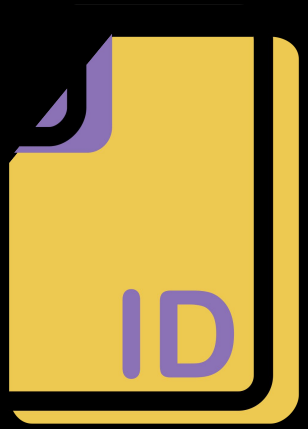
Sentiment Analysis of SXSW Tweets

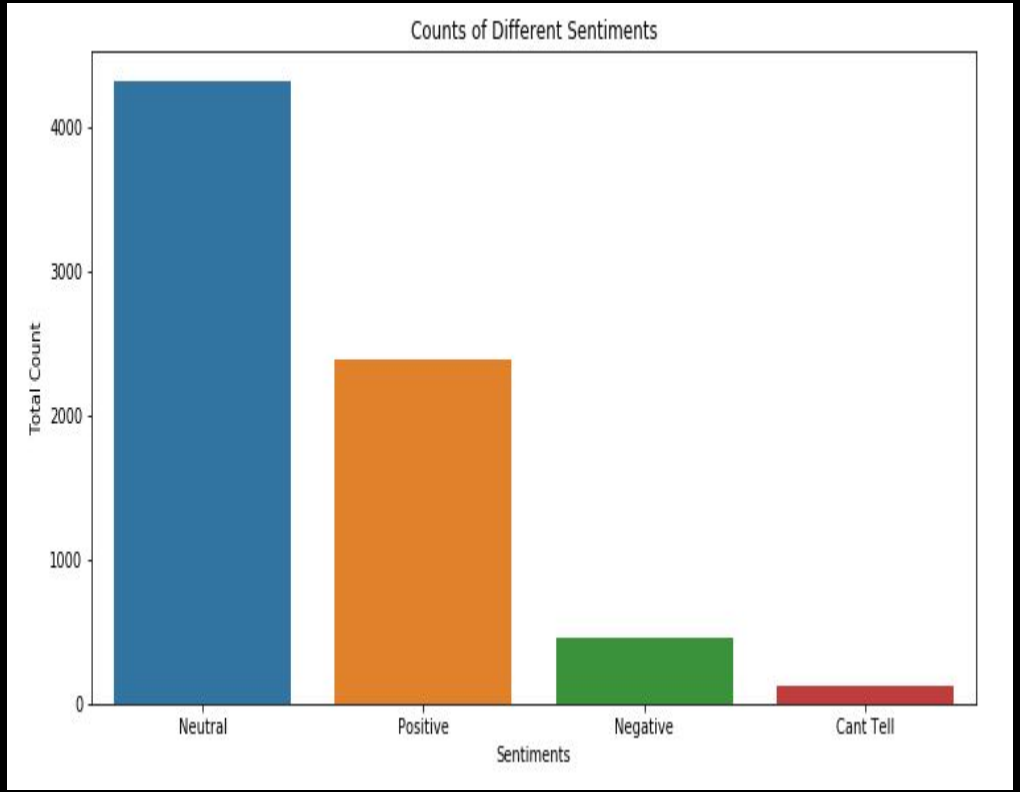
- SXSW organizes conferences, trade shows, festivals and other events.
- Twitter is the most accessible microblogging platform for end users which makes it popular for organisations to reach the pop culture.
- The dataset contains thousands of evaluated tweets from an SXSW tech event, making it ideal platform to analyse sentiments of the audience and gain insights from a customer perspective.

