T w i t t e r S e n t i m e n t Analysis

Marta Mas

Raúl Medina

Master in Data Science, Kschool

Abstract

Sentiment Analysis is the most common text classification tool that analyses an incoming message and tells us the underlying sentiment. Twitter contains huge data about users and the relationships between them. The linguistic analysis of Twitter has been applied as a tool to understand the behavior between individuals or organizations. The main purpose of this work has been to carry out text processing and sentiment analysis of Twitter data. The Twitter sentiment analysis has been applied over five different American airlines and their corresponding polarity (positive or negative). Three supervised classifiers have been used: Random Forest, Logistic Regression and SVM. The results show that the classifier that has provided the best results has been SVM: 0.84 of both accuracy, precision, recall and F1. However, the results do not vary considerably if we compare them with those obtained with the Logistic Regression classifier: 0.83 accuracy, 0.83 precision, 0.84 recall and 0.83 F1.

INDEX

1. Introduction

Currently, as Reyes, Rosso and Veale (2013) argue: "Web-based technologies have become a significant source of data in a variety of scientific and humanistic disciplines, and provide a rich vein of information that is easily mined" (pp. 239). In addition, Reyes et. al (2013) also states: "User-generated Web 2.0 content (such as text, audio and images) provides knowledge that is topical, task-specific, and dynamically updated to broadly reflect trends, behavior patterns and social preferences" (pp. 240). Therefore, nowadays, social networks are a really useful source of information for research purposes on the actual use of the language, since users can express themselves freely, informally and spontaneously through them.

Twitter is a social network that allows its users to post text messages, called tweets, which are displayed on the user's homepage. This type of service corresponds to a form of blogging, microblogging, whose main feature is its simplicity and ability to synthesize. The linguistic analysis of Twitter has been applied as a tool to understand the behavior between individuals or organizations. Twitter contains huge data about users and the relationships between them. To analyze and extract useful information from these huge data, special mining tools based on graphics that can easily model the structure of social networks are required. Some of these analysis tools are available with their own features and benefits.

Also, with regard to the relationship between Twitter and the analysis of sentiment, as stated by Liu (2012):

Opinions and its related concepts such as sentiments, evaluations, attitudes, and emotions are the subjects of study of sentiment analysis and opinion mining. The inception and rapid growth of the field coincide with those of the social media on the Web, e.g., reviews, forum discussions, blogs, microblogs, Twitter, and social networks, because for the first time in human history, we have a huge volume of opinionated data recorded in

digital forms. (pp. 5). Under the same point of view, Hernández et al. (2016) argue that the large amount of information that is transmitted from social networks and microblogs, such as Twitter, increasingly attracts the attention of researchers in the area of sentiment analysis.

Sentiment Analysis is the most common text classification tool that analyses an incoming message and tells whether the underlying sentiment is positive, negative our neutral. Therefore, sentiment analysis allows us to identify and extract subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service. Finally, in this project, we propose to create a text classification model capable of classifying a given text taking into account its polarity. First, we decided to make the distinction between positive or negative and neutral polarity. However, afterwards, as we will explain later, we decided to make the distinction between positive polarity and negative polarity, grouping the neutral polarity and the positive polarity as a single polarity. Therefore, the main objective of this project is to analyze and create a model that predicts, through the use of different classifiers of Machine Learning, the popularity, in Twitter, of different competing airlines: America, Delta, Southwest, United, US Airways and Virgin America.

2. Methodology

2.1 Data cleaning with Python

The data originally came from Crowdflower's Data for Everyone library. As the original source says, a sentiment analysis job about the problems of each major U.S. airline. Twitter data was scraped from February of 2015 and contributors were asked to first classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service”). The data we're providing on Kaggle is a slightly reformatted version of the original source. It includes both a CSV file and SQLite database.

The dataset is composed of fifteen Fields, of which we have mainly used three:

- Airline: the name of the airline (America, Delta, Southwest, United, US Airways and Virgin America). This field has allowed us to analyze the number of tweets, taking into account their polarity, of each of the companies. This field has been generally used for the exploratory analysis of the data, as well as for its visualization through Tableau.
- Text: the text of each of the tweets written by users on Twitter. This field has been used to create the text classification model and the exploratory analysis of the data. Likewise, a preprocessing of each of the tweets has been carried out as well as their vectorization.
- Airline_sentiment: the polarity of each of the tweets: positive, neutral, negative. This field has been used for the exploratory analysis of the data and the creation of the classification model.

