Sentiment Analysis

Technical Report

20 Sep 2019

Problem statement

We want to help our digital marketing team to identify how customers express their opinions about our airline and other competitors easily and efficiently. So, the digital marketing team can spot negative or positive speech then escalate it to the responsible team.

Thousands of customers use social media platforms, especially Twitter, to express their opinions. The opinion of our customers is important and influences our business in the airline market. Therefore, we need to know what customers hate about us and their complains to solve it or what they like about us so we can keep and develop it.

We will use Machine Learning and Deep Learning techniques to identify the opinion of our customers. That will be the fuel that will help the business move forward and increase revenue.

SOLUTION ***3:12***

Exploratory Data Analysis

We need to Explore our data (corpus) to answer some important questions

- What is the form of the data?
  - Structured
    - CSV file
    - SQL DB
  - Unstructured
    - Text
    - Images
- Does the data have labels?
- How many instances does the data have?
  - Small Dataset – less than 10000 Instances
  - Large Dataset – more than 10000 Instances
- How many features does the data have?
  - Multidimensional Dataset
    - Small Dataset – less than 10 dimensions
    - Large Dataset – more than 10 dimensions
- What is the type of each feature?
  - Quantitative (Interval)
  - Qualitative (Nominal)
- How to describe the numerical data?
  - Sample/Population
  - Central Tendency (Mean, Median or Mode)

- Are there any Null/Missing values?
- What does The Visualization tell us?

Findings

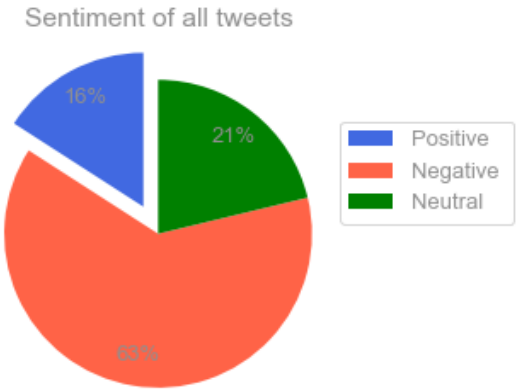
The data we will work on is

- Unstructured data in a CSV file
- The data is Labeled
- A Large Dataset – it has 1400+ instances
- A Large Multidimensional Dataset – it has 15 features
- There are 2 types of Features
  - 1 Quantitative Feature (Interval)
    - Retweet Count
  - 14 Qualitative Features (Only 4 important features)
    - Airline Sentiment
    - Negative Reason
    - Airline
    - Text
- The Dataset is a Sample of 14640 instances, and we can describe it as follows:
  - Retweet Count
    - There are 767 instances have values greater than
    - only 5% of the tweets were retweeted
    - The maximum number of retweet count is 44

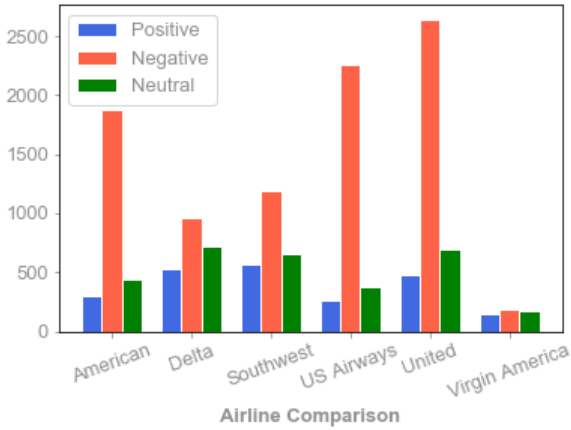
- There are no Null/ missing values

Data Visualization

- Figure 1,2 (sentiments of all Tweets - Retweets)