Features in Raw Data



General Observations

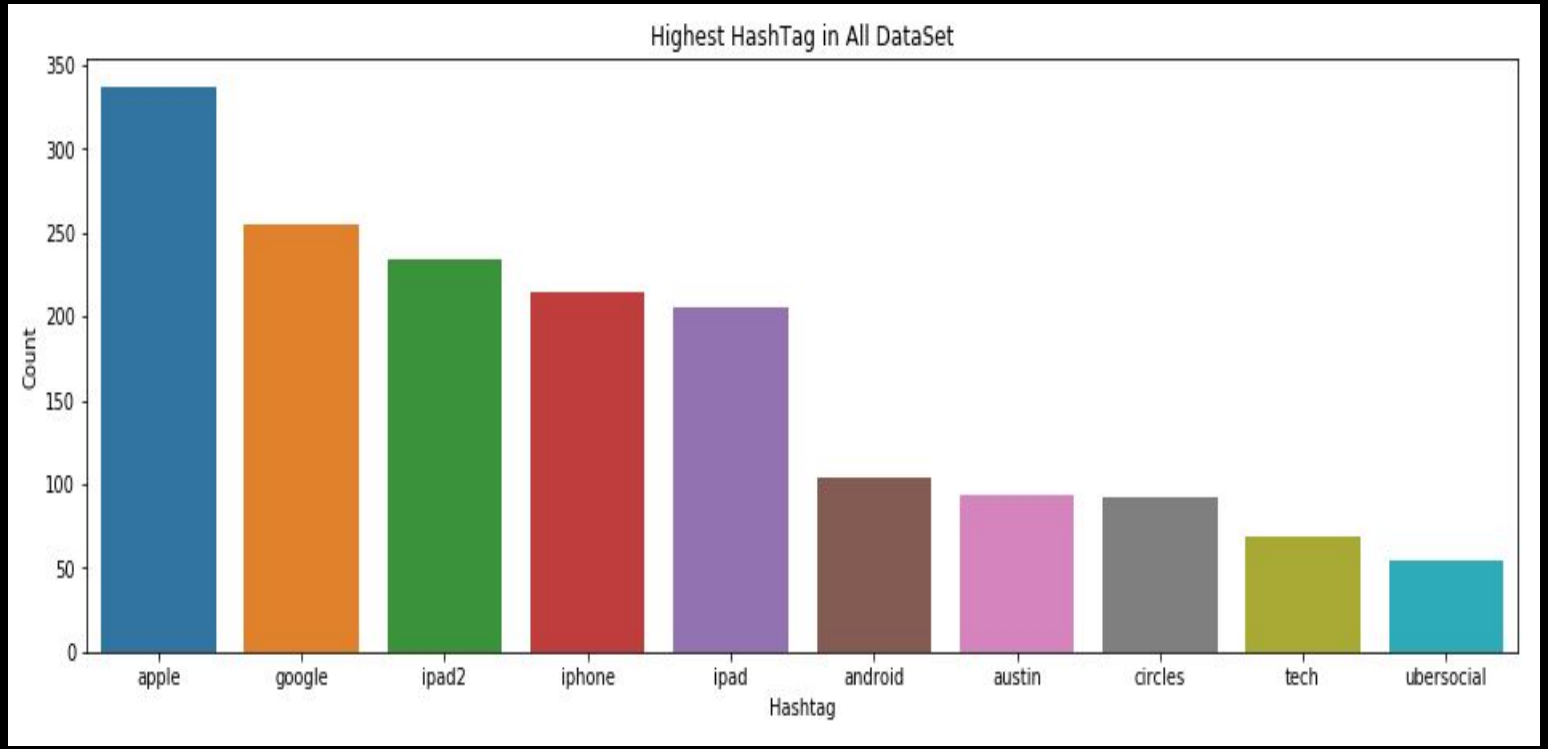
- The most common sentiment, out of the 7274 tweets in the train data, was identified as neutral
- SXSW events in general seem to gather a positive sentiment
- The imbalance in data was treated separately