However, despite the fact that the corpus was composed of a set of negative, positive and neutral tweets, it was decided to group as positive tweets those labeled as both neutral and positive. In this way, our corpus consists of two sets of data, text with negative polarity and text with positive polarity. This decision was made due to the lack of a necessary balancing of the data, there was a greater amount of negative data (9178) compared to positive data (2363) or neutral (3099). In this way, finally, the data set with negative polarity remained (9178) while the data set with positive polarity was (5462). Likewise, it was found that the classifiers provided best results in the classification.

On the other hand, the computer programming language used in the project has been Python 3.7.1 at all times and the libraries mainly used to carry out the project have been:

- Pandas: to easily manipulate the data.
- Numpy: for scientific computing.
- Wordcloud: data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. For generating word cloud in Python, modules needed are: matplotlib, pandas and wordcloud.
- Seaborn: data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
- Scikit-learn: to use machine learning methods and implement the different classifiers.
- Matplotlib: to make graphics.
- Emoji: to represent in text format each of the emojis of the tweets.
- Nltk: to eliminate the stopwords.
- Re: to create regular expressions and preprocess the text.
- Itertools: to draw the confusion matrix
- Tweet-preprocessor 0.5.0: to eliminate the links, mentions and hashtags of each of the tweets.
- Pywsd.utils: to lematize each of the words in the tweets (preprocessing).

Finally, some example of the corpus used would be the following:

- Positive tweet: I ✂ ! flying @VirginAmerica. ！ ！ 👍

- Negative tweet: @VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces & they have little recourse

2.2 Data visualization

Firstly, the dataset (Tweets.csv) is imported and all the fields that compose it are visualized to observe what type of data is composed of it and to be able to select which ones will be used in the project. Then, an exploratory analysis is carried out and the amount of data contained in each of the fields is counted and the empty fields are displayed.

In addition, the number of negative and positive tweets in the dataset is displayed. Two of the three main fields of the project ('airline' and 'airline_sentiment') are selected to analyze the amount of each of the polarities of the tweets in relation to each of the airlines. All this is done through the Matplotlib library with a plot bar.

Finally, Wordcloud is used, which allows us to make a visual representation of text data. It displays a list of words, the importance of each beeing shown with font size or color. This format is useful for quickly perceiving the most prominent terms.

2.3 Supervised Machine Learning

To carry out the project, a model based on supervised machine learning is created because we have input variables (x : the text of the vectorized tweets) and an output variable (Y : the polarity label) and we use an algorithm to learn the mapping function from the input to the output. The goal is to approximate the mapping function so well that when we have new input data (x) that we can predict the output variables (Y) for that data. Also, the problem we propose to solve is not a problem of regression but of classification, the variable output is a category: positive or negative. Also, the supervised machine learning algorithms that we use for text classification are: Random Forest, Logistic Regression, Support Vector Machine (SVM) and Decision Tree classifier.

3. Products Obtained

The products / notebooks obtained when carrying out the project were the following:

- *TfM_Explo_Desc_V1*: An exploratory analysis of the data is carried out (tweets.csv), each of the fields that make up the dataset are explored and those that are considered useful for the project are selected. In addition, the fields selected as useful are also related to each other and the results are represented by graphs. Finally, the Wordcloud library is used for representing text in which the size of each word indicates its frequency or importance.

- *TfM_Sentiment_Analysis_V2*: The data is imported, a dataframe is created with the columns that were considered useful for the project and the data is cleaned. To begin with, those 'neutral' labels that belong to the 'airline_sentiment' field are replaced by the 'positive' name, so that the set of labels is only positive / negative. Next, the preprocessing of the text is carried out. However, for each of the previous steps it is verified that the extracted results are the correct ones. Regarding the preprocessing of the text of the tweets, the main procedures have been the following:
    - Elimination of hastags, mentions and link
    - Elimination of emojis
    - Treatment of linguistic contractions
    - Treatment of emoticons
    - Treatment of punctuation marks, uppercase characters, etc.
    - Lemmatization

