Course: PR **Date:**12/12/2020

M.Navyasri **S20180020222**

P.Sruti **S20180020233**

Y.Vishnu Sreya **S20180020262**

Project Report

Sentiment Analysis

Project description

Sentiment analysis refers to analysing the feelings or opinions using data in text format or images.Sentiment analysis helps companies in their decision-making process for example if public sentiment towards a product is not so good, a company may try to modify the product.

So,the main goal of the project is to detect the patterns in the data and find the sentiment of the sentence we provide.

Problem Description and approach

Given tweets about six US airlines, the task is to predict whether a tweet contains positive, negative, or neutral sentiment about the airline. This is a supervised learning task which classifies text strings into positive, neutral and negative categories.we have used four traditional machine learning algorithms to reach the output.

Implementation:

1. Dataset :

Data - this link contains the data , we got the data from a reference

Shape of the data - `(14640, 15)`

Number of features - 14 + 1 (output - airline_sentiment)

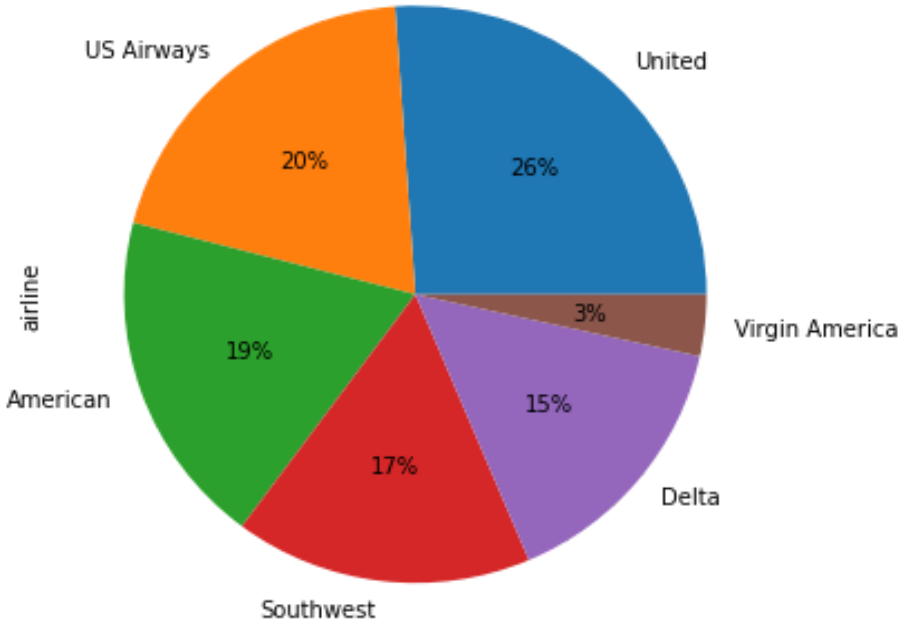
```
airline_tweets.columns
```

```
Index(['tweet_id', 'airline_sentiment', 'airline_sentiment_confidence',
       'negativereason', 'negativereason_confidence', 'airline',
       'airline_sentiment_gold', 'name', 'negativereason_gold',
       'retweet_count', 'text', 'tweet_coord', 'tweet_created',
       'tweet_location', 'user_timezone'],
      dtype='object')
```

