Trend Analysis in Aerospace Industry using Twitter

Project – Online Social Network Analysis

Goutham Kannan
A20361163

12/2/2015

Introduction:

The exponential upsurge in the usage of Social Media Platforms such as Twitter, Facebook by people across the globe has made the Technological and Engineering companies, to use it as a medium to market their existing products, advertise a new product, talk about their unique technologies, bully competitors by show casing the technological might, engage with people and customers, broadcast job openings etc. The main idea behind this project is to use trap this potential user data and get a meaningful information out of it. Finding the trends in Engineering and Technological field is very difficult; mostly it can only be opined by the experts in that field. How to get this information to common man? This is the main problem that will be addressed in this project. The Tweets from major Aerospace and Aviation players are collected and potential trends are to be identified from it. Finding the trends is not as simple as tracking the hash tag trending words.

Hypotheses: Higher the number of tweets classified under a class in given set of tweets, then that particular class is highly popular and trending.

Data Collection:

In order to do any data analysis, valid datasets are very crucial. So we have to collect trustworthy data to have a better prediction. The data collected is from the Twitter pages of Top Aerospace companies and Regulators (i.e. Verified twitter handles). They are FAANews – Federal Aviation Agency, AirbusGroup – Airbus, Boeing , GulfstreamAero- Gulfstream, LockheedMartin ,BoeingAirplanes ,icao - International Civil Aviation Organization, EASA – European Aviation Safety Agency, Bombardier, thalesgroup – Thales, GEAviation – General Electric Aviation, Honeywell_Aero- Honeywell Aerospace, UTC –United Technologies, RockwellCollins, finmeccanicaweb – Finmeccanica, Raytheon, BAESystemsAir – BAESystems.

Twitter API has a field USER_TIMELINE, using this we can collect up to 3200 public tweets per user handle, at the rate of 300 tweets per request and the next request can be send after 15 minutes from the previous request. So used a function in python to request data from Twitter, in case of a rate error we wait for 15 minutes and then send a request again and exit after the 16(16*200=3200) pages of data is collected. The collected data is stored in the form of '.txt' files and used pickle serialization to save the data for further usage.

Methods:

1. Cleaning the data from the collected data, tweets that are not in English are cleaned for better classification.

2. Manually labeled the data in to 9 categories, The nine categories are “Accidents and Safety”, ”Careers”, ”Cost Effectiveness”, ”Military and Defense”, “Drones”, “Space Research”, “Aircraft System”, “Other/Miscellaneous”. These labels were chosen based on the recommendation from https://www.asme.org/engineering-topics/articles/aerospace-defense/top-5-aerospace-trends-now-future ; ASME is the American Society of Mechanical Engineers.

3. Using the tokenized words of the labeled data, created a feature matrix, during the tokenization the URLS/Links were removed to keep the degrees of freedom in feature matrix in check.

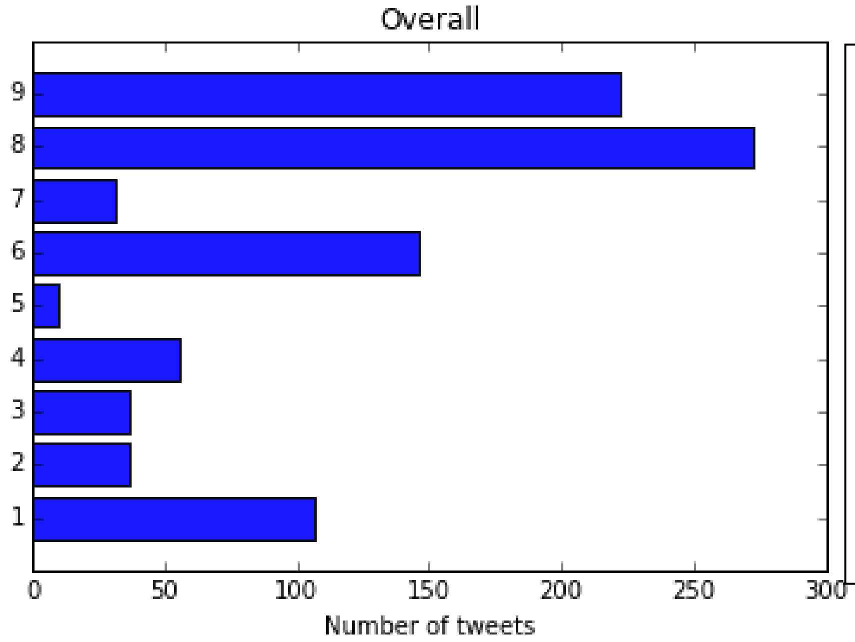
4. Fitted the model using the Features matrix for all labeled tweets and the labels, and the accuracy of the training data is found using Cross Validation.