- *TfM_Sentiment_Analysis_Classifiers_V3*: Mainly the preprocessing of each of the tweets is done again and some graphs are made to provide clearer and more precise information about the type of data with which we are going to work. Next, a Bag of Words is created with the 6000 most frequent tokens without stopwords and CountVectorizer is used to convert to collection of text documents to a matrix of token counts. Then, with this obtained data, a TF-IDF model is created to transform a count matrix (X) to a normalized tf-idf representation (X). Tf means term-frequency while tf-idf means term-frequency times inverse document-frequency. Subsequently, the data obtained from the TF-IDF model, as well as their respective labels with the respective sentiment of each of the tweets, are divided into a training set and a test set. Finally, each of the classifiers is applied and results of its correct operation are extracted through a series of graphs of their respective matrices of confusion.

- *TfM_Sentiment_Analysis_Opt_Classifiers_V4*: This is the last notebook created for the creation of the classification model. The main objective of this notebook is to optimize the results obtained so far and provided in the TfM_Sentiment_Analysis_Classifiers_V3 notebook. The main difference between this notebook and the previous notebook is that it uses the Tfidfvectorizer model and the previous one uses Tfidftransformer. The differences between the two modules can be quite confusing: with Tfidftransformer we will systematically compute the word counts with CountVectorizer, generate Inverse Document Frequency (IDF) values and then compute Tf-idf scores. However with Tfidfvectorizer we will do all three steps at once. Under the hood, it computes the word counts, IDF values, and Tf-idf scores all using the same dataset.

4. State of the Art

Sentiment analysis is a challenge of the Natural Language Processing (NLP), text analytics and computational linguistics. In a general sense, sentiment analysis determines the opinion regarding the object/subject in discussion. In our case, it determines the sentiment of a twitter user regarding his experience with a certain airline. Airlines can use sentiment extremity and opinion point acknowledgment to pick up a more profound comprehension and the general extent of estimations. These experiences can progress focused insight, accomplish better brand picture, enhance client benefit, and upgrade competitiveness. In the airlines is hard to gather information about clients' input by polls, yet Twitter gives a sound information source to them to do client opinion examination.

Sentiment analysis has been practiced on a variety of topics, for example, for movie reviews or product reviews. However, the present project focuses on tweets written by Twitter users. Regarding the models created so far on the analysis of sentiments of Twitter can be confirmed that there are a lot of projects that address the problem, which can be seen for example in the Kaggle community, page through which we have extracted the data to carry out the project. The techniques used to address the classification problem have been varied. As Thakkar and Patel (2015) say "Opinion mining (sentiment extraction) is employed on Twitter posts by means of following techniques:

- Lexical analysis
- Machine learning based analysis
- Hybrid/Combined analysis"

5. Results Obtained

5.1. Data Set's Description

The shape and typology of informed fields is described in this section.

Shape of the data set:

-Number of Columns: 14640

-Number of Rows: 15

The name of the columns:

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

The main fields that provide information to our study are:

-airline. This variable shows us the main airlines registers in the dataset. The companies studied are:

- 'Virgin America'
- 'United' 'Southwest'
- 'Delta'

- 'US Airways'
- 'American'.

-airline_sentiment. This variable shows us the main sentiments register in the dataset. These sentiments are classified as:

- 'neutral'
- 'positive'
- 'negative'.

-negativereason. The main negative reason are:

- Bad Flight
- Cancelled Flight
- Customer Service Issue
- Damaged Luggage
- Flight Booking Problems
- Late Flight
- Lost Luggage
- Longlines