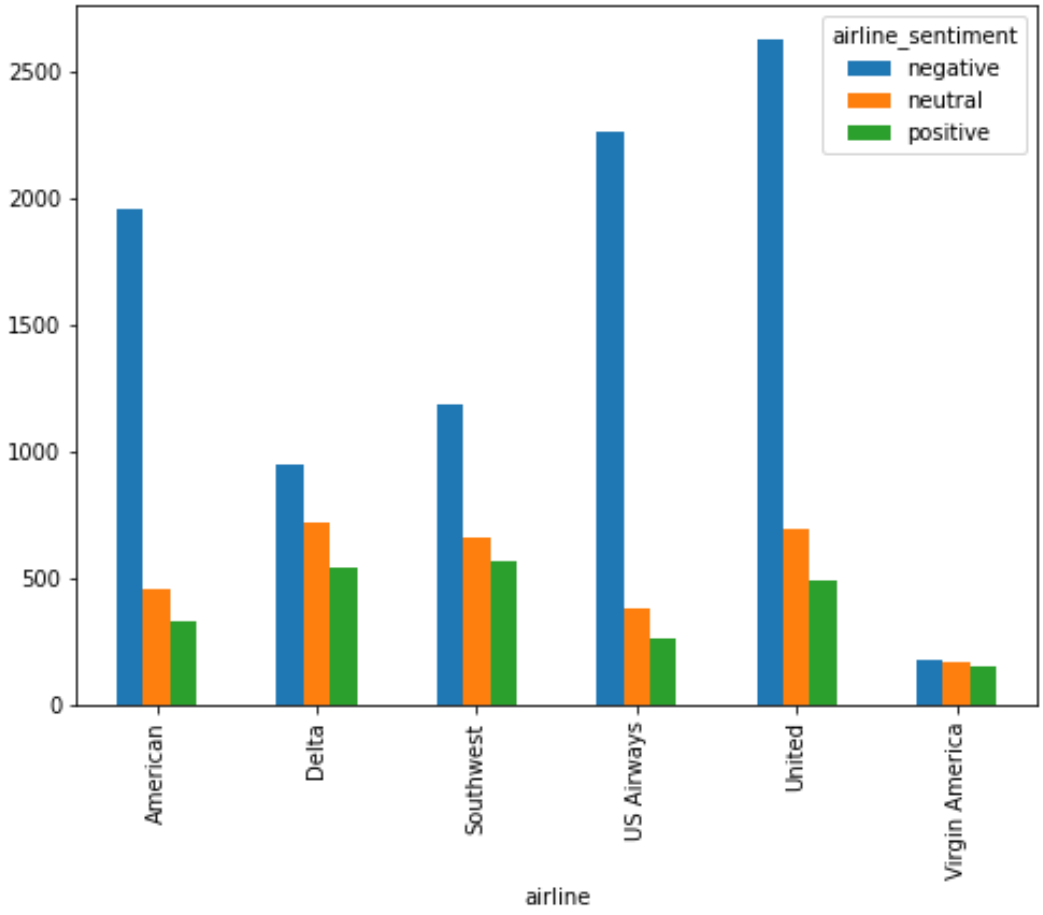
Number of tweets (examples) - 14640

2. Data Analysis:

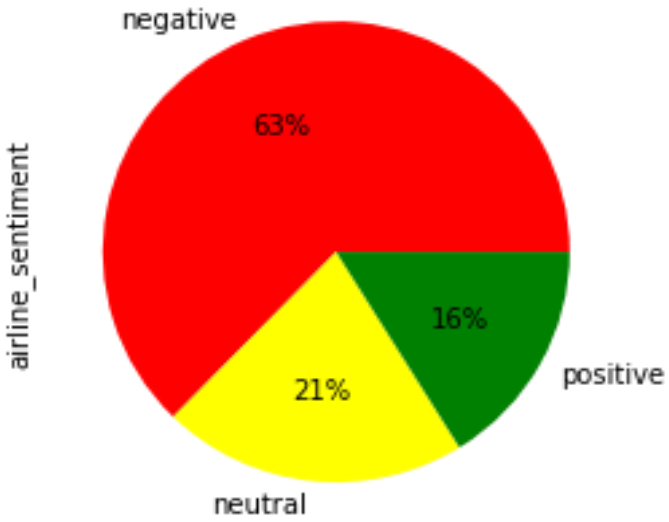
- We can observe the components of each airline in whole data



- Below graph tells about the distribution of 3 classes(sentiments) in all airlines



- Distribution of sentiment across all tweets



3. Data Cleaning

Tweets contain many slang words and punctuation marks. We need to clean our tweets before they can be used for training the machine learning model.

To train the model , we only need tweet data . So we are extracting model input data (text column) and model output (airline_sentiment column).

We preprocess the data using regex :

- Removing all special characters
- Removing single characters from the start
- Substituting multiple spaces with single space
- Removing prefixed b
- Converting all the characters to lowercase

4. Text in Numeric form

Statistical machine learning models use mathematics as its background and it works only on numerical values.So we have to change the tweet text into numbers, we can use different methods for this purpose like bag of words , word2vec , TI-DIF scheme .we have used TI-DIF method in our project.

- **TI -DIF method :**
  words that occur less in all the documents and more in individual documents contribute more towards classification.

