Web Analytics BIA 660-B

FINAL PROJECT REPORT

Sentiment Analysis of Airlines – United, Southwest and Virgin America

Team Members

Paresh Maheshwari

Juhi Gurbani

Rishabh Jain

INDEX

Introduction

Sentiment Analysis is the process of determining whether a piece of writing (product/movie review, tweet, etc.) is positive, negative or neutral. It can be used to identify the customer or follower's attitude towards a brand using variables such as context, tone, emotion, etc. Marketers can use sentiment analysis to research public opinion of their company and products, or to analyze customer satisfaction. Organizations can also use this analysis to gather critical feedback about problems in newly released products.

Sentiment analysis not only helps companies understand how they are doing with their customers, it also gives them a better picture of how they stack up against their competitors.

Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. The applications of sentiment analysis are broad and powerful. The ability to extract insights from social data is a practice that is being widely adopted by organizations across the world. Shifts in sentiment on social media have been shown to correlate with shifts in the stock market. The Trump administration used sentiment analysis to gauge public opinion to policy announcements and campaign messages ahead of 2016 presidential election.

Motivation/Objective

Everyday, over 100,000 flights carry passengers to and from destinations all around the world, and it's safe to say air travel brings out a mixed bag of emotions in people. Through social media, customers now have a platform to express what's on their mind while they are travelling, creating a real-time stream of customers opinion on social networks. Therefore, we thought it would be interesting to collect and analyze tweets about airlines to see how passengers use twitters as a platform to voice their opinion. We intend to scrap tweets for three airlines, namely United airlines, Southwest and Virgin America to get a better understanding how each airline is doing and then we can also compare their individual analyses with each other to get a better understanding regarding the competition levels among these airlines.

Web Scraping

Web scraping or web data extraction is data scrapping used for extracting data from websites. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser.

Methodology used by us to scrap data from Twitter involved the use of Library(Tweepy), which was connected to Twitter API and the data was downloaded and written in CSV form.

Text Pre Processing

The basic objectives of Natural Language Processing are as follows:

1. Split documents into words, punctuation symbols or segments
2. Understand context of the text
3. Extract key features or meaningful context to be used for further text mining tasks

Basic Processing steps are as follows –

1. Tokenization – split documents into individual words and punctuation symbols
2. Remove stop words and filter tokens
3. POS (part of speech) Tagging
4. Normalization: Lemmatization

Tokenization –

Tokenization is the process of breaking a stream of textual content up into words, terms, symbols, or some other meaningful elements called tokens.

N-grams –

A combination of multiple words together is called N-grams. ngrams returns a tuple of n successive words.

We have divided the tweet or the text into –

1. Unigram - Single word
2. Bigram - Two consecutive words
3. Trigram - Three consecutive words

Remove stopwords –

- Stop words are a set of commonly used words, have very little meaning, and cannot differentiate a text from others, such as "and", "the" etc.
- They are typically ignored in NLP Processing
- Stopwords have been removed in the text and the text has been converted in lower case as well

Part of speech tagging –

- Part of speech tagging is the process of marking up a word in a text as corresponding to a part of speech (e.g. nouns, verbs, adjectives, adverbs etc.), based on both its definition and its context - adjacent and related words in a phrase, sentence, or paragraph.
- Part of speech tagging helps in disambiguation and phrase extraction.
- We have taken out the following from the text –
- NN: noun, common, singular
- NNP: noun, proper, singular
- NNPS: noun, proper, plural
- NNS: noun, common, plural

Normalization –

- Normalization converts a list of words in different surface forms to a more uniform form –
- a word with different forms, e.g. organize, organizes, organized, and organizing families of derivationally related words with similar meanings, such as democracy, democratic, and democratization.
- Normalization improves text matching and reduces feature space
- **Lemmatization –**
    - Lemmatization is determining the lemma for a given word.
    - A lemma is a word which stands at the head of a definition in a dictionary, e.g. run (lemma), runs, ran and running (inflections).
    - Lemmatization is a complex task involving understanding context and determining the part of speech of a word in a sentence.

Sentiment Analysis –

- Sentiment Analysis is determining a particular sentiment of a given text – positive, negative or neutral.
- It can identify the orientation of opinion in a piece of text.
- It can predict many different emotions – happiness, sadness, anger.
- Sentiment analysis is basically the process of determining the attitude or the emotion of the writer - positive or negative or neutral.
- Given a tweet about movies, airlines or any product, it can be automatically classified in categories.

- It helps to determine the computational tone behind words.
- The majority of sentiment analysis approaches take one of two forms: polarity-based, where pieces of texts are classified as either positive or negative, or valence-based, where the intensity of the sentiment is taken into account. For example, the words 'good' and 'excellent' would be treated the same in a polarity-based approach, whereas 'excellent' would be treated as more positive than 'good' in a valence-based approach.
- Sentiment analysis has been done using TextBlob and VADER.