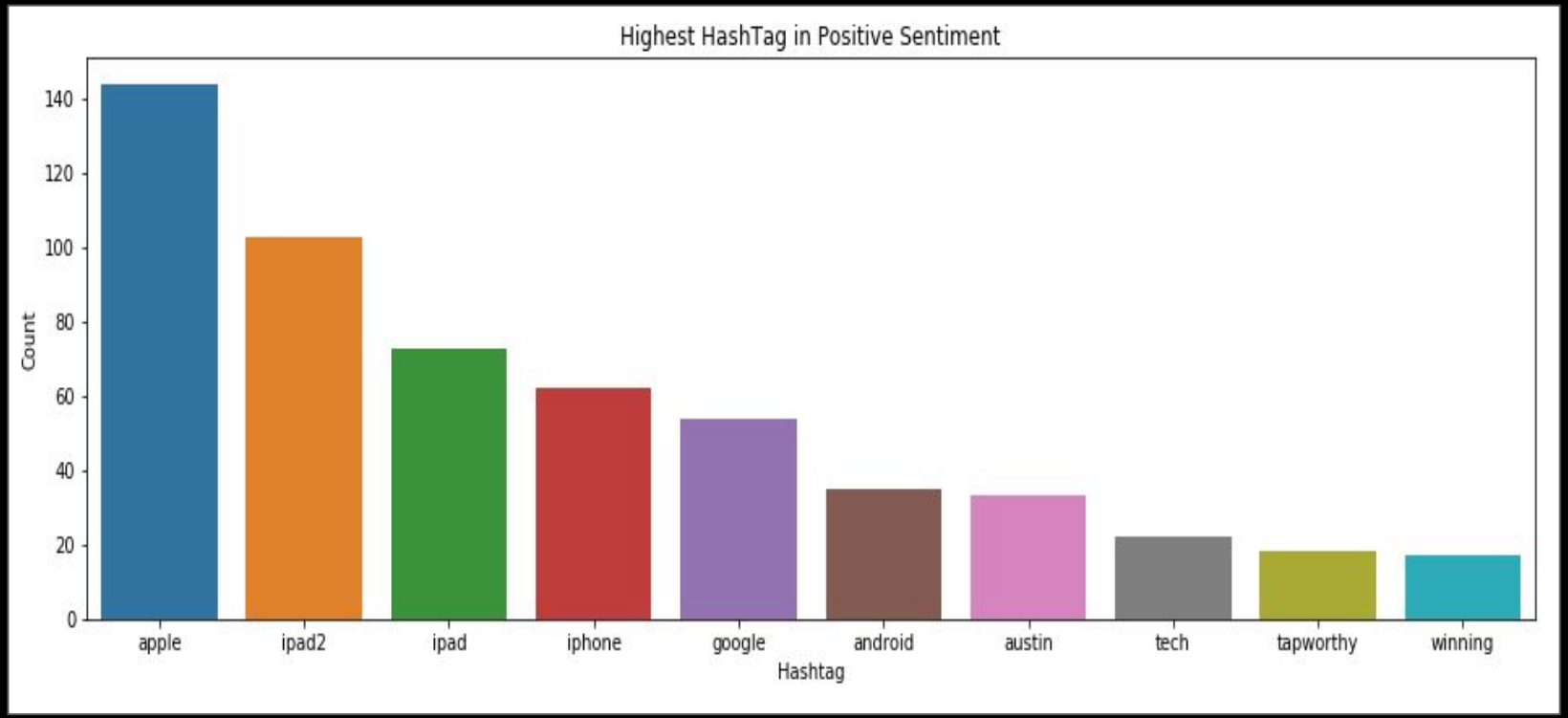
Positive tweets

Negative tweets

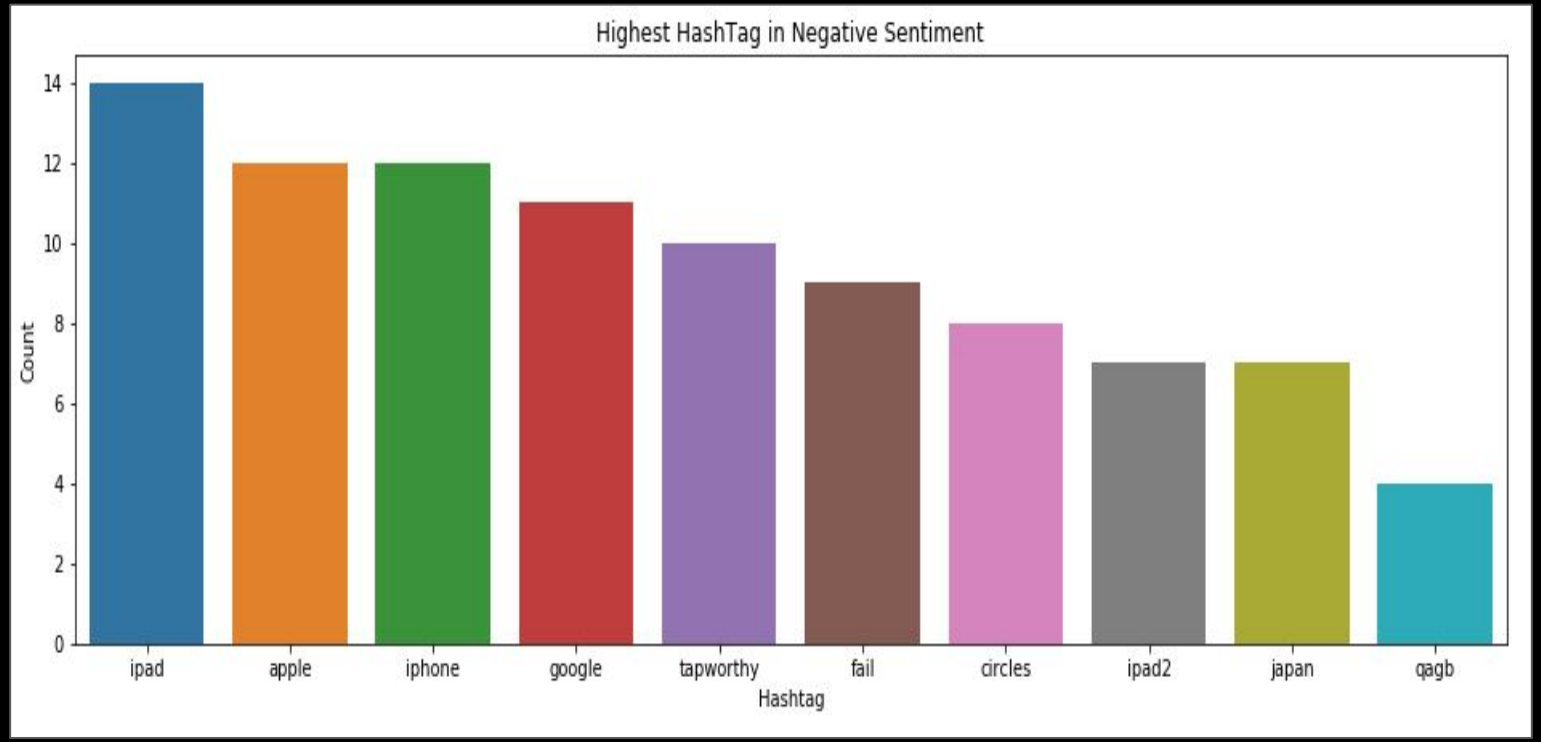
Top Overall HashTag

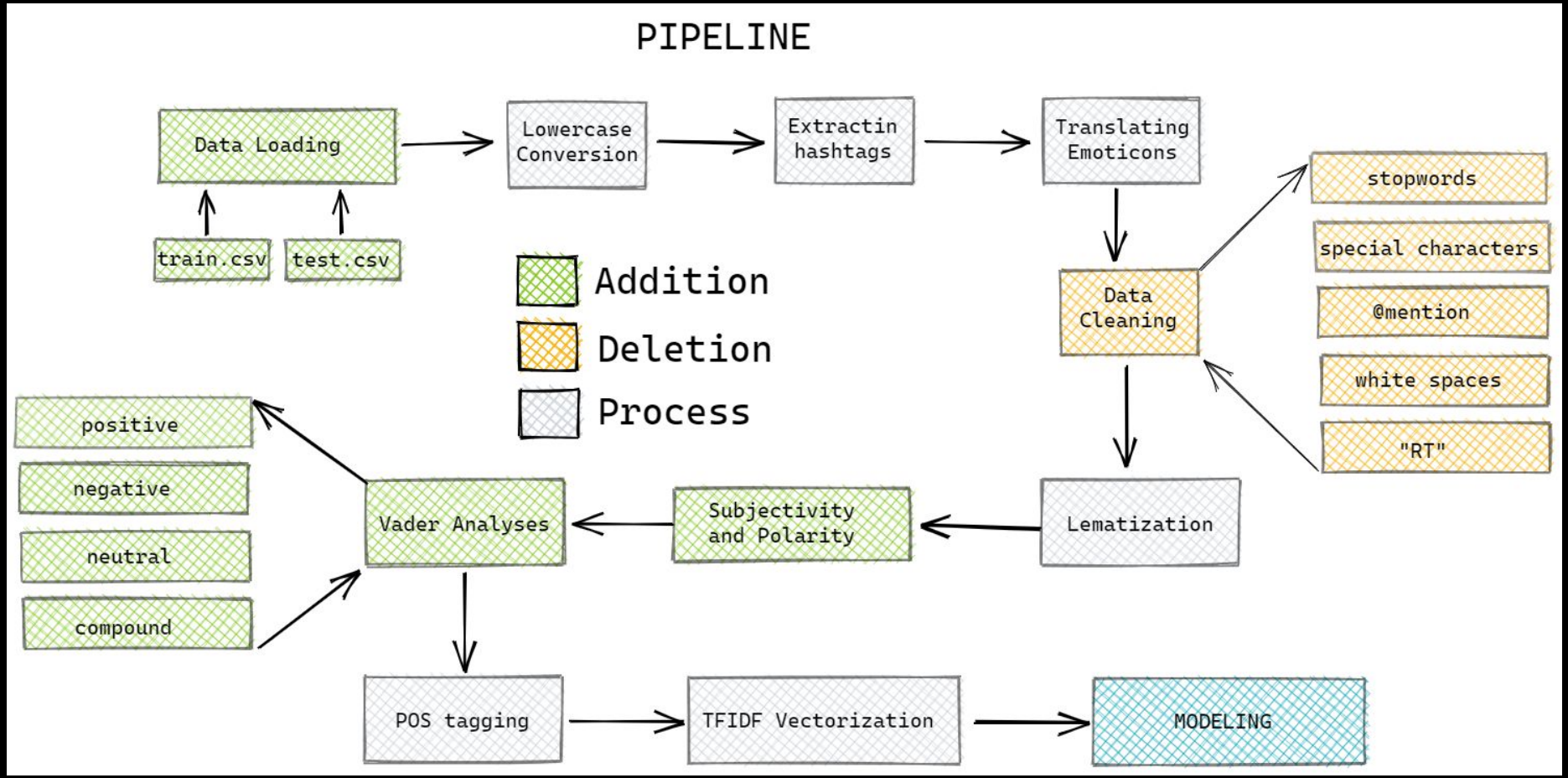


What People Liked About The Event?



What People Disliked About The Event?




PIPELINE
Data Loading
train.csv
test.csv
Lowercase Conversion
Extractin hashtags
Translating Emoticons
Addition
Deletion
Process
Data Cleaning
stopwords
special characters
@mention
white spaces
"RT"
Lematization
Subjectivity and Polarity
Vader Analyses
positive
negative
neutral
compound
POS tagging
TFIDF Vectorization
MODELING

Models and Approaches

Models:

- Logistic Regression
- Decision-Tree Classifier
- Random Forest Classifier
- Linear SVM
- Extra Trees Classifier

Approach:

- Used PorterStemmer and SnowballStemmer to stem words.
- Used Tf-Idf as Vectorizer.

Model Tuning

- We used Software-Engineering Principle of Reusability and Built a Comprehensive Function for Model tuning in which we can pass All models for Reusability in Future.
- In 2nd Advance Iteration, we tried Hyper-Parameter Tuning with GridSearchCV and used SMOTE to balance data and converted to as separate Function.

