 64% True Negatives
- 82% True Positives
- 79% Accuracy

03 K-Nearest-Neighbors

- 57% True Negatives
- 80% True Positives
- 77% Accuracy

04 Support Vector Machine

- 66% True Negatives
- 86% True Positives
- 83% Accuracy

05 Adaboost Classifier

- 67% True Negatives
- 74% True Positives
- 73% Accuracy

06 Stacking Classifier

- 56% True Negatives
- 91% True Positives
- 86% Accuracy

Similar Words

```wv.most_similar()```

Below are some selected words from running equations on words i.e (1+2=3)

Iphone minus android:

- stock, interface, easier

It seems that Apple should focus on their interface, and keeping things easier.

Android minus Apple:

- choice, contest, events

It seems that apple may want to focus on events and contests.
```

Conclusion

Best Performing Model

Gradient Boosting for balanced recall, and Stacking for the highest overall accuracy.

How Accurate?

Overall Testing Accuracy

- Gradient Boosting: 76%
- Stacking Classifier: 86%

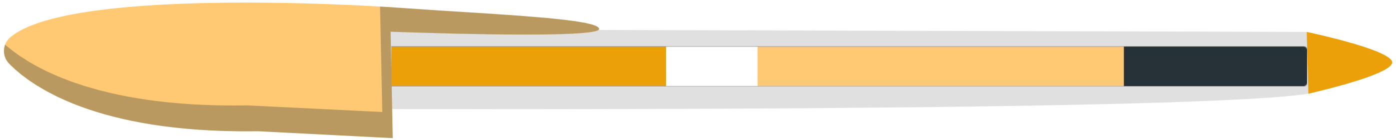
Company Focus?

Android

- Contests
- Events

Iphone

- Interface
- Ease of use



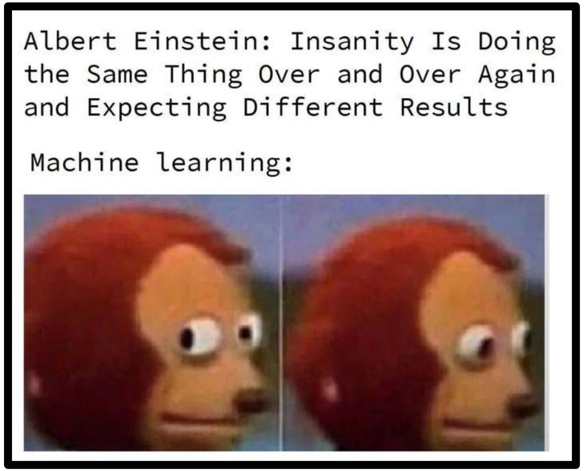
Future Work

- Build an sklearn pipeline and grid search with tokenizer and vectorizer parameters along with classifier parameters.
- Build neural networks for the data, and use the Oscar API for parameter tuning.
- More tuning on the noted models.

THANKS

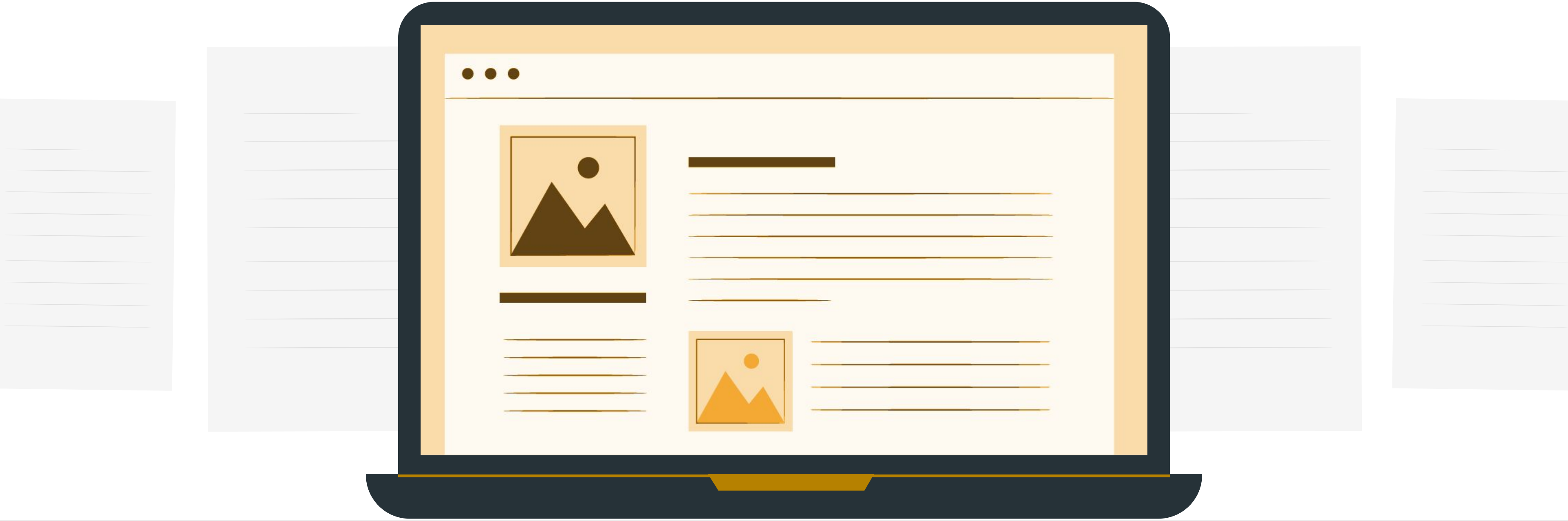
Do you have any questions?