5.2. Data Volumetry

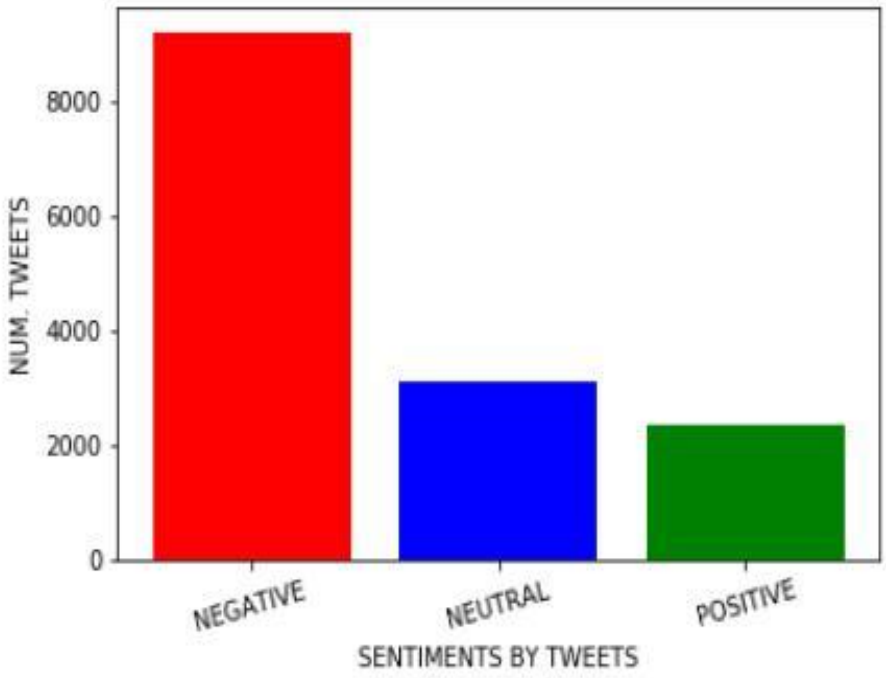
The nulls values about each field are:

Field	N. Null Values
tweet_id	0
airline_sentiment	0
airline_sentiment_confidence	0
negativereason	5462
negativereason_confidence	4118
airline	0
airline_sentiment_gold	14600
name	0
negativereason_gold	14608
retweet_count	0
text	0
tweet_coord	13621
tweet_created	0
tweet_location	4733
user_timezone	4820

5.3. Plot Data

Sentiments by Tweets: This plot shows the number of Tweets by Sentiment.

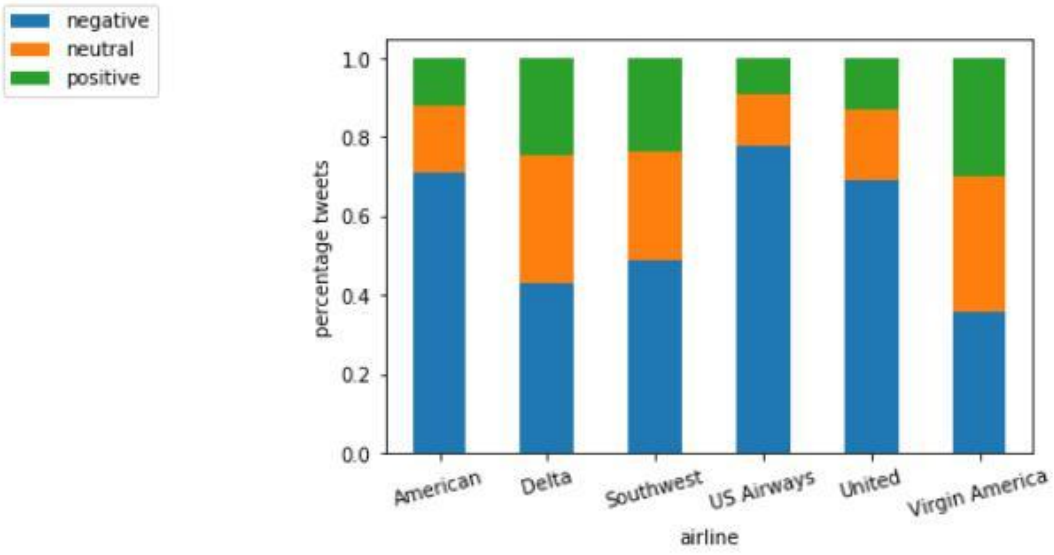
Negative: 9178
Neutral: 3099
Positive: 2363



!

