Twitter US Airline Sentiment Analysis

Dexiang Xu Gaoyang Ye Liang Shi Mengheng Hu Yue Sun

dxx140230 gxy140830 lxs143430 mxh144730 yxs146930

Abstract

Twitter is a popular social net site where people often express their emotions or record their life moments. On the other side, Airline company provide flight services to their passengers whose feedback means a lot. Therefore, twitters including passengers' sentiments or comments are especially valuable for the airline companies, since airline companies would be able to know what their passengers' sentiments about the flights and what else airline companies can do to improve their services.

I. Design

Summary of problem definition

Our project focuses on analyzing how travelers in February 2015 expressed their feelings on Twitter. Machine learning analysis based on processed data is also involved, as well as some data mining from raw data set.

The whole project is composed of two parts:

Part 1: Original data analysis. It consists of two tasks.

Task 1:

1. Overall sentiment (negative, neutral, positive percentage)
2. Distribution of tweets per airline
3. Proportion of negative/neutral positive sentiment tweets per airline
4. Reasons for negative sentiment tweets
5. Reasons for negative sentiment per airline
6. Tweet location exploration
7. Tweet time zone and location study

Task 2:

1. Find top 300 negative words in the negative comments in decreasing order
2. Find top 300 positive words in the positive comments in decreasing order
3. Choose 5-10 from top negative words, find their top 100 co-occurrence words in negative comments.
4. Choose 5-10 from top positive words, find their top 100 co-occurrence words in the positive words

Part 2: Machine Learning analysis. It also consists of two tasks.

Task 1:

1. Preprocess the class label of original data set, which is 3 class labels including positive, neutral and negative and compare the accuracy of different models.

2. preprocess the class label of original data set, which is 12 class labels this time including positive, neutral and 10 different kinds of negative reasons. Compare the accuracy of different models and the accuracy of previous part.

Task 2:

Build different models to predict whether a tweet is positive, neutral or negative, and if it is negative, give the type of the reason of it being negative.

Our assumptions include: 1. All the twitters are true reflections of passengers' sentiments. 2. The information we got from those twitters really means something for the airline companies.

Our limitations include: 1. Only twitters in February 2015, and if more data, it'll be better. 2. Only 10 reasons for the negative attitude, but in fact, there are more reasons which will lead to negative comments.

Description of input data

Part 1

Tweets.csv:

Original data for our project.
This file contains 15 attributes: tweet_id, airline_sentiment, airline_sentiment_confidence, Negativereason, negativereason_confidence, airline, airline_sentiment_gold, name, negativereason_gold, retweet_count, text, tweet_coord, tweet_coord, tweet_location, user_timezone.

Tweets_Updata.csv:

Data after some process including remove of unformatted or null values.

Part 2

wordbag.txt:

Negative word bags extract from tweets.csv, it is for machine learning analysis purpose.
Below are the steps to get the wordbag:

Step 1:

Go through all the data, and create a dictionary which contains all the words appeals in the twitters, and save the dictionary as W.

Step 2:

Go through the data again, for each tweet, count the words frequency in the order of dictionary W. Therefore, after go through all tweets, we would get a M*N matrix which M is the number of all tweets and N is the length of W.

Step 3:

Use this matrix as our training data.

Other data used in this project are in the input_data folder within the source code report.

