MARKET SENTIMENT ANALYSIS WITH NLP

SUMAN GAUTAM

X-SIGHT

- X-Sight is an analytical company which provides analytical solutions to the major companies regarding there products, market analysis, sales analysis. Pine-Apple has hired X-Sight to perform a large-scale market sentiment analysis on their products.

- In order to perform such analysis, X-Sight is relying on Machine Learning to predict public sentiment from text data. For this purpose, X-Sight is looking into twitter text data to predict if the given text has positive or negative sentiment towards a particular brand. More importantly, they want to test whether the project is feasible to invest more on large scale data collection and modelling.

METHODS

- Data Acquisition and Exploration

- Data Analysis and Preprocessing
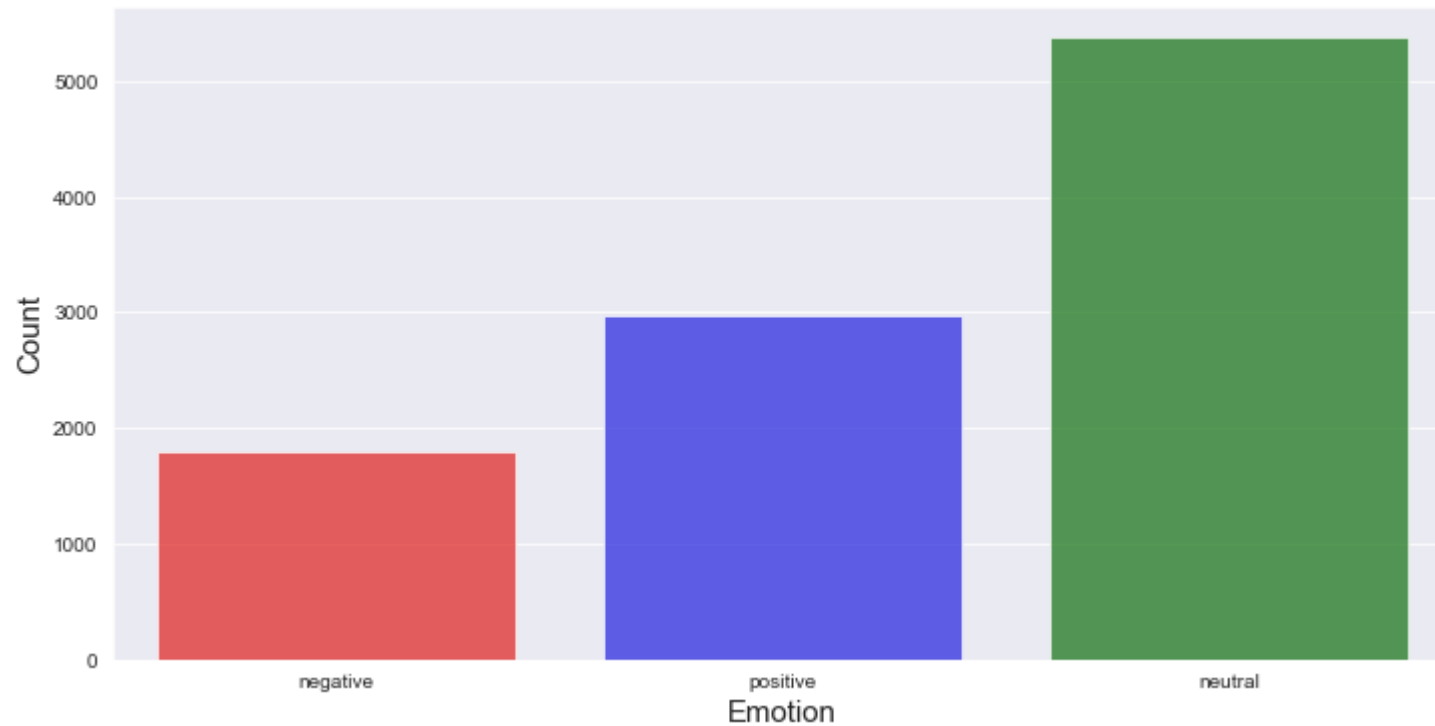
- Model prediction and evaluation