5. Splitting the data

We split the data into 80-20 % such that training data is 80% of whole data and 20% is testing data.

6. Training the Model

We have used 5 machine learning algorithms to train our model

1. Random Forest classifier
2. Linear SVM
3. Gaussian SVM
4. Logistic Regression
5. K-nearest neighbours

7. Predictions :

We predict the outputs of input test data using all 5 models to calculate the accuracies and confusion matrix

Analysis

- Random Forest :

```
confusion_matrix
[[1723  108   39]
 [ 326  248   40]
 [ 132   58  254]]
accuracy
0.7599043715846995
```

- Linear SVM

```
confusion matrix
[[1697  129   44]
 [ 264  306   44]
 [ 107   56  281]]
accuracy
0.7800546448087432
```

- Gaussian SVM

```
Confusion_matrix
[[1870    0    0]
 [ 614    0    0]
 [ 444    0    0]]
Accuracy
0.6386612021857924
```

- Logistic Regression

```
Confusion-matrix
[[1732  115   23]
 [ 268  304   42]
 [ 118   54  272]]
Accuracy
0.7882513661202186
```

- K -nearest neighbours

```
Confusion matrix
[[1500  316   54]
 [ 252  310   52]
 [ 118  103  223]]
Accuracy
0.6943306010928961
```

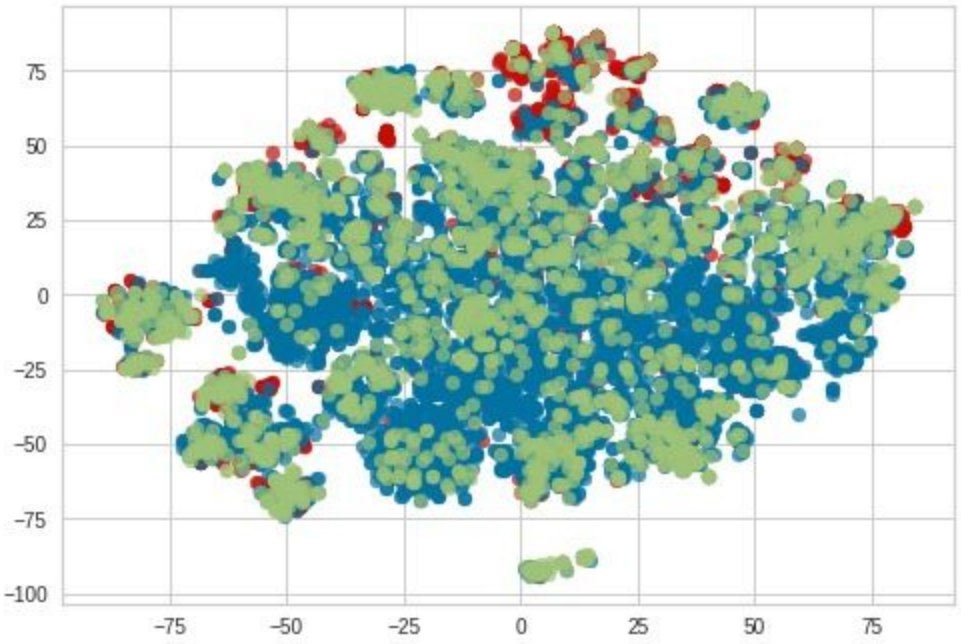
Results

Comparing accuracies of all trained models ,
we can say that **Logistic regression** gives better performance than remaining algorithms
Accuracies of all models in descending order

MODEL	ACCURACY
Logistic Regression	0.788213661202186
Linear SVM	0.780054644808732
Random Forest	0.759904357158470
K-nearest neighbours	0.6943306010928961
Gaussian SVM	0.6386612021857924

Why Logistic Regression works best :

The shape of training data after the TF-IDF method gives us 2701 features.
So we do dimensionality reduction using Principal Component analysis (PCA) on training data and after plotting it , we got the following plot
In 3D space, we can see they are linearly separable
So,Logistic Regression works best



**** working code ****
https://github.com/artemis-2701/sentiment-Analysis

Contributions :

1 . P.Sruti - Data cleaning , Exploratory Data Analysis + Report
2 . M.NavyaSri - Text to numeric form + models+predictions + Report
3. Y.VishnuSreya - Models + prediction, evaluation + Report