Big Data strategy and algorithm

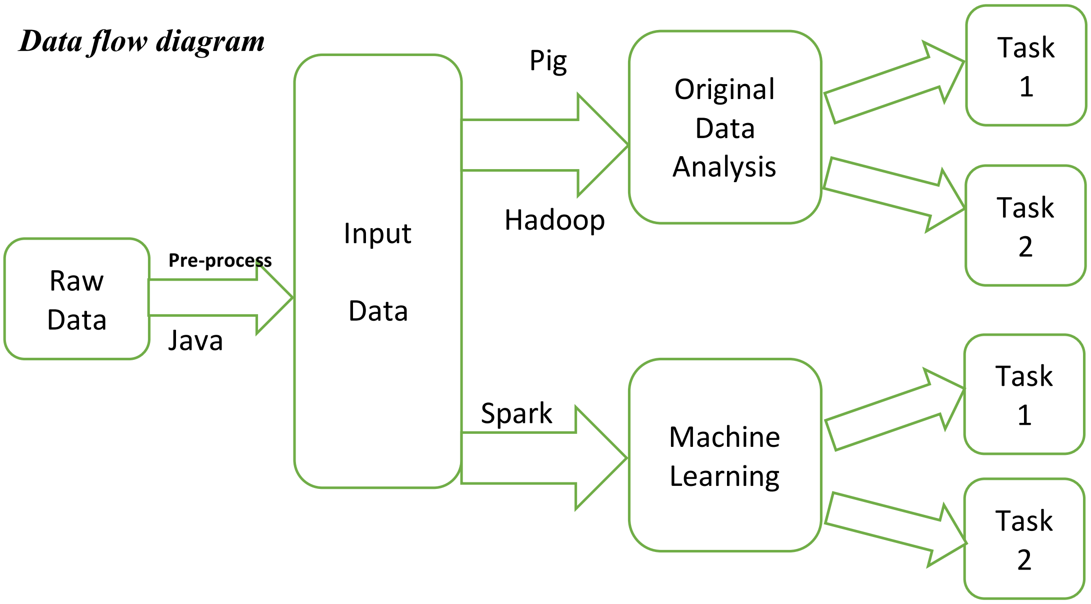
For **Part1_Task1**, we use Apache Pig to retrieve information from the data set, and Tableau to visualize the result.

For **Part1_Task2**, we use Hadoop MapReduce to retrieve information from the data set, and Tableau to visualize the result.

For **Part2_Task1**, we use Spark and spark.mllib package for machining learning. Algorithms used here contain Logistic Regression(LR), Decision Trees(DT), Random Forests(RF) and Naïve Bayes(NB). Preprocess the class label of original data set, which is 3 class labels including positive, neutral and negative and compare the accuracy of different models. Then, preprocess the class label of original data set, which is 12 class labels this time including positive, neutral and 10 different kinds of negative reasons. Compare the accuracy of different models and the accuracy of previous part.

For **Part2_Task2**, we use scikit-learn package(sklearn.neural_network) and Spark Multilayer perceptron classifier to compare the result of neural network. And we use naiveBayesClassifier package and MLlib Naïve Bayes to compare the result of Naïve Bayes. Algorithms here include Naïve Bayes(NB) and Artificial Neural Network(ANN)

Data flow diagram



Strategy Robust

1. Set all the missing data in the data to null value, such that the system can detect them correctly.

2. Avoid overfitting:

It mainly aims at Machine Learning part. For neural network, the biggest problem is overfitting. To avoid overfitting problem, we use python scikit-learn package to compare the accuracy with Spark Multilayer perceptron classifier. The reason we can do this test is scikit-learn package is scikit-learn package has some inner mechanism to avoid overfitting. Once the accuracy of Spark Multilayer perceptron classifier has the close result to scikit-learn package, we can assume that it has no overfitting in the model. As for Naïve Bayes, there is no problem for overfitting.

II. Analysis of Results

Part 1, Task 1:

1. Overall sentiment analysis.

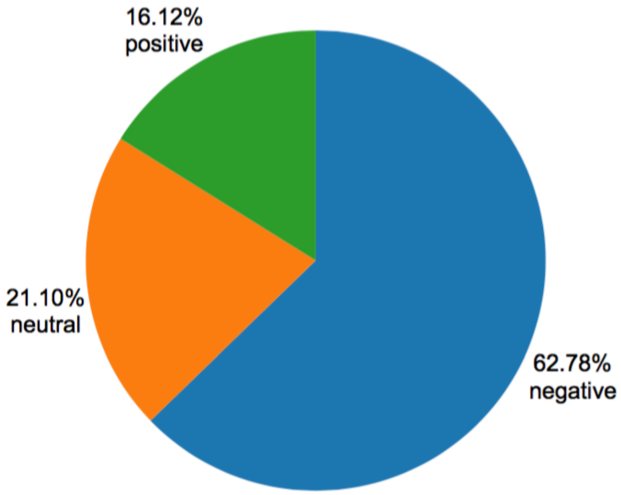


Figure1. Overall sentiment

We can see from above that negative tweets take the lead which has 62.78% of the total tweets.

