**Student Name(s): Aliasghar Arabi, Saeede Sadat Asadi Kakhki, Bernardo Najlis**

**Student Number(s): #500606766, #500774214, #500744793**

**Course Title:  DS8006 – Social Media Analytics**

| Assignment # | Group Project – Social Media Analytics |
|---|---|
| **Due Date** | **Apr-03-2017** |
| **Group #**  (if applicable) | **1** |

**Return all assignments through the Course Assignment Submission Page
(Unless otherwise specified by the instructor).**

**Summary**

In this project we used Twitter to collect observations about six different Canadian Airlines (Air Canada, Porter, WestJet, Jetlines, Sunwing, Newleaf) and analyze their usage of Social Media. We performed quantitative (exploratory data analysis, metadata analysis, network analysis) and qualitative analysis (text analytics, sentiment detection, topic modelling) and combined all these in order to obtain conclusions and make recommendations about the different airlines behavior in their social media channels. We discovered that different airlines have disparate levels of engagement with their customers and sentiment in their Twitter accounts, the most common negative topic mentioned is about delays in their flights and an airline doesn't have to be a major carrier in order to have good control of the message in their social media channels.

| Group Member | Responsibilities |
|---|---|
| Aliasghar Arabi | <ul><li>Collected data and Exploratory