Tweet Sentiment Classification

Phase 4 Project by David Boyd

Business Problem/Use Case

90% of data is in unstructured formats, such as text. By being able to analyse sentiment analysis across social media channels brands can better understand whether they need to invest in **brand marketing** or what **pain points** customers might have about their products.

This information is pivotal to maintain companies **CAC** and reducing the chances of **customers churning** due to poor sentiment in either a brand/product.

Data

There are **9,082** records in the dataset, with the distribution being 33% **positive**, 60% **neutral** and 7% **negative** categorised

There are **3** different features which are listed below

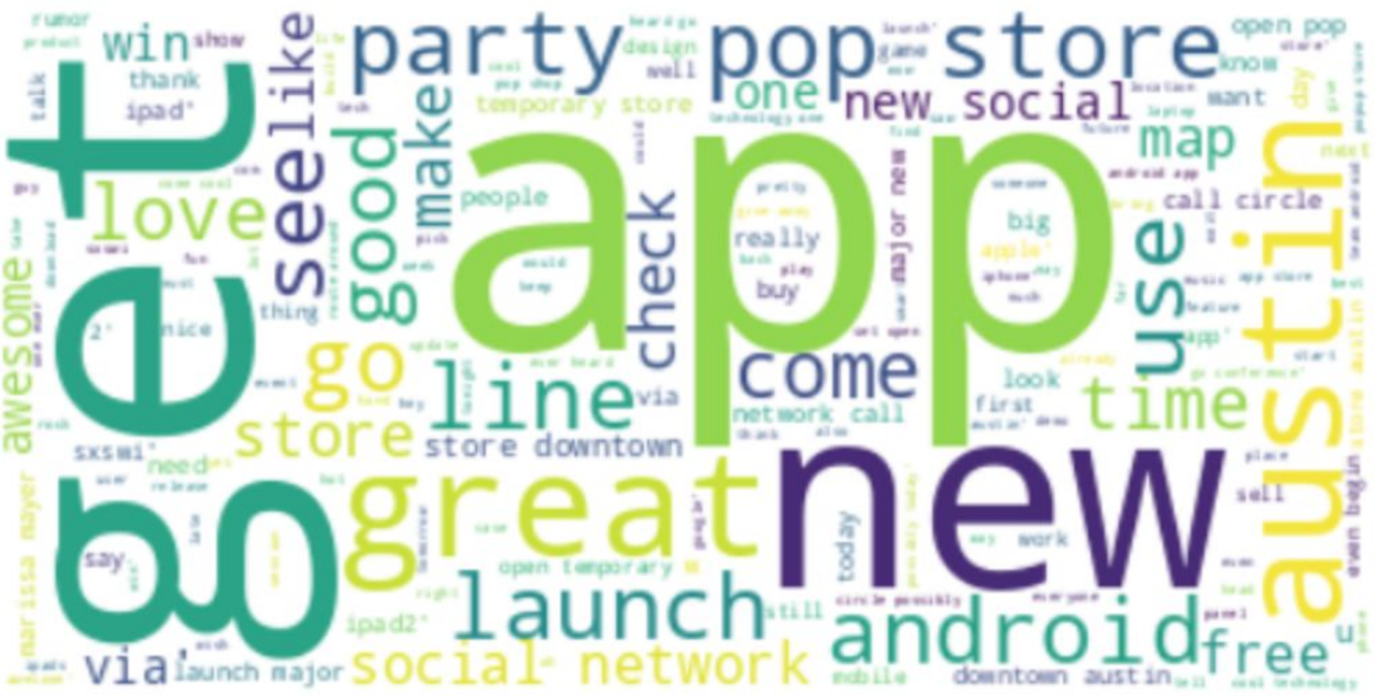
- `tweet_text` - The column contains the contents of the tweet that was captured
- `emotion_in_tweet_is_directed_at` - This column identifies whether the tweet was targeted to a brand or a specific product. There are a lot of NULL values in this column, so it's not super helpful initially
- `is_there_an_emotion_directed_at_a_brand_or_product` - This column helps to identify the sentiment of the tweet