Vanilla Models With SnowBall Stemmer

Model Name	Train Score	Test Score
Logistic Regression	0.65	0.61
Multinomial Naive Bayes	0.55	0.53
Linear SVM	0.64	0.61
Decision Tree Classifier	0.60	0.54
Random Forest Classifier	0.63	0.54

Models with Porter Stemmer

Tweets processed with TF-IDF, SMOTE and Vader analyses

Model Name	Train Score	Test Score
Logistic Regression	0.639	0.644
Linear SVM	0.638	0.636
Decision Tree Classifier	0.602	0.591
Random Forest Classifier	0.658	0.644

Models with Subjective and Polarity

Tweets processed with TF-IDF, SMOTE and Vader analyses

Model Name	Train Score	Test Score
Logistic Regression	0.641	0.610
Linear SVM	0.634	0.642
Decision Tree Classifier	0.612	0.584
Random Forest Classifier	0.656	0.654
ExtraTrees Classifier	0.664	0.643

*** Models with GridSearchCV, Subjectivity and Polarity and Vader as Feature Engineering**

Final Results

- Logistic Regression and Random Forest Classifier proved to be better than the rest
- Despite Feature Engineering and Hyper-Parameter tuning we could only get a slight increase of 2-3%

Insights

- Most popular hashtags were #apple, #ipad and #google
- Marketing of the event can bee seen as a success due to higher use of #sxsw tags which accounted for more than 7300 as compared to #apple which was just used 340 times
- The ratio of positive to negative hashtags was more or less 10 :1
  For ex: for every #apple used in negative tweets there were 10 positive tweets using the same hashtags

Suggestions

- We can create twitter User Engagement Drive to involve Audience during shows Exhibit. Example: Contest, Game
- We can create a Database of these users based on the product/company liking and can provide them to our Client Company so they can provide these users offers and Exclusive tickets for future events .

Problems That Were Solved

- Identifying Sentiments about a brand or product.
- Top Tweeted Hashtags during the Event
- General polarity of audience and future prospects

Problems That Can Be Solved

- Identify Sentiments at a particular period of time ?
- Identifying Whether user was present at the Event ?
- Source of Event Information
  Example: How did User got Invite (Email/ Newspaper/App)

Next Steps

- Neural Network and Deep Learning approaches for sentiment analysis.
- Add features like timestamp, geo-tagging and user-source-platform

Thank you!