5. In order to improve the accuracy, the several experiments like varying the min_df, max_df, regularization and the corresponding coefficient. After several experiments it was found that K=10, min_df = 2 and max_df = 0.6 and L2 Regularization is chosen to limit the features’ degree of freedom with a co-efficient of 1. Using this model setting we get an accuracy of 58% for a training data set of 918 tweets.

6. Latest 54 tweets from all handles that sums up to 918 tweets are taken as testing data. We predict the classes for this new data using the trained model.

7. When Logistic Regression is used as the classifier we get an accuracy of 61% and Multinomial Naive Bayes classifier predicts with an accuracy of 71%.This accuracy is measured using a small labeled set of 100 tweets from the test data. These labels are compared with the predicted labels to get the accuracy value.

8. From the classified data a graph is drawn depicting, “How many Tweets where classified into each category”



Class 9: Others

Class 8: Aircraft Systems

Class 7: Space

Class 6: Marketing

Class 5: Drones

Class 4: Military and Defense

Class 3: Cost/Money

Class 2: Careers

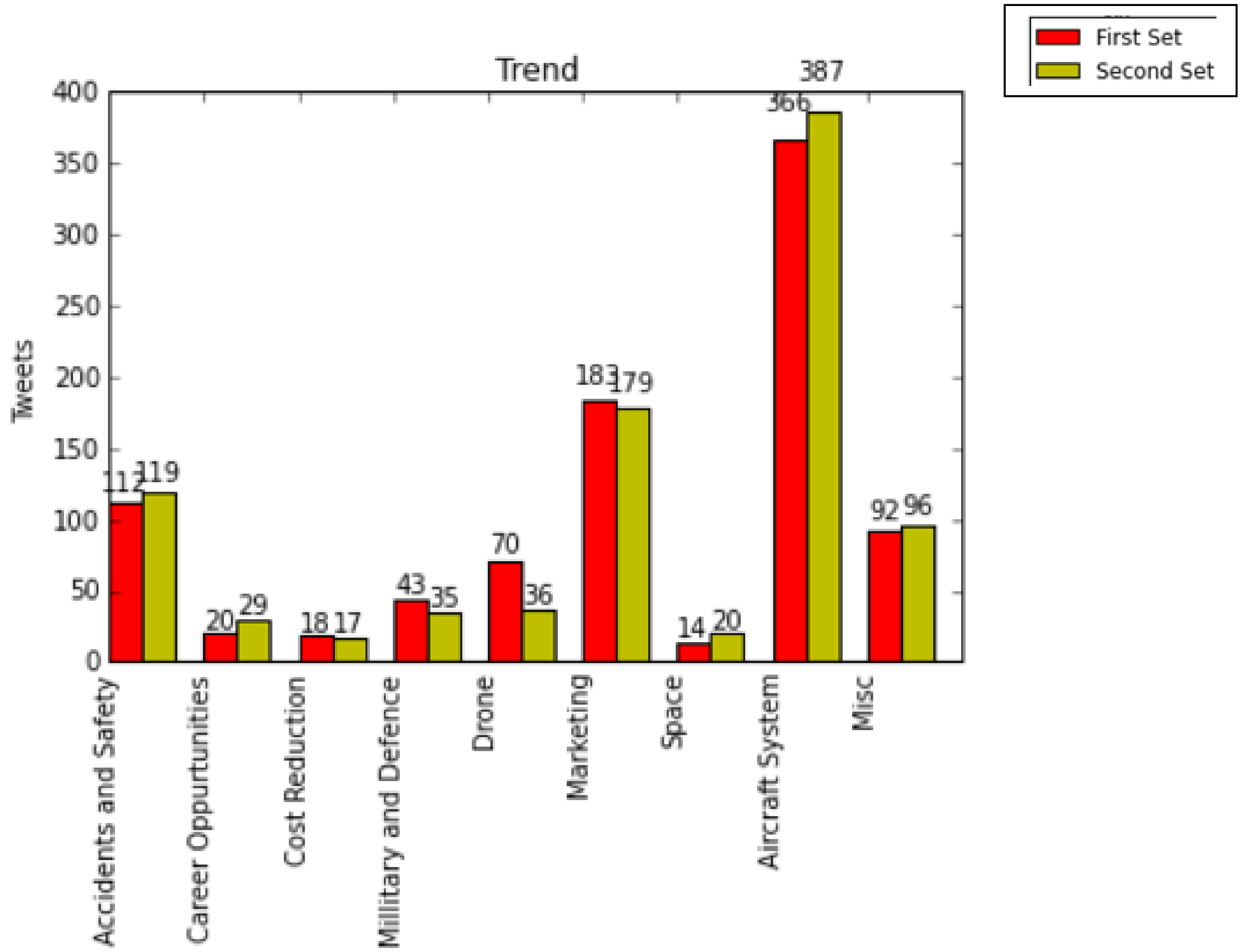
Class 1: Accidents and Safety

From this distribution of tweets we can see that maximum number of tweets was classified into the category of Aircraft Systems which as per expectations as an Aircraft has more than 1000 systems. As flying involves a lot of risk, safety is paramount in the field of Aerospace. These two are the trending fields in the industry.