2. Percentage of tweets per airline.

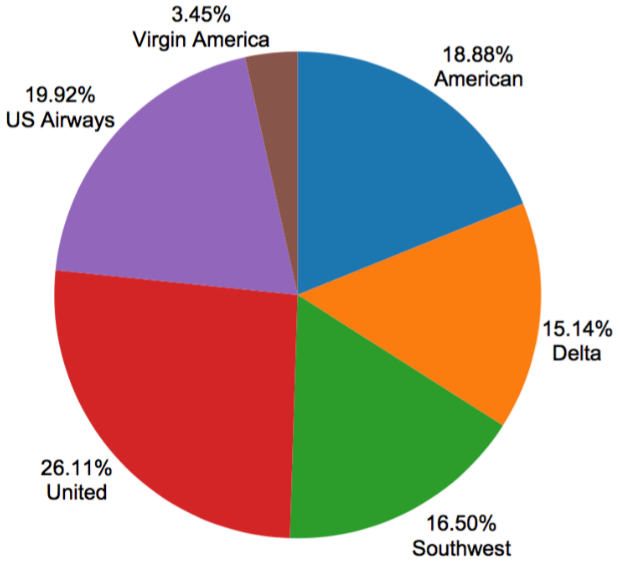


Figure2. Percentage of tweets per airline

from the chart above we can see that United Airline has the most tweets regarding their services, US Airways takes the second place which has 20% tweets, and then American Airline in the third place with 19% tweets

3. Proportion of negative, neutral positive sentiment tweets per airline.

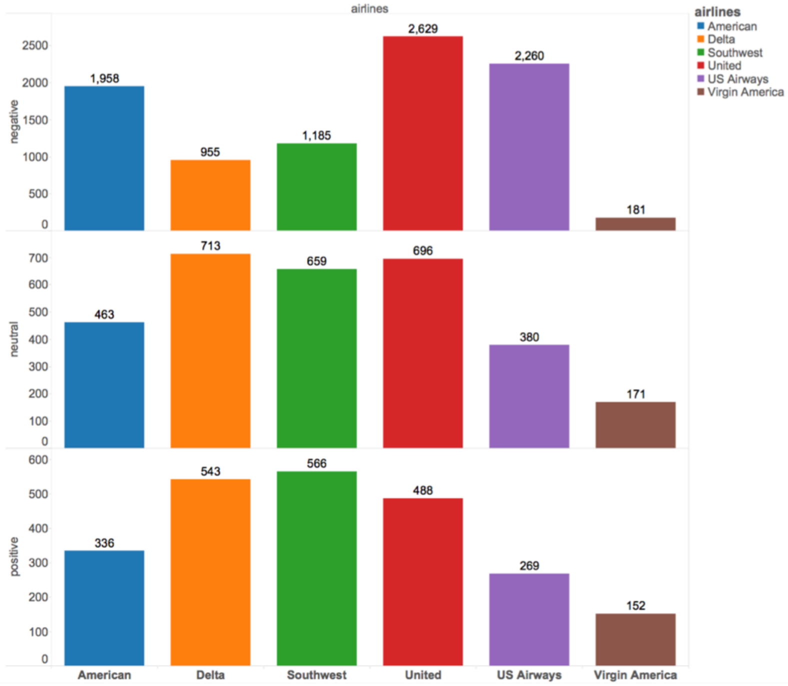


Figure3. Proportion of negative, neutral positive sentiment tweets per airline

Observing from the output, we can see that United Airline received the highest negative tweets which has 2629 out of 14500 tweets about their services, and Southwest has the highest positive tweets.