INITIAL DATA EXPLORATION
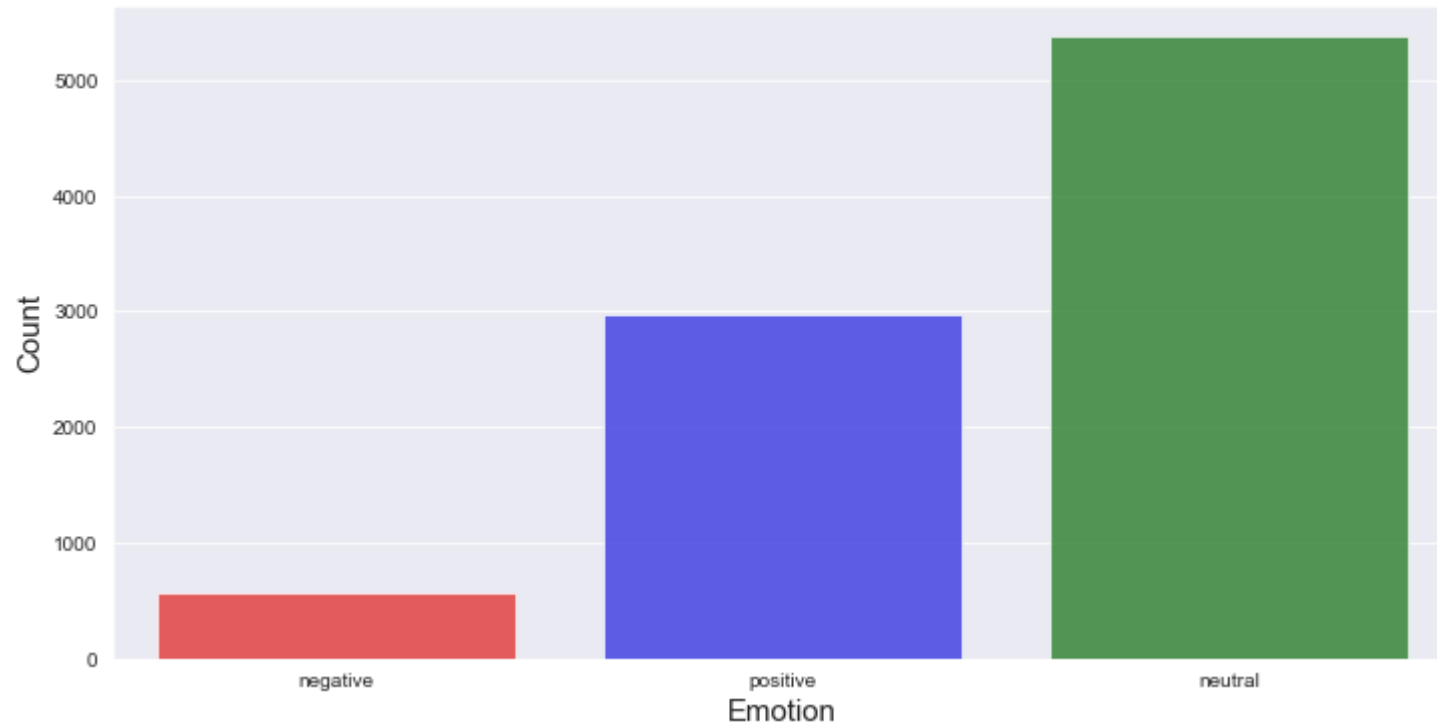
- Twitter Data
- ~9,000 tweets on products of two major brands: Apple and Google
- Class Imbalance!