The fields like Drones, Space research has very limited share in the tweet space this indicates that they are not in trend and they can be emerging field or obsolete fields. The hypothesis under which this project doesn’t help to differentiate between the downward or upward trends. In order find the direction of the trends we need to classify more data based on a timeline (Say yearly or half yearly).

9. Predict the classes for two set of tweets set2 latest 918 tweets and set 1 - immediate past 918 tweets**.**

Trend Graph:

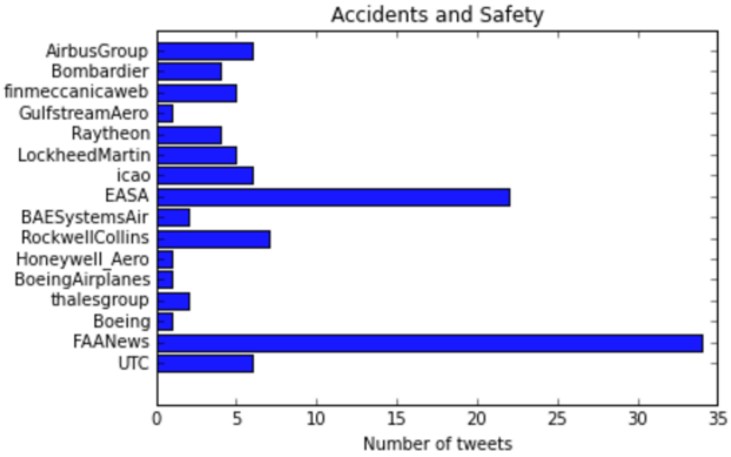


From above graph we can see very slight increase and decrease trend in each segment, Accidents and Safety (112 and 119), Career Opportunities (20 and 29), Marketing (183 and 179), Space (14, 20). As time difference between set 1 and set 2 is less than 2 months, we can't see much variations in the result. When the same analysis is done on data with difference in time frame of 4 or 6 months, then we will be able to visualize the growth or decline of every segment.

Claims to support the Experiment results

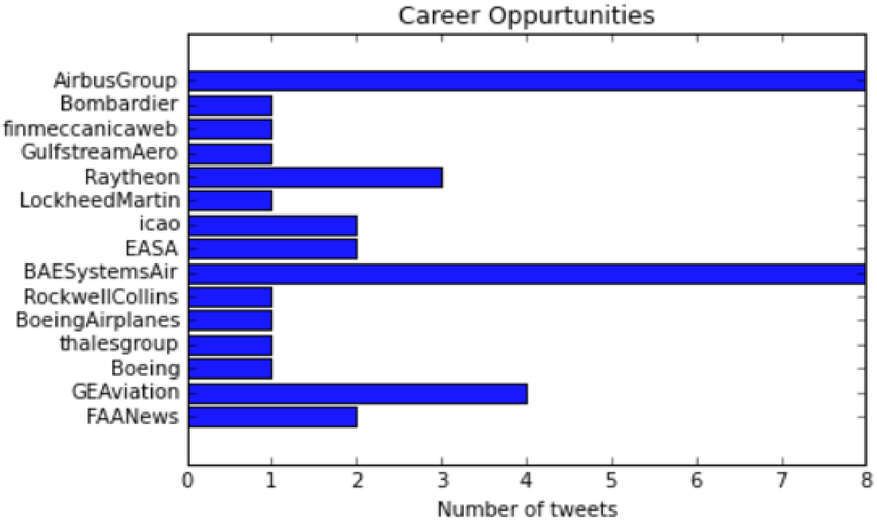
In order prove the effectiveness of the classifier , distribution of tweets classified for every 'Class' is plotted , Y-axis denotes the companies whose tweets belong to that class , X-axis is the number of tweets.

