Twitter Sentiment Analysis:

South By Southwest
(SXSW)

•••

Henry Van Gorp

Outline

- Business Problem
- Data Understanding
- Natural Language Processing
- Modeling Process
- Final Model
- Conclusion and Recommendation

BUSINESS PROBLEM

- South By Southwest (SXSW) is one of the largest festivals in the world. Taking place in Austin, Texas it consists of many exhibitions. Technology based companies attend the event to showcase their brand any new technologies.

- Google and Apple attend South By Southwest. The conference is looking to see how they can utilize twitter data to build a model for these companies to utilize in seeing if a tweet about their company/product is positive, neutral or negative.

- By building a model that can use twitter data from one of the worlds largest technology conferences they will better understand what is looked upon positively and what is looked upon negatively from potential buyers in the future.

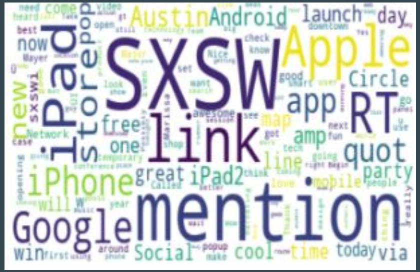
DATA UNDERSTANDING

- The dataset comes from CrowdFlower via data.world.

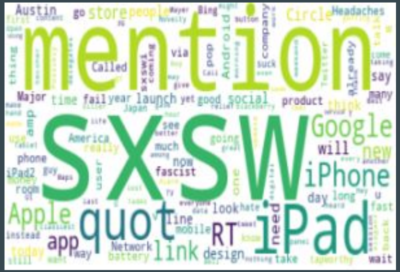
- Human Raters rated the sentiment in over 9,000 Tweets as positive, negative, or neither (neutral). There was a large class imbalance with the dataset:
  - **Neutral Tweets:** 5,545
  - **Positive Tweets:** 2,978
  - **Negative Tweets:** 570

- The dataset shows that this twitter set was pulled from those who were attending the SXSW conference due to the amount of mentions in the tweets.

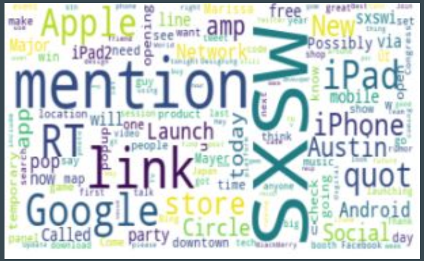
POSITIVE WORD CLOUD



NEGATIVE WORD CLOUD



NEUTRAL WORD CLOUD

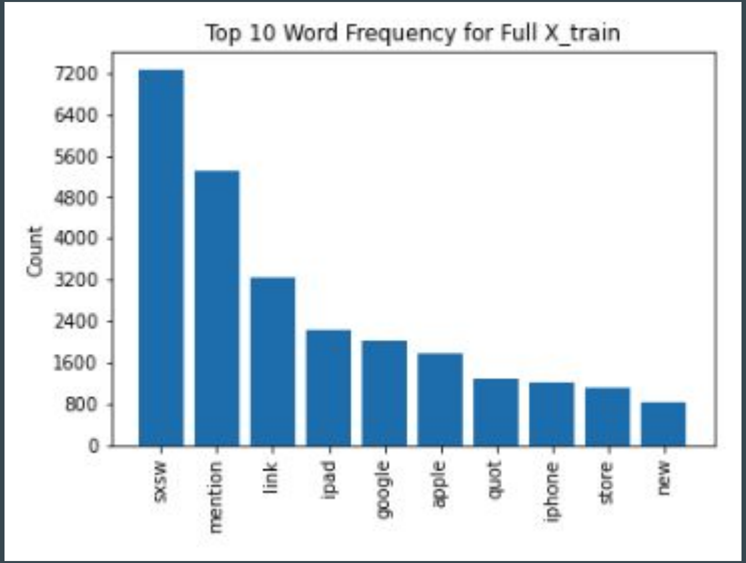


NATURAL LANGUAGE PROCESSING

Natural Language Processing was utilized to clean the twitter data. This process included:

- Standardizing
- Tokenizing

After cleaning the text data, you could look at frequency of words in the full dataset.



MODELING

Three different models were created before choosing the best fit model. Before modeling was done, the data was vectorized using TF-IDF Vectorizer.

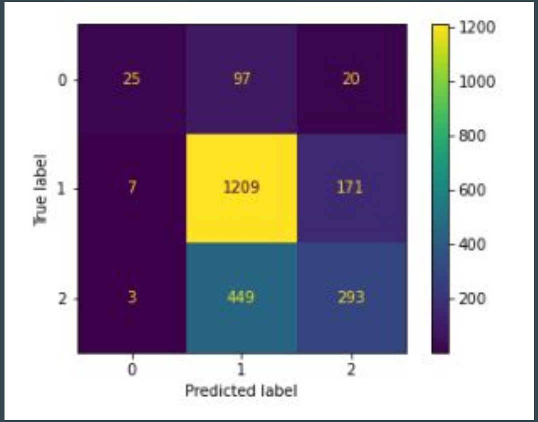
Model Types:

- Baseline Model with Multinomial NB
- Random Forest
- XGBoost

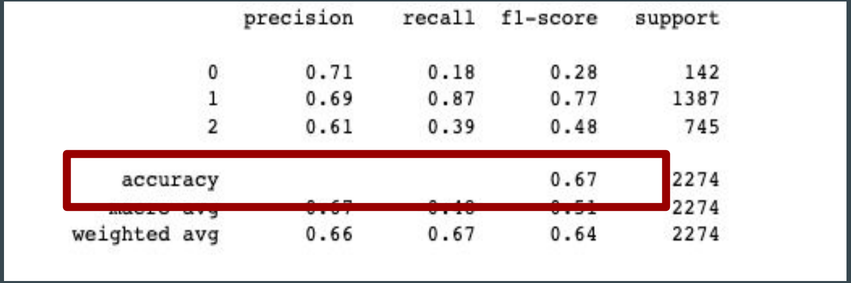
FINAL MODEL

- The best performing model was Random Forest with default parameters.
- A confusion matrix was utilized to showcase largest area of mislabeled tweets.
- A classification report was run on the best model. This model has an accuracy of 67%.

Confusion Matrix For Best Model



Classification Report For Best Model



```
              precision    recall  f1-score   support

           0       0.71      0.18      0.28       142
           1       0.69      0.87      0.77      1387
           2       0.61      0.39      0.48       745

    accuracy                           0.67      2274
   macro avg       0.67      0.48      0.51      2274
weighted avg       0.66      0.67      0.64      2274
```

CONCLUSION & RECOMMENDATIONS

- The best fit model was built with Random Forest.
- This model will accurately predict the emotion of a tweet 67% of the time.
- There was a huge class imbalance causing Class 0 (negative emotion) to perform poorly. Class 1 (neutral emotion) performed the best in the final model.
- It is recommended to utilize this in showcasing another way in which companies can benefit from the South By Southwest conference. Specifically technology based companies.
- It is also recommended to utilize this model with more data to increase the scores.