TextBlob -

- TextBlob is a python library for processing textual data.
- The sentiment function of TextBlob returns the polarity
- Polarity is float which lies in the range of [-1,1]
- 1 - positive statement and 0 - negative statement
- As observed, we have obtained 44.86% tweets that are negative, 32.74% tweets that are positive and 22.74% tweets that are neutral.
- The features of TextBlob are as follows –
  - Spelling correction
  - Creating a short summary of a text
  - Translation and language detection

VADER -

- VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that

is specifically attuned to sentiments expressed in social media.

- Each of the words ae rated whether its positive or negative, and how positive or negative.
- It works well on short textual sentiments as well.
- It is a simple algorithm and gives the freedom to understand the pipeline and customize it.
- We have calculated the VADER score and the sentiment of the text has been predicted as well.
- As observed, we have obtained 61.16% tweets that are negative, 9% tweets that are positive and 29.09% tweets that are neutral.

VADER as compared to TextBlob has a higher percentage of negative tweets. It is because VADER doesn't just do simple matching between the words in the text and in its lexicon. It also considers certain things about the way the words are written as well as their context. VADER recognizes capitalization, which increases the intensity of both positive and negative words. It also recognizes exclamation marks. For example, 'extremely bad' would increase the negative intensity of a sentence, but 'bad' would decrease it.

We have manually labeled a subset of samples and have not considered the tweets having a neutral sentiment.

We took the positive and negative sentiments and based on the polarity obtained by using TextBlob and VADER score, we have used the classifiers – Linear SVM and Naïve-Bayes. Linear SVM

and Naïve-Bayes classifier has been used to obtain the Classification report, precision, recall, f1score, accuracy score and the Receiver Operating Characteristic Curve.

Linear SVM –

- Linear Support Vector Machine(SVM) is a machine learning algorithm which can be used for classification. We have divided the data into training and test dataset. OnevsRestClassifier has been used It can be used for multiclass/multilabel strategy as well.
- It is a robust machine-learning algorithm to solve prediction problems as it maximizes margin. SVMs are effective when the number of features is quite large. It works well both on linear and non-linear data. It handles high dimensional data well. It works really well with clear margin of separation and is effective in high dimensional spaces.

- SVM follows the following structure for implementation –

1. Import the Library
2. Object Creation
3. Fit the model
4. Prediction

Naïve-Bayes –

- Naive Bayes classification technique is based on the Bayesian theorem - finding functions describing the probability of belonging to a given class feature
- It aggregates information using conditional probability with an assumption of independence among features

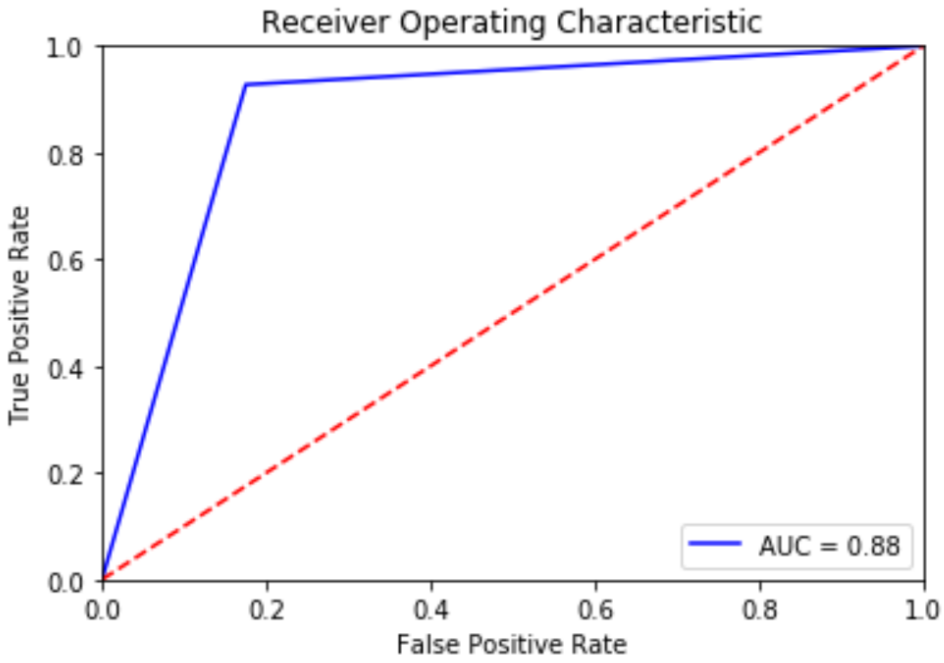
Analysis of Experiment Results-

Linear SVM Classification Report –

Results obtained when Sentiment Analysis has been done using TextBlob -

	precision	recall	f1-score	support
0	0.89	0.82	0.86	611
1	0.88	0.93	0.90	868
avg / total	0.89	0.89	0.88	1479