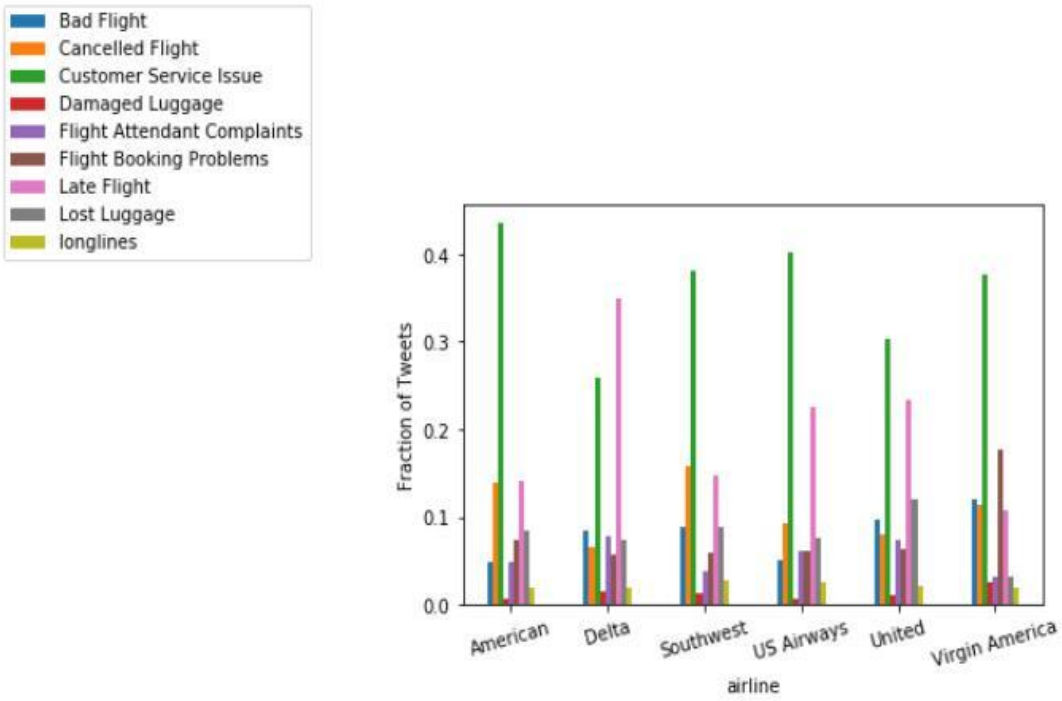
Sentiments by Airline: This plot shows the percentage of Sentiment by Airline

	Airline Sentiment		
Airline Name	**Negative**	**Neutral**	**Positive**
American	0.710402	0.167814	0.121783
Delta	0.429793	0.325383	0.244824
Southwest	0.490083	0.274380	0.235537
US Airways	0.776862	0.130793	0.092345
United	0.688906	0.182365	0.128728
Virgin America	0.359127	0.339286	0.301587



!

Negative Sentiments for Airline: This plot shows the percentage of Negative Sentiments for Airline.



!

First Negative Sentiment reason common in all companies is Customer Service Issue.

Second Negative Sentiment reason common in all companies is Late Flight.

The third most frequent Negative Sentiment reason is divided between two:

- Cancelled Flight.
- Bad Flight.

5.4. WordCloud

The main paragraphs that appear in the tweets, in order of highest to lowest percentage, are:

-flight
-bag
-time

A WordCloud has been made for each sentiment analyzed. The objective is to be able to visualize which are the main words in each sentiment.

5.4.1. WordCloud Negative Sentiment

Note that the main words that appear in the WordCloud Negative Sentiment are:

-flight
-bag
-hour
-time
-plane

5.4.2. WordCloud Positive Sentiment

Note that the main words that appear in the WordCloud Positive Sentiment are:

-thank

-flight

-great

-time

-good

5.4.3. WordCloud Neutral Sentiment

Note that the main words that appear in the WordCloud Neutral Sentiment are:

-flight

-thank

-need

-help

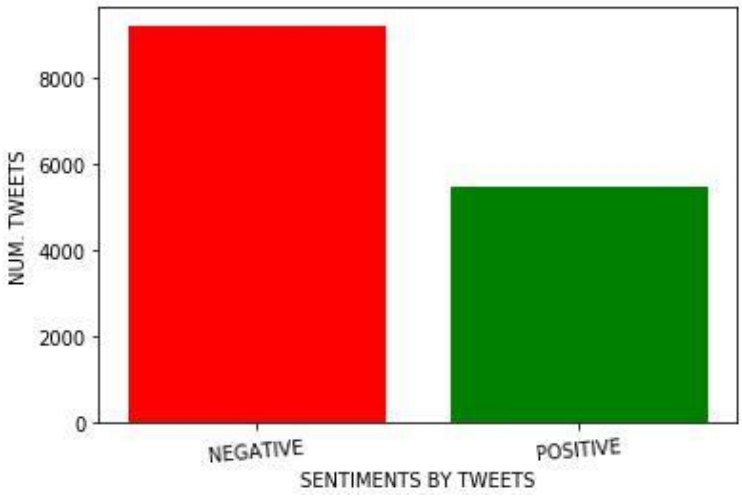
-please

-time

5.5. Positive and Negative Analysis

In this work we are interested in comparing negative sentiment, that are the most common, with another opposite polarity. In the section above it is observed that the neutral and positive sentiments contain mostly the same type of words. There is a considerable imbalance between the sentiments classified as negative and the other two sentiments (neutral and positive). Moreover, when performing supervised classification with machine learning algorithms a recommended practice is to work with a balanced classification dataset. Therefore, after join positive and neutral sentiments the result of plot Sentiments by Tweets change. This plot shows the number of Tweets by Sentiment:

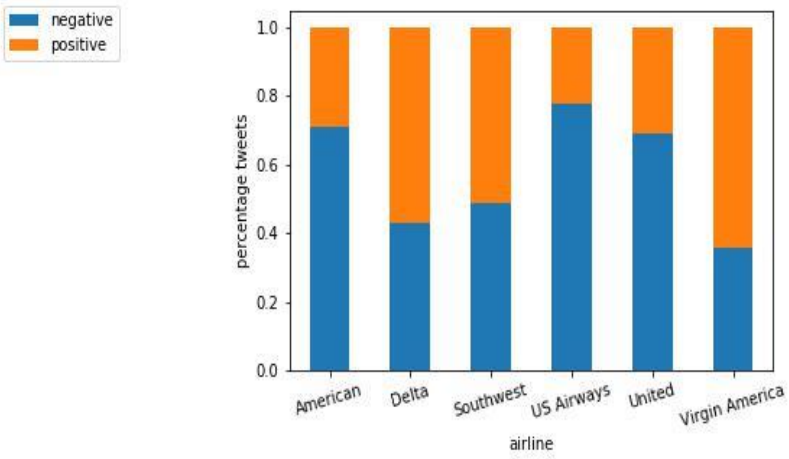
Negative: 9178
Positive: 5462



!

Sentiments by Airline: This plot shows the percentage of Sentiment by Airline.

	Airline Sentiment	
Airline Name	**Negative**	**Positive**
American	0.710402	0.289598
Delta	0.429793	0.570207
Southwest	0.490083	0.509917
US Airways	0.776862	0.223138
United	0.688906	0.311094
Virgin America	0.359127	0.640873



!

Three companies, Delta, Southwest and Virgin America, are observed that are equal or below 50% in negative comments.

5.6. Evaluation of Supervised Learning Models.

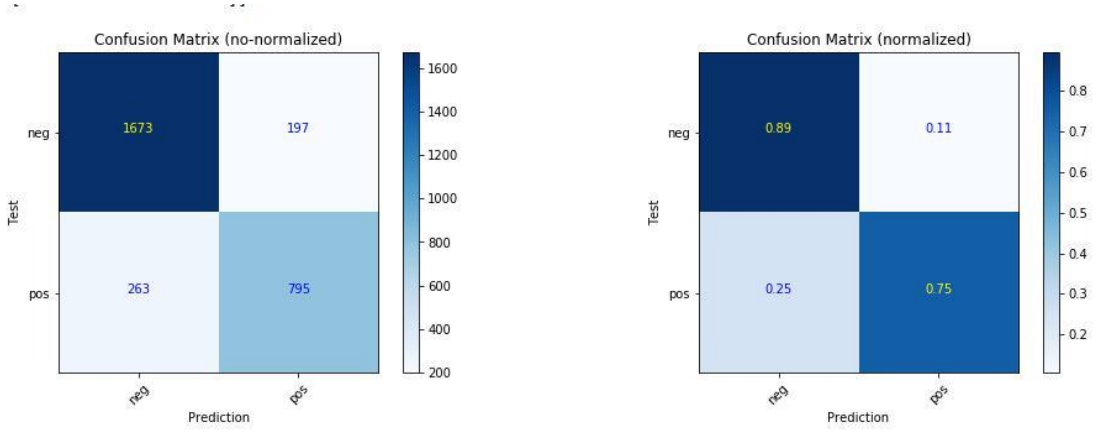
A comparison of different models of machine learning is carried out: SVM, Logistic Regression, Random Forest.