Also, an interesting point is that all companies have negative tweets more than positive tweets we can see graph below

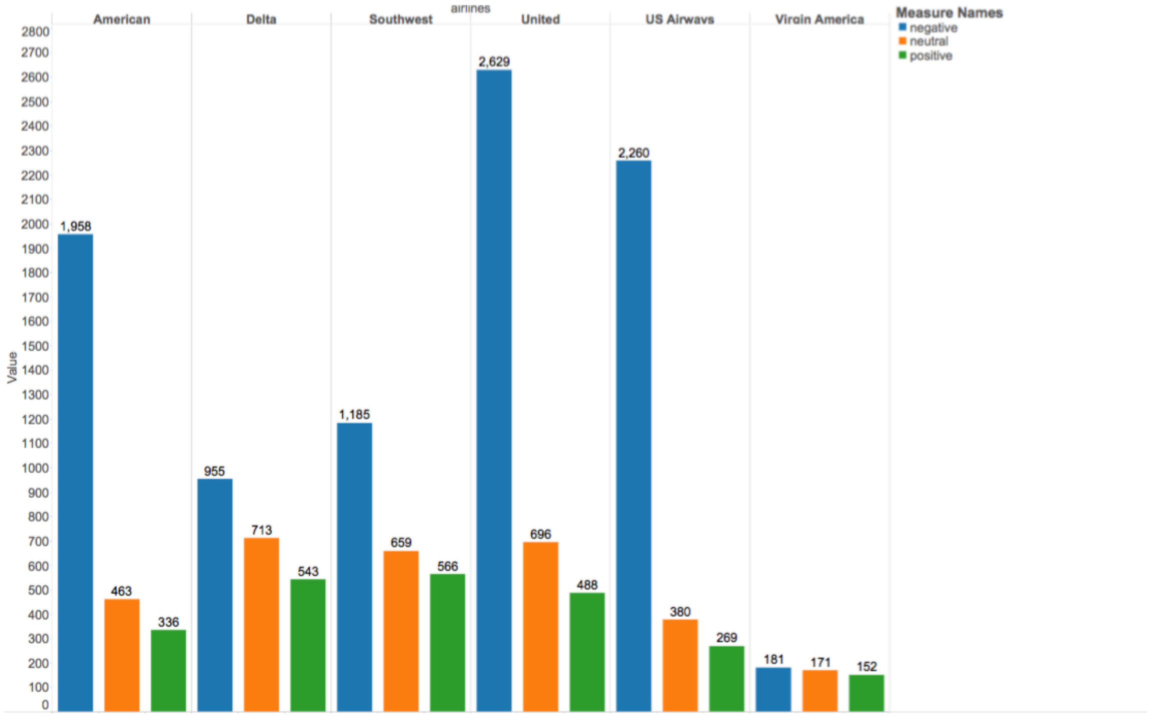


Figure4. Proportion of negative, neutral positive sentiment tweets per airline

4. Negative reasons

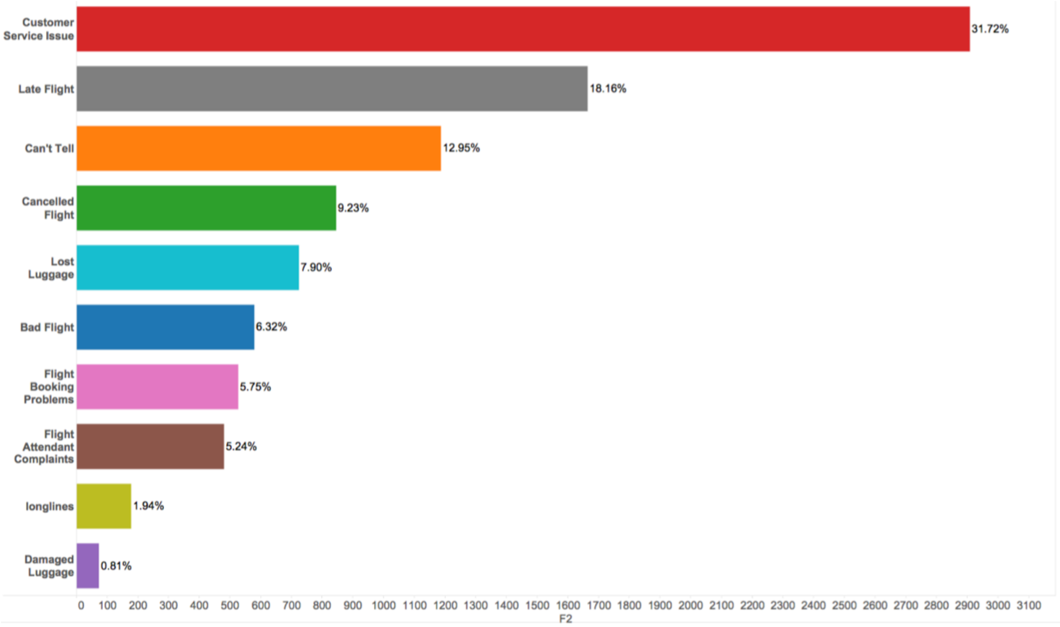
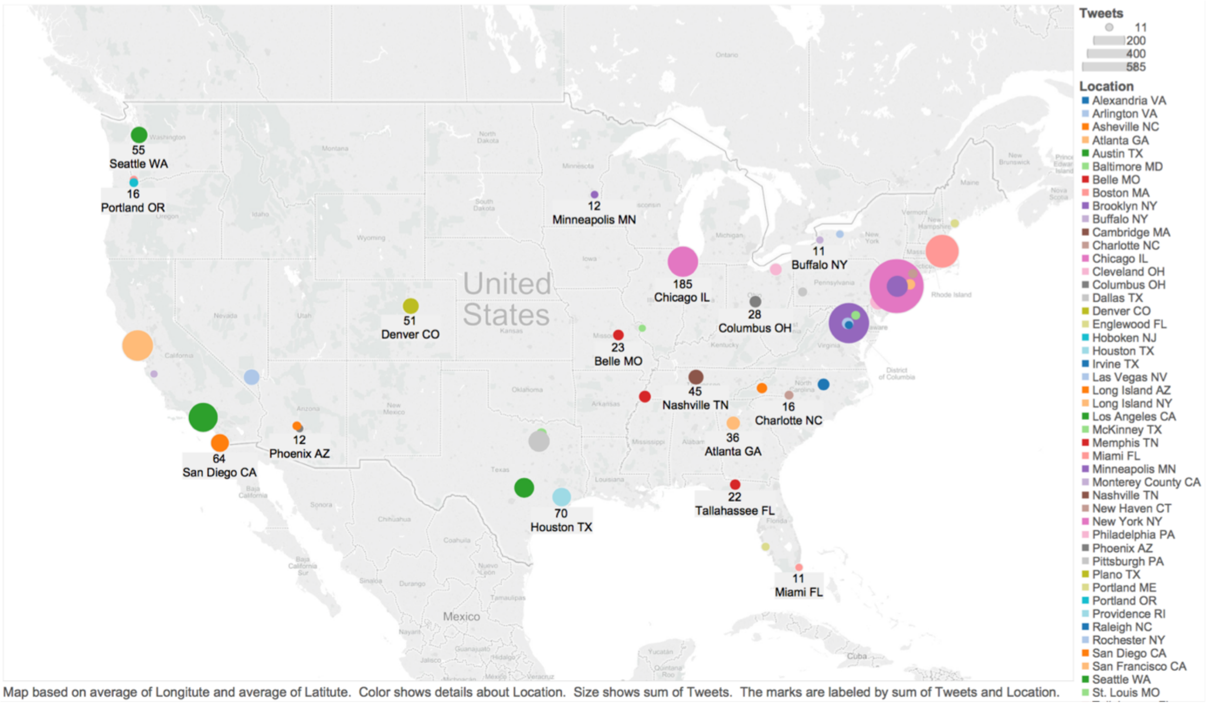


Figure5. Negative reasons

We can see from above that most tweets are about customer service issue, and light delay is the second important negative reason.

5. Location Study



Map based on average of Longitute and average of Latitute. Color shows details about Location. Size shows sum of Tweets. The marks are labeled by sum of Tweets and Location.