Receiver Operating Characteristic Curve –

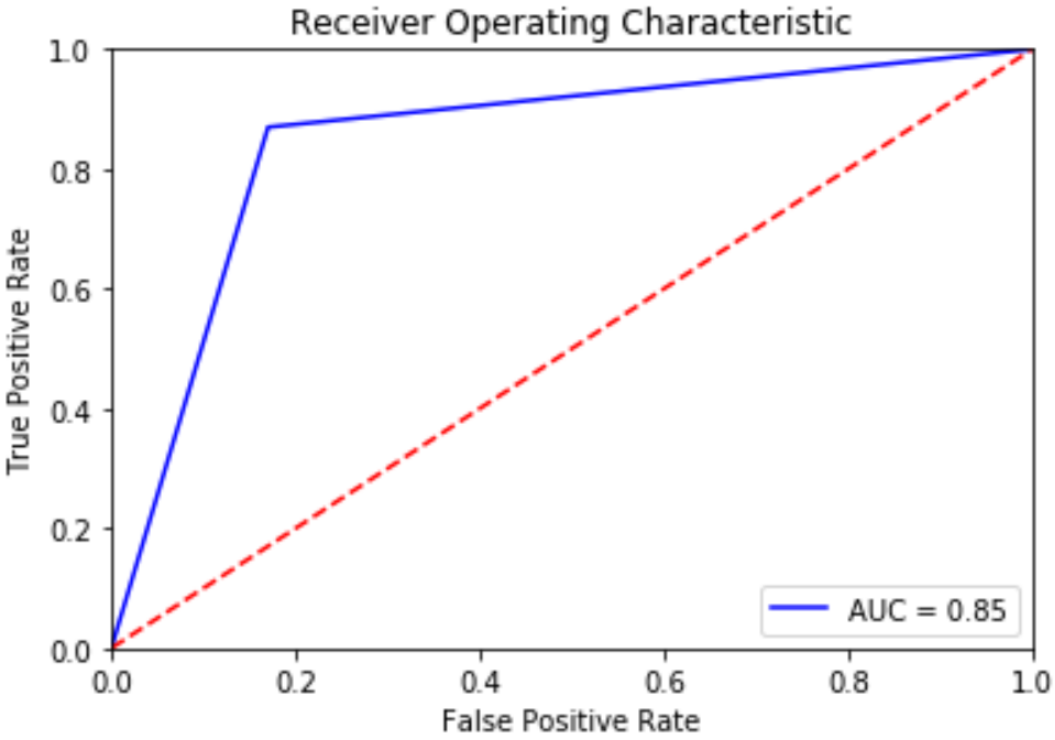


Naïve Bayes Classification Report -

Results obtained when Sentiment Analysis has been done using TextBlob -

	precision	recall	f1-score	support
0	0.82	0.83	0.82	611
1	0.88	0.87	0.87	868
avg / total	0.85	0.85	0.85	1479

Receiver Operating Characteristic Curve –



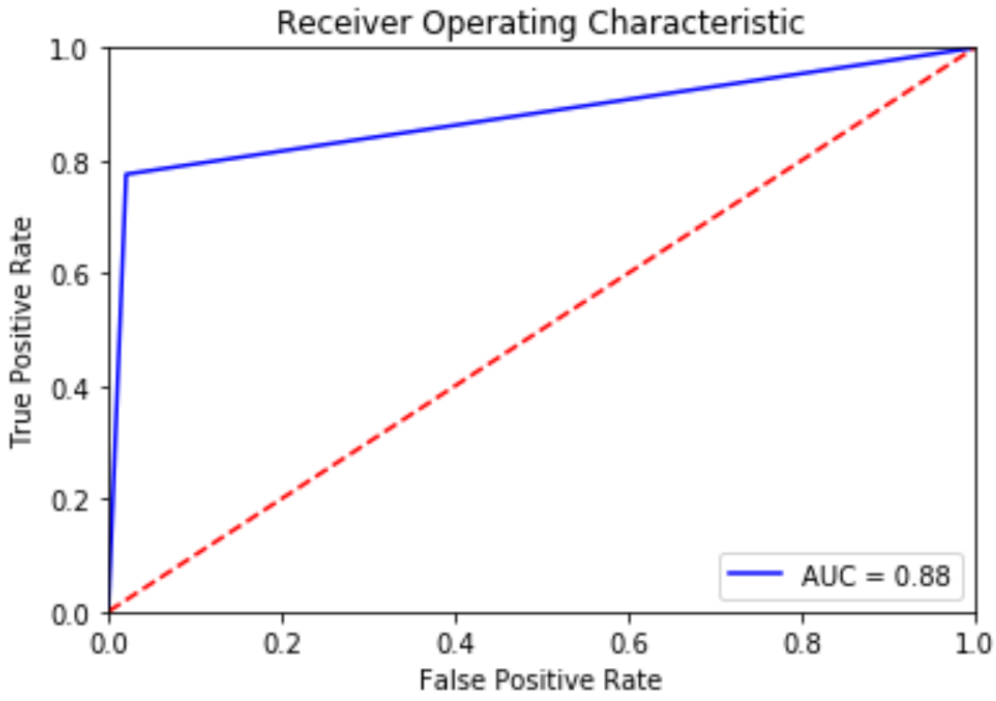
The accuracy obtained is 89% while using Linear SVM as compare d to Naïve-Bayes, the AUC for Linear SVM is 0.88 whereas for Naïv