Graph 1: Accidents and Safety



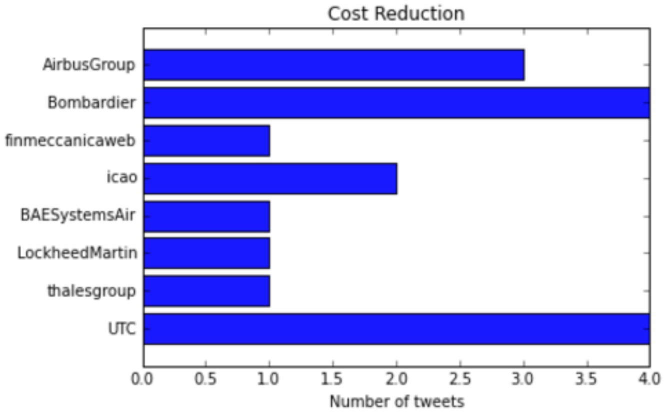
Classifier has predicted more tweets from FAA (Federal Avionics Agency: https://www.faa.gov/) and EASA (European Avionics Safety Agency: https://www.faa.gov/). FAA and EASA are the top safety regulators in the Aviation Industry.

Graph 2: Careers



From the Career website of Airbus we can see that more than 187 job openings were posted in last few months and the same was posted in social media Airbus Career:http://www.airbusgroup.com/int/en/people-careers/jobs-and-applications/search-for-vacancies.html?queryStr=&country=gb&country=us

Graph 3: Cost and Money Related

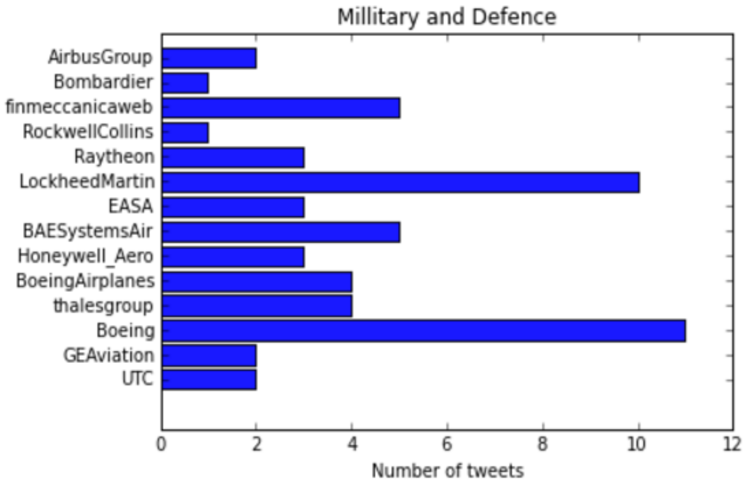


The second most prominent in this regard in Bombardier, we can see from the companies press release than, recently they have released their "Strategic and Transformation Objectives through 2020" and acquisition of a smaller company. So the tweets were classified under the Cost/Money category.