Methodology

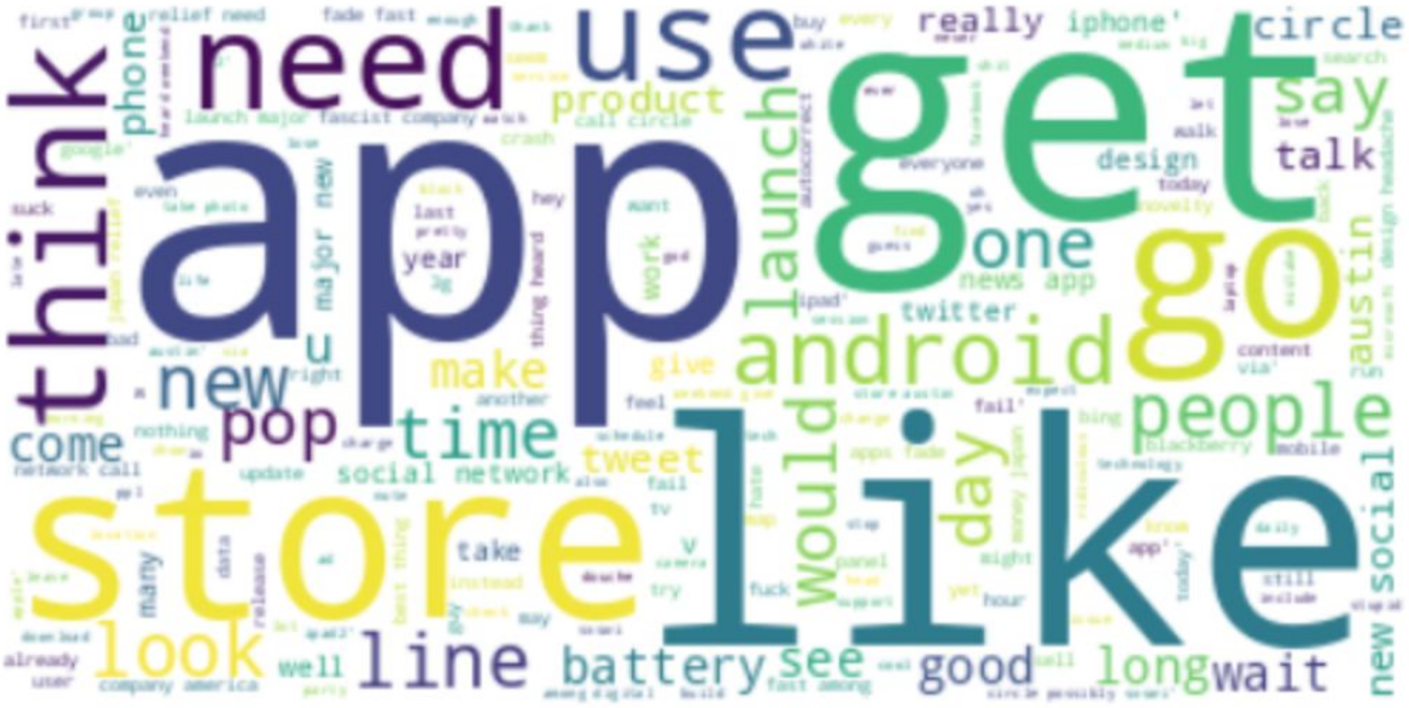
- Did a range of data cleaning to improve the accuracy of model performance
- Created a classification model to predict sentiment classification
  - First started out as a **binary classification** model then moved to **multi-classification** model
- Trialled out several different models to assess performance

Metric used to assess performance was called **Recall** - This is because we want to limit our false negatives as labelling positive sentiment tweets as negative as it reduces the size of the problem companies might be facing when it comes to brand sentiment.

Most frequent words - Positive



Most frequent words - Negative



Model Performance

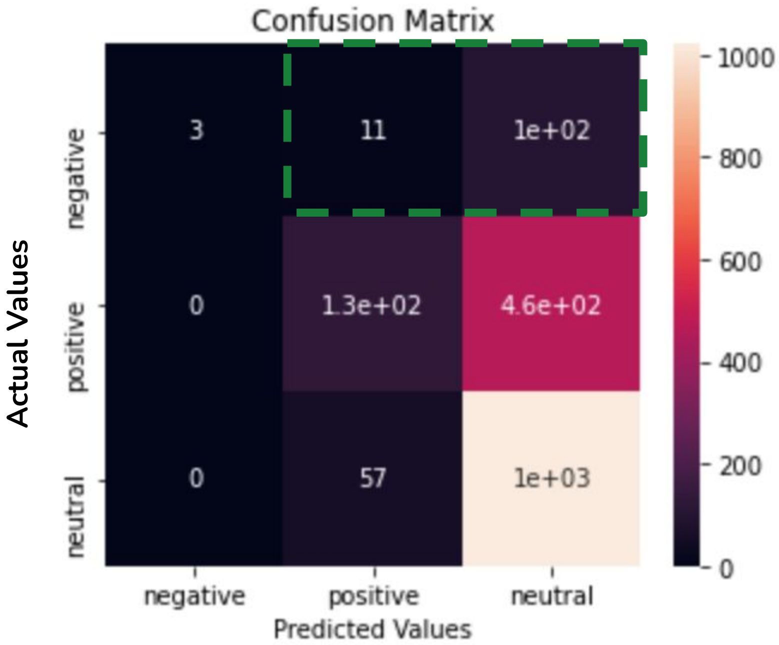
The final model performance drove the following metrics:

- **Recall** - 65%

The two columns to focus more on are the:

- Predicted **positive** but actually **negative**
- Predicted **neutral** but actually **negative**

Both of these categories will hide the impact on a company that has negative tweets and make them seem in a better state than what they actually are



Future Improvements

The main issue faced when performing this classification project was:

- **Imbalanced dataset** when it came to negative sentiment tweets, which meant that the importance was understated in all models, with most optimising towards neutral tweets, then positive and performing poorly to identify negative sentiment tweets.

Other improvements can be to look at re-building the model using more advanced neural network techniques once the data imbalance problem has been resolved. Some options are LSTM, LDA, GRU, CNN

How can you use this at your business

Below are a few suggestions on how this can be implemented into your company to add value, once the accuracy & recall score is above a specifically defined threshold

- Set up a **monitoring system** that flags when negative sentiment social comments are made for a customer support team to reach out to better understand potential bug issues, these should then be picked up in a bi-monthly report to Product/Tech teams
- Add in a new column into your **user base** to identify whether they are a positive/negative customer to perform analysis on impact **against other key business metrics**.

Any Questions?

Github: https://github.com/db495
LinkedIn:
https://www.linkedin.com/in/david-boyd-16245ba1/