sam@skelouse.com
214-724-6699
skelouse.com

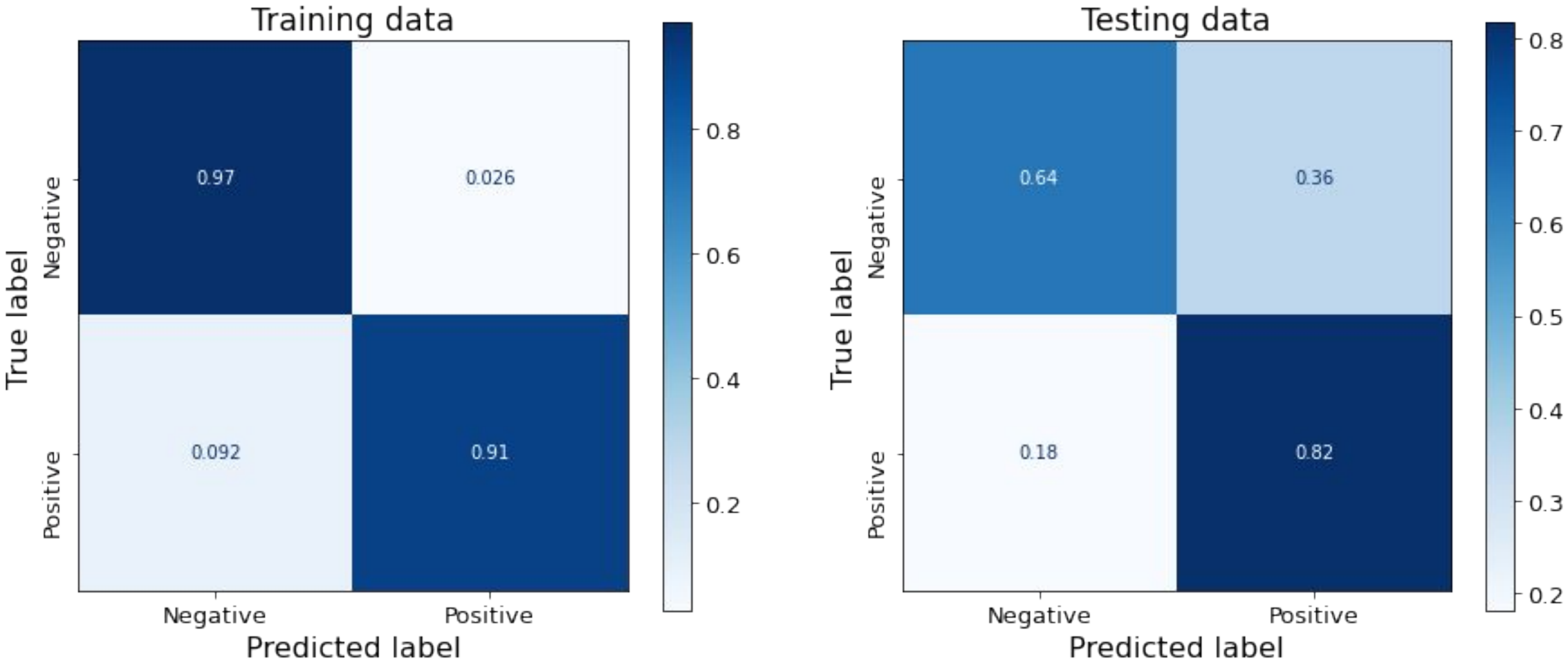


Appendix

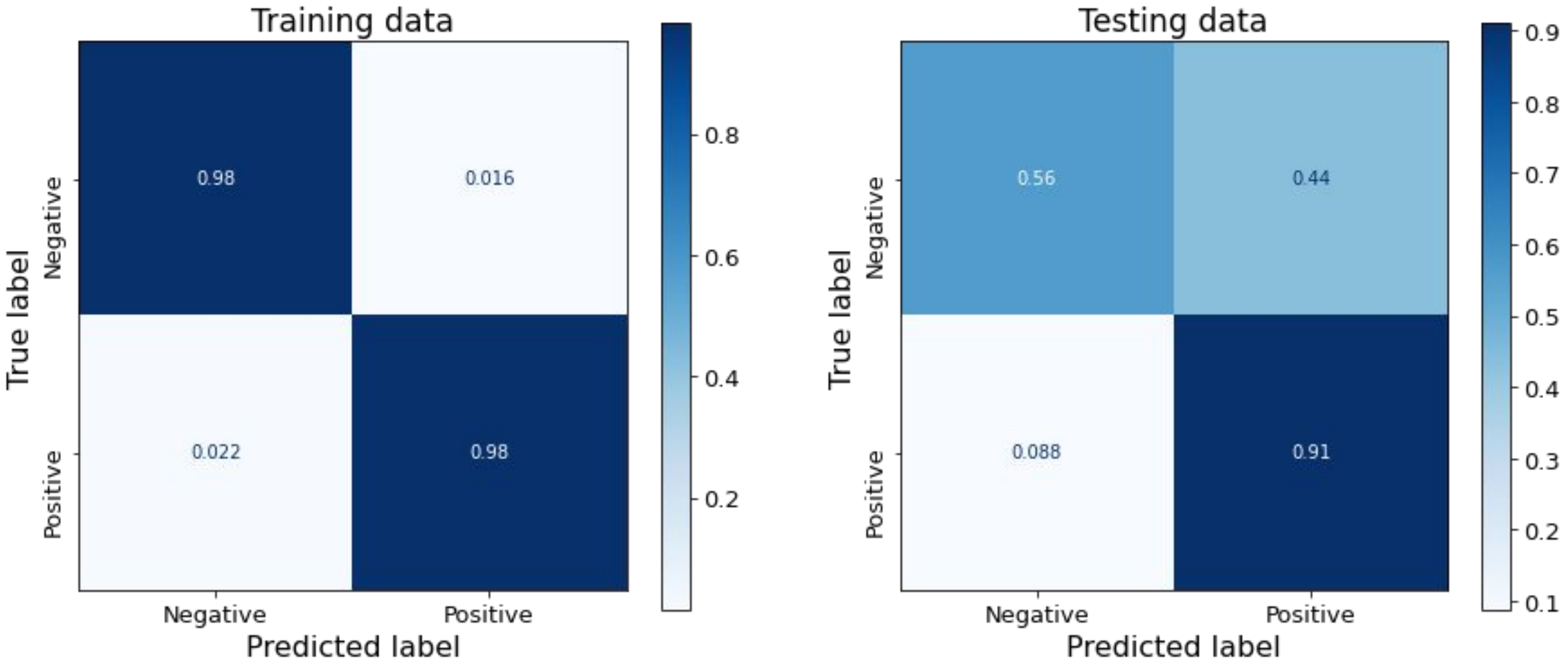
Insert your multimedia content here. You can replace the image in the screen with your own work. Just delete this one and add yours



Gradient Boosting Classifier Confusion Matrices



Stacking Classifier Confusion Matrices



Word Analysis

```
1              0.309063
stock          0.265474
able           0.256434
winner         0.252086
barry          0.238984
going          0.236802
content        0.235509
cbatsxsw       0.231650
entire         0.230150
interface      0.230088
miss           0.226862
make           0.224043
hotel          0.218620
must           0.217708
etchasketch    0.214270
original       0.210932
heck           0.210874
almost         0.207319
easier         0.207113
someone        0.206420
dtype: float64
```

```
xoom           0.300534
working        0.258659
done           0.255140
hootsuite      0.246845
rocks          0.242207
events         0.230156
room           0.227836
woot           0.224581
contest        0.221765
part           0.217028
blocks         0.212759
choice         0.211492
fwd            0.211405
end            0.210556
featured       0.200980
ps             0.198907
walked         0.198528
call           0.197890
blogger        0.195487
excited        0.192184
dtype: float64
```