DATA AUGMENTATION

- Twitter Data

- >1,000 tweets on 'Negative' Sentiment data was added from a different source
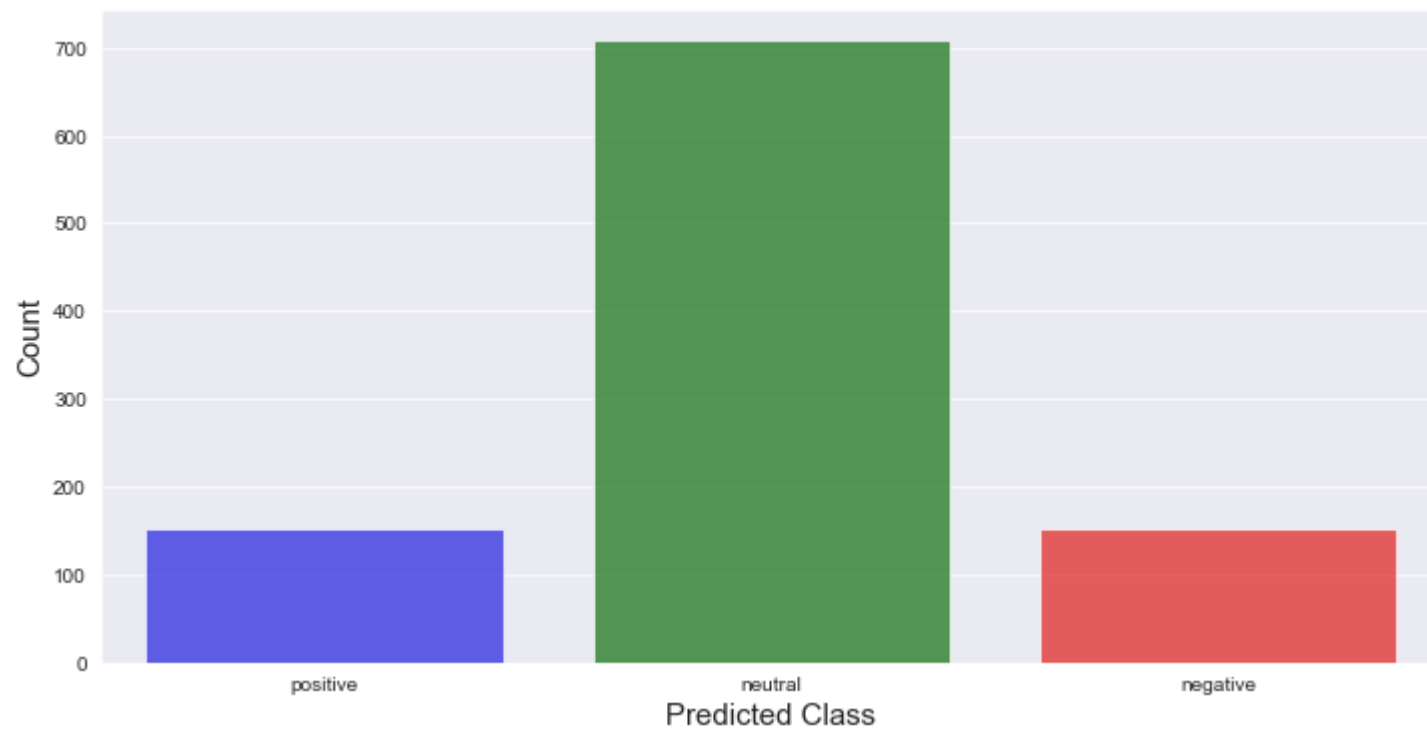
MODELING

- Traditional Models:
  - Naïve Bayes Classification (Baseline)
  - Random Forest Classification
- Neural Networks:
  - LSTM
  - GRU

Model	Accuracy
Naïve Bayes	~65%
GRU	~70%
LSTM	~70%

SENTIMENT CLASSIFICATION

| Tweets	Actual Sentiment	Predicted Sentiment
interrupt your regularly scheduled geek programming with big news google circles	positive	neutral
anybody seen the 6th apple store yet	neutral	neutral
apple the most elegant fascist company america flipboard	negative	negative

PREDICTION STATISTICS

CLASS IMBALANCE

Model GRU	F1 Score Before Augmentation	F1 Score After Augmentation
Positive	56%	49%
Negative	13% (0% in some test!!)	74%
Neutral	67%	78%