e-Bayes is 0.85. Linear SVM is a better suited model when Sentim ent Analysis has been done using TextBlob.

Linear SVM Classification Report -
Results obtained when Sentiment Analysis has been done using VADER –

	precision	recall	f1-score	support
0	0.93	0.98	0.95	779
1	0.93	0.78	0.84	263
avg / total	0.93	0.93	0.93	1042

Receiver Operating Characteristic Curve –

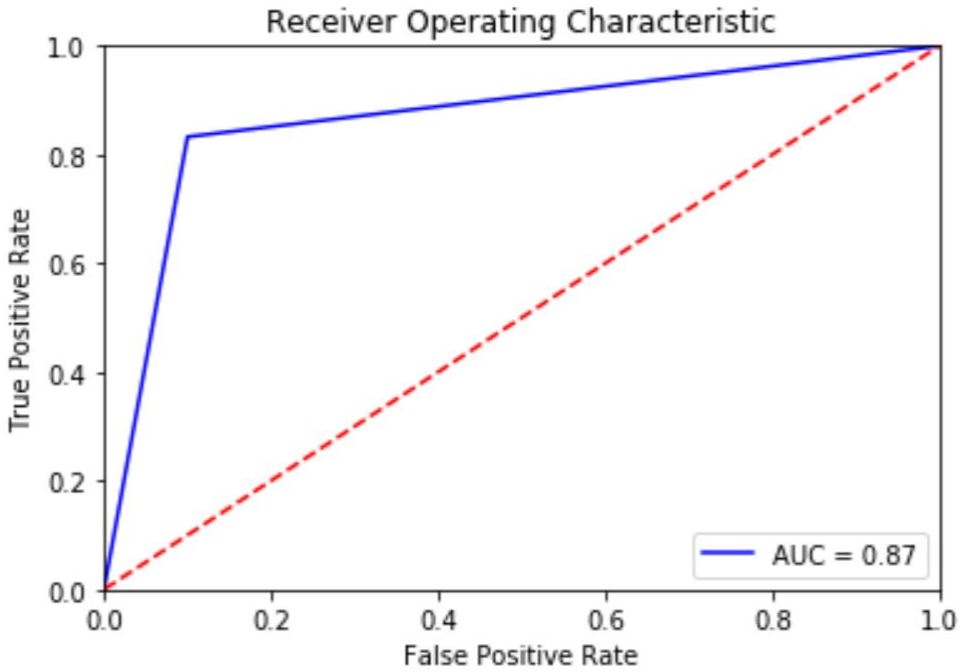


Naïve Bayes Classification Report -

Results obtained when Sentiment Analysis has been done using VADER –

	precision	recall	f1-score	support
0	0.94	0.90	0.92	779
1	0.74	0.83	0.78	263
avg / total	0.89	0.88	0.89	1042

Receiver Operating Characteristic Curve –



The accuracy obtained is 93% while using Linear SVM as compare d to Naïve-Bayes, the AUC for Linear SVM is 0.88 whereas for Naïv e-Bayes is 0.87. Linear SVM is a better suited model when Sentim ent Analysis has been done using VADER.

Recommendations

- United airline has more traffic on its twitter handle than Southwest and Virgin America Airlines.
- The point of concern for United airlines management should be the large extent of negative sentiment or usage of negative expressions by the users.
- United Airlines management can also train their social media team in being more proactive and sensitive towards users concern and encourage the use of personalized responses in place of scripted ones.

Future Scope

- We would use Clustering algorithm such as k-means clustering.
- We would also consider tweets having neutral sentiments.

- Airlines also have other social media platforms like Facebook and Instagram where customers comment and express their emotions.
- We intend to incorporate data from these other platforms to our existing database.