Observing from the map above we can see that most tweets are coming from east and west region, and Chicago is another region has most tweets. Airline companies should focus on flights between west and east.

Part 1, Task 2
Negative:

	Top 1		Top 2		Top 3		Top 4		Top 5	
customer	service	538	United	289	For	260	USAirways	241	AmericanAir	224
delayed	Flight	258	United	163	My	135	USAirways	123	JetBlue	75
flight	United	886	Cancelled	747	USAirways	745	AmericanAir	690	SouthwestAir	589
help	Air	230	USAirway	214	United	207	Flight	202	Helpful	202
hold	Airways	275	Flight	170	Been	179	Cancelled	120	Hour	109
service	Customer	538	United	307	USAirways	259	You	229	Flight	156
time	United	259	Flight	232	USAirway	218	My	217	AmericanAir	168

Positive

	Top 1		Top 2		Top 3		Top 4		Top 5	
best	Southwestair	52	Jetblue	37	My	37	Flight	29	Airline	28
customer	Service	400	United	263	For	260	You	239	Your	199
great	Flight	61	Jetblue	55	Southwestair	52	Airways	47	United	46
love	Southwest	83	Jetblue	56	Would	38	Me	35	Airline	35
thank	Thanks	553	United	213	Airline	216	JetBlue	153	USAirways	141
thanks	United	134	Air	117	JetBlue	114	Service	34	Time	26
service	Customer	538	United	307	Usairways	259	You	229	Americanairline	225

For above tables, first column is top 10 words from either positive or negative top words which make sense. And right 5 columns represent co-occurrence words.

Analysis:
(1) We find the jetblue and Southwest air has happened to appear quite frequently with positive top words, specifically for southwest air, it has come to top 5 co-occurrence with best, great, for 52, 52 time. Jetblue is even better, is has come to the top 5 of best, great, love for 27, 55, 46 times. So we know jetblue and southwest air are quite popular with good words, we have a sense of positive feeling of it.
(2) If I am a customer, I would like to choose the southwest and jetblue at the same price and equally time schedule. If I am the worker of airline company, maybe I need to think about what is the difference between us and southwest air company, do we need to consider jetblue?
(3) We suggest do not draw any conclusion further than that, just based independently on this part, this part is serviced at a clue or hint, but not strong evidence, we have machine learning part in the following.

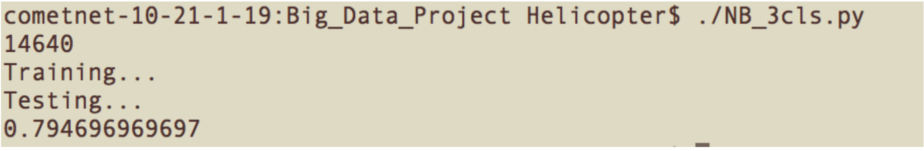
Part 2, Task 1
Accuracy:

	3 class	12 class
Decision trees	0.6207	0.4143
Naïve Bayes	0.6550	0.3785
Random forests	0.6332	0.2869
Logistic regression	0.5252	0.4993

As can be seen in the accuracy, our model has a significant better result in 3 class labels than that of 12 class labels.This is also a common situation, since the probability of random guess for 3 class label data is 0.34, while the probability of random guess for 12 class label data is 0.125. It's hard to get good accuracy for 12 class label data. Because, for the 3 class label situation, it generally has a huge different between tweets which have different sentiments, and it is much easier for both human and computer to label them properly. But for 12 class label situation, even human will have a hard time to label different tweets, especially to classify the reason of negative sentiments. Since a negative Tweets may have complained 2 reasons at the same time. So in the original data set this is a feature named airline sentiment confidence and negative reason confidence, people who created the files cannot perfectly label them. So it is extremely easy for a machine learning algorithm to mismatch negative reasons for an ambiguity tweet. That's why we got a low accuracy towards 12 class label situation.