Press Release: http://ir.bombardier.com/en/press-releases/press-releases

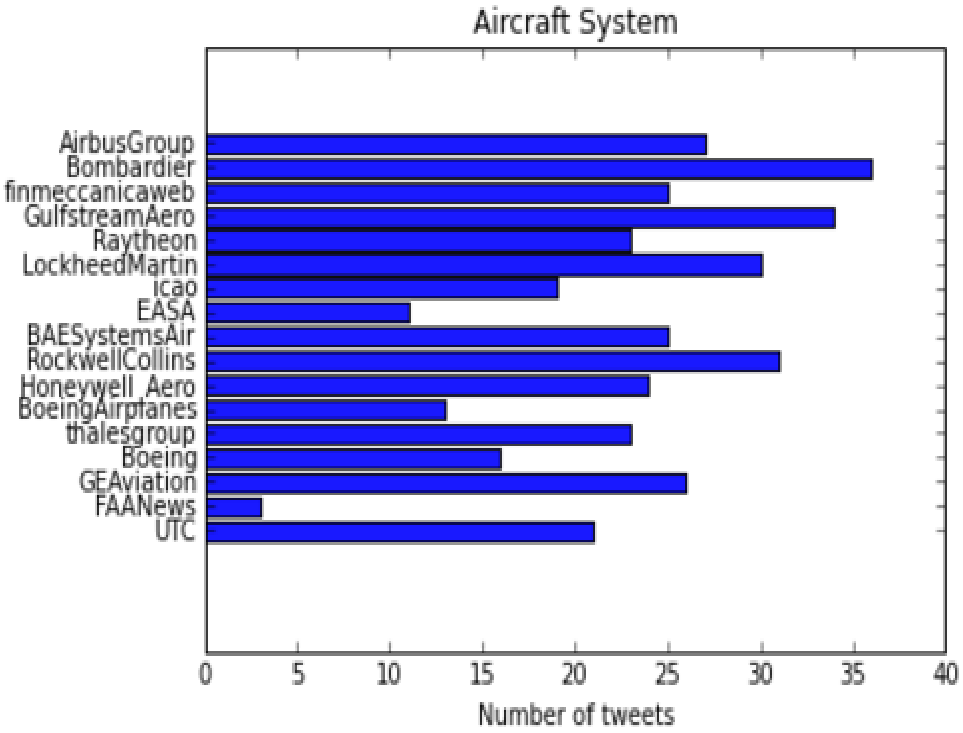
Graph 4: Military and Defense

The top 2 military aircraft manufactures in US are Lockheed Martin and Boeing, The same has been reflected in the classification of tweets.



List of Manufactures:
http://web.stanford.edu/class/e297a/U.S.%20Defense%20Industry%20and%20Arms%20Sales.htm

Graph 7: Aircraft Systems



From the graph we can see that two regulators FAA and EASA (Non manufactures) have the least amount of graphs classified under them and all top Aircraft System manufactures have higher tweets under their name.

Case where Hypotheses fails

In the hypotheses it said that more number of tweets signifies about a field then the field is trending and popular but the sentiments of the tweets are not considered. For Example if the tweet is “Drone are not the good for deserts” – this tweet is intended to say that Drones are not effective but in this project the sentiments are not considered we just classify this tweet under the Drone label.

Related work

Traditionally, the forecast of trends in any field is done by measuring the growth of Patents and Research paper submission. Mostly textual analysis is done on the amount of papers and patents in a particular field from then on there are two ways of forecasting. First, Experts in the field of study are gathered and they take decisions regarding the growth of a particular filed. Second, is by modeling and simulation the trends are plotted and they are extrapolated to make the predictions. The second approach is a data intensive approach. The paper titled,” Forecasting with twitter data, ACM TRANSACTIONS ON INTELLIGENT SYSTEMS AND TECHNOLOGY 5(1) · NOVEMBER” [1] explains how twitter data can be used to predict the movie rating or the trends in stock market. In this project a similar approach is used to find the trends in the field of Aerospace and Aviation.

Conclusions and Future Work:

To conclude, we can confidently say that mining data from companies' social media pages can be used to measure and predict the growth of technologies. But in order to clearly visualize the trends this analysis should done on quarterly or half yearly basis.

Another important revelation is that most of the companies use Social Media as a medium to broadcast the job openings and internship avenues.

This trend analysis can be done can on a time scale like quarterly / Half Yearly and based on the trend graph future trends can be forecasted using methods like Delphi method, Trend Extrapolation, Gompertz and Fisher-Pry Substitution Analysis.[2]

Another improvement in the existing project would be find the clusters who are all working on the same products at a given time, this can be achieved using the tweet created time, mentions field and by parsing tweets to find the involvement of other companies in their tweets.

References

1. Forecasting with twitter data, ACM TRANSACTIONS ON INTELLIGENT SYSTEMS AND TECHNOLOGY 5(1) · NOVEMBER"

2. Existing Technology Forecasting Methodologies, http://www.nap.edu/read/12557/chapter/4