The purpose of this section is to analyze the accuracy, precision, recall and F1-Score to determine which classifier is the most appropriate. To test the validity of our models, the dataset has been separated into a training set and a test set.

	Values obtained Models vs Test set			
Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
SVM	0.84	0.84	0.84	0.84
Logistic Regression	0.83	0.83	0.84	0.83
Random Forest	0.8	0.81	0.81	0.8

SVM and Logistic Regression have the best result. However, it's necessary to review the matrix confusionn to decide the best solution.

5.6.1. SVM Matrix Confusion:



!

Confusion matrix for SVM:

- Prediction (Total Negatives): 1936.

1673 have been predicted correctly.

- Prediction(Total Positive): 992.

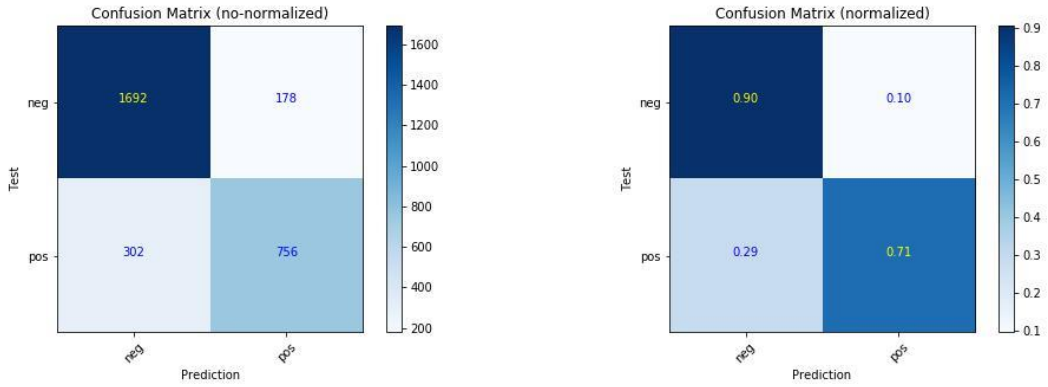
795 have been predicted correctly.

-The Accuracy, (Pos+Neg) / Total, in this case corresponds to:

(1673+795) / 2928 = 2.468/2.928 = 0.8428

84% have been correctly classified.

5.6.2. Logistic Regression Matrix Confusion:



!

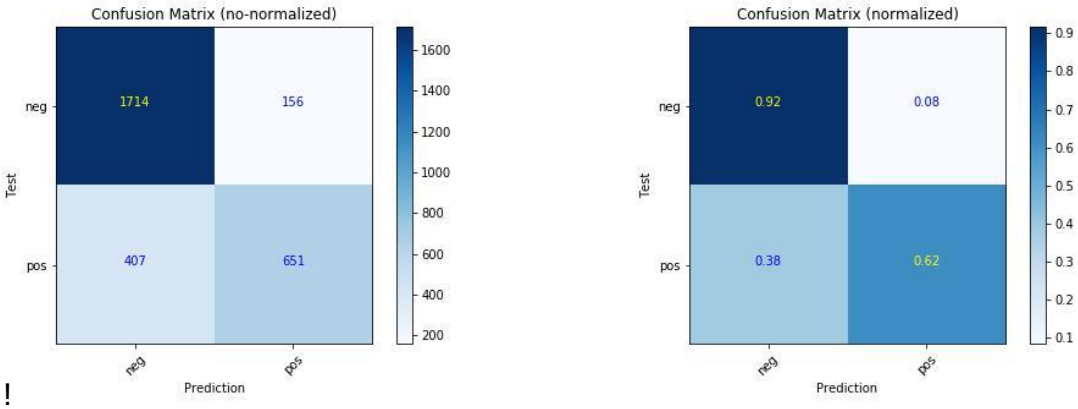
- Prediction (Total Negatives): 1994.

  1692 have been predicted correctly.

- Prediction (Total Positives): 934.

  756 have been predicted correctly.

5.6.3. Random Forest Matrix Confusion:



!

- Prediction (Total Negatives):

  2121. 1714 Have been successful.

- Prediction (Total Positives): 807.