MODELING OVERFIT

LSTM
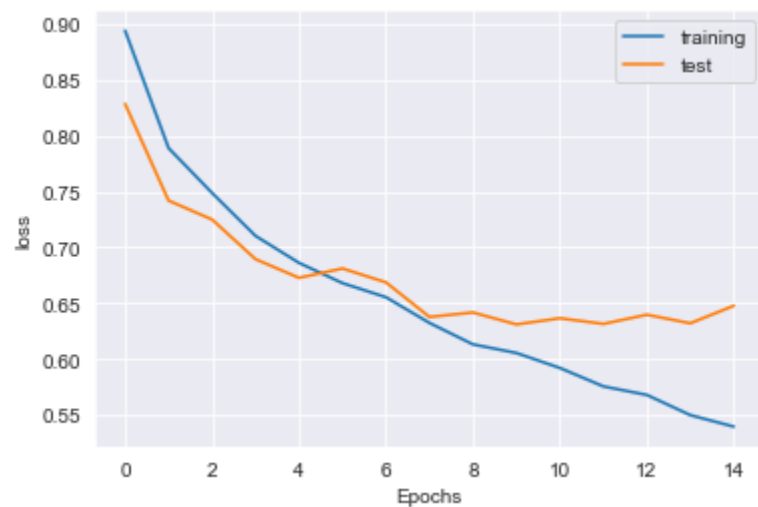
GRU
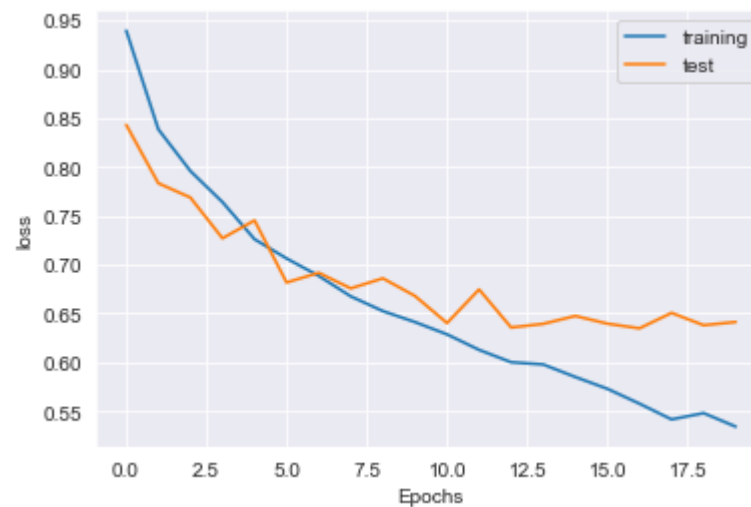
CONCLUSION

- The final model was able to predict a given tweet sentiment on a 70% accuracy level. Given the size of the data this is a relatively good prediction when compared against the 65% if accuracy of the baseline model.

- Current study shows large class imbalance that may bias our prediction. Data augmentation has shown to reduce the adverse affect of imbalance dataset.

- Two neural networks: LSTM and GRU, both produced similar results. Both have similar accuracy and most importantly both displayed significant training overfit.

ACTIONABLE INSIGHTS

- Current modeling shows the impact of class imbalance is significant in prediction accuracy. Therefore, for large scale production, it is recommended have more balanced dataset as much as possible.

- The modeling results have shown that there is a strong training overfit given the smaller size of the dataset. It is recommended to collect more relevant dataset to reduce the training overfit.

- It is also suggested to look into the tweets that has been misclassified, whether it is labeling issue issue or required more text processing. This will be critical for the future modeling improvement.

FUTURE RECOMMENDATION

- Text preprocessing plays a huge impact on modeling accuracies. It is highly recommended to perform more extensive analysis and preprocessing of the text data prior to modeling.

- Some state-of-art architecture such as BERT and GPT-3 may produce more accurate result. Therefore, testing and implementation of such architecture is suggested for further improvement.

QUESTIONS!