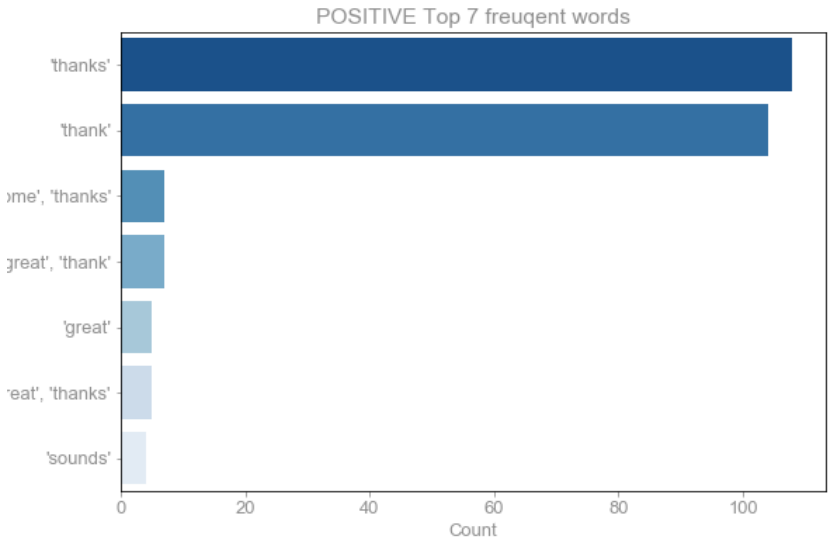


- Figure 3



- Figure 4,5 (Positive- Negative most frequent words)

- Figure 6,7 (Positive- Negative Top 7 frequent words)



Data Preparation

- The data had many features that we would not need in this task. So, only 5 features were kept during EDA phase.
    - Retweet Count
    - Airline Sentiment
    - Negative Reason
    - Airline
    - Text
- The Corpus needed to be cleaned to create the Sentiment Lexicon
    - Emojis, symbols, stop words… etc were removed
    - Empty tweets were removed
    - The sentences were tokenized
- The Sentiment Lexicon became ready to be used

Note:

- The sentiment lexicon needs to be converted to a numerical form to be feed into the model

Machine Learning Modelling

Corpus preparation

The corpus needs to be converted to numerical values to be feed to the model.

Used Techniques:

- Convert the corpus into a vector
  - TFIDF Vectorizer
- Convert labels into numerical categories
  - Label Encoder
  - One Hot Encoder

The corpus is ready and now we can start building the model but there are some steps involved to choose the best algorithm.

1-Choosing the algorithm

From EDA and Visualization, we conclude that

- It is a Supervised Learning task (Labeled data)
- It is a batch learning task (Large dataset, but can be fitted to the memory)
- It is a classification NLP task

Different classification ML models can be applied such as Random Forest or Logistic Regression… etc. to choose the best model. In this task we will use only Random Forest Classifier for the time constrains.

Building Random Forest Classifier

Used Libraries:

- Sciekit-Learn

Method:

1. Dividing the dataset into two sets
   - Training set
   - Test set
2. Creating an instance of Random Forest Classifier
3. Train the Random Forest with the training set
   - .8 of the corpus
4. Testing the Random Forest with the test set
   - .2 of the corpus
5. Hyperparameter tuning
   - Randomized Search
6. Training and testing the new model after tuning
7. Comparing between the first model and tuned model

Results:

```
First model accuracy:  73.4%
Tuned model accuracy:  73.9%
There is .5% increase in the accuracy
```

Validation

- Using the test set the model was Validated.
- The model needs to be tuned to give better results.

Deep Learning Architecture

We will use **Recurrent Neural Network (LSTM)** as it is gives good results in Sentiment Analysis.

Corpus preparation

Used techniques

- Creating words embedding
  - GloVe (Global Vector)
- Convert labels into numerical categories
  - Label Encoder
  - One Hot Encoder

Building the model

Used Libraries:

- Keras

Method:

1. Dividing the dataset into two sets
   - Training set .8
   - Test set .2
2. Creating an instance of the model
   - Sequential()
3. Fine tuning the model
   - Activation functions
     - sigmoid
   - Loss Function
     - categorical_crossentropy
   - Metrics
     - Accuracy

Validation

- Using a Validation Set of .2 of the training set the model was Validated

Results:

```
Accuracy = 63.86
Loss = 0.90
```

The model needs to be tuned to give better results

Saving the model

The model was saved as **pickle** file, to be used in deployment.