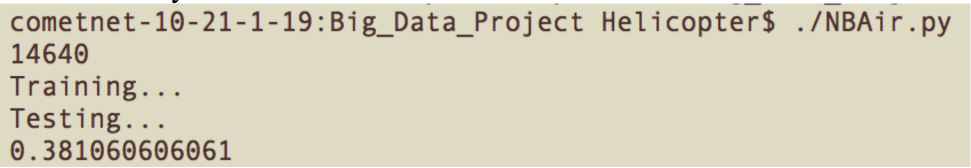
Part 2, Task 2
Accuracy
Naïve Bayes:

If we only predict three classes (Positive, Negative, and Neutral), it could reach the accuracy of 0.794696969697.



```
cometnet-10-21-1-19:Big_Data_Project Helicopter$ ./NB_3cls.py
14640
Training...
Testing...
0.794696969697
```

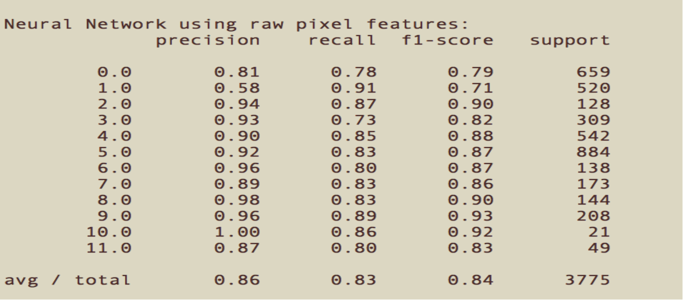
If we only predict all twelve classes (including all classification of different reasons of negative), it could reach the accuracy of 0.381060606061, which is relatively good because some of the different reasons are very similar.



```
cometnet-10-21-1-19:Big_Data_Project Helicopter$ ./NBAir.py
14640
Training...
Testing...
0.381060606061
```

Neural Network:

Neural network could reach the accuracy of 86%, which is the best result we got among all the methods. Because neural network is extremely well performance with multiple classes output.



```
Neural Network using raw pixel features:
             precision    recall  f1-score   support

        0.0       0.81      0.78      0.79       659
        1.0       0.58      0.91      0.71       520
        2.0       0.94      0.87      0.90       128
        3.0       0.93      0.73      0.82       309
        4.0       0.90      0.85      0.88       542
        5.0       0.92      0.83      0.87       884
        6.0       0.96      0.80      0.87       138
        7.0       0.89      0.83      0.86       173
        8.0       0.98      0.83      0.90       144
        9.0       0.96      0.89      0.93       208
       10.0       1.00      0.86      0.92        21
       11.0       0.87      0.80      0.83        49

avg / total       0.86      0.83      0.84      3775
```

And the result seems to be quite acceptable.

III. Conclusion

Big Data is a very popular concept now, it seems that every one know about it, but actually few of them truly know about big data. Through this project, we are shocked by the power of big data technology. Because our data set is quite huge, without tools like Hadoop or Spark, the data analysis work will be hard. For this project, with the help of big data, we not only succeeded in original data analyzing but also got good performance at the machining learning part. We just can’t imagine how time would we take if running this project on local computers. Therefore, big data is efficient.

Through this project, the work like data collecting, data preprocessing, result analyzing and predictions ensure all of us get a deep understanding of big data technology concept and good mastery of big data skills.

As for the improvements, given to the time limitation, even we have used several machining learning algorithms, but maybe we could try more algorithms like BP, CNN and so on. Besides, if we could get more data, we believe more explorations would be done.

IV. Role of Each Team Member

Dexiang Xu: Machine Learning part, task 1.

Gaoyang Ye: Machine Learning part, task 2.

Laing Shi: Part of Original data analysis, and report writing.

Mengheng Hu: Original data analysis, task 1.

Yue Sun: Original data analysis, task 2.

V. Reference

[1] Anurag Nagar, course slides, University of Texas at Dallas, 2016.

[2] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Introduction to Data Mining, Addison-Wesley April 2005.

[3] Jimmy Lin and Chris Dyer, Data-Intensive Text Processing with MapReduce, Morgan & Claypool Publishers, 2010.

[4] Anand Rajaraman and Jeff Ullman, Mining of Massive Datasets, Cambridge Press.

[5] http://en.wikipedia.org/wiki/Machine_learning