  651 Have been successful.

6. Conclusions

Sentiment Analysis is the most common text classification tool that analyses an incoming message and tells whether the underlying sentiment is positive, negative or neutral. Therefore, sentiment analysis allows us to identify and extract subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service. In this project, we propose to create a text classification model capable of classifying a given text taking into account its polarity. We decided to make the distinction between positive polarity and negative polarity, grouping the neutral polarity and the positive polarity as a single polarity. Therefore, the main objective of this project is to analyze and create a model that predicts, through the use of different classifiers of Machine Learning, the popularity, in Twitter, of different competing airlines: America, Delta, Southwest, United, US Airways and Virgin America. Therefore, the main purpose of this work has been to carry out text processing and sentiment analysis of Twitter data.

To create a classification model, a preprocessing of the text (tweets) has been carried out and the tf-idf model has been applied to extract features and convert textual data into a numeric form. tf-idf, short is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The tf–idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general.

After training the classification model, its correct operation has been verified with the test set. A comparison of different classifiers of machine learning is carried out: SVM, Logistic Regression, Random Forest. After the accuracy, precision, recall and F1-Score has been analyzed to determine which is the most appropriate classifier.

Finally, it has been possible to verify that the classifier that has provided the best results (0.84 accuracy, recall, and F1) has been the SVM classifier,

however, the Logistic Regression classifier has also provided good classification results. Likewise, as a possible improvement of the project a deep learning model could be proposed, however, it has not been possible to apply in this project due lack of time.Natural Language Processing with Deep Learning in Python would have been the approach to obtain better results. When not being able to make a statement at a functional level for its later development, it has not been possible to carry out the notebooks with classifiers by applying Deep Learning. By applying deep learning in our project on Natural Language Process, we will find the following advantages:

- Word2vec is interesting because it magically maps words to a vector space where you can find analogies.

- Recursive Neural Networks, can help us solve the problem of negation in sentiment analysis. Recursive neural networks exploit the fact that sentences have a tree structure, and we can finally get away from naively using bag-of-words.

7. Tableau

This section shows us the links to view through dashboards, how sentiment analysis has been applied over 5 different American airlines and their corresponding polarity. The links are the following:

Tweets Airline Sentiment vs. Airline

https://public.tableau.com/profile/raul3410#!/vizhome/Num_AirlineSentimentVs_Airline/Num_AirlineSentimentvs_Airline?publish=yes

Negative Analysis

https://public.tableau.com/profile/raul3410#!/vizhome/NegativeTweetAnalysis/Dashboard1?publish=yes

Records and Reasons by Sentiment

https://public.tableau.com/profile/raul3410#!/vizhome/Records_Reasons_by_Sentiment/RecordsandReasonsbySentiment?publish=yes

Number of Retweets ~ Airline Sentiment by Airline

https://public.tableau.com/profile/raul3410#!/vizhome/TweetsvsRetweets2/Num_ofRetweets_AirlineSentimentbyAirline?publish=yes

Tweets vs Retweets ~ Airline Sentiment by Airline

https://public.tableau.com/profile/raul3410#!/vizhome/TweetsvsRetweets3/TweetsvsRetweets?publish=yes

8. Bibliography

Hernández Farías, D.I., Patti, V., y Rosso, P. (2016). Irony Detection in Twitter: The Role of Affective Content. ACM Transactions on Internet Technology, 16(3), 1-24. DOI: http://dx.doi.org/10.1145/2930663

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167. DOI:https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Reyes, A., Rosso, P., y Veale, T. (marzo, 2013). A multidimensional approach for detecting irony in Twitter. Language Resources & Evaluation, 47(1), 239-268. DOI: http://dx.doi.org/10.1007/s10579-012-9196-x

Thakkar H, Patel D. Approaches for sentiment analysis on twitter: A state-of-art study. arXiv preprint arXiv:1512.01043. Accessed 3 Dec 2015.

9. Global Work Planning

The following table details the key points that have been carried out for the realization of the TFM called Twitter Sentiment Analysis. The planning takes place from April 01, 2019 to June 15, 2019. The main points dealt with, which can be seen broken down by sub-tasks in the box below, are:

-Initial Proposal Review.
-Bibliography, State of the Art.
-First Analysis.
-Functional Solution.
-Final Master's Writing
-TFM delivery.
-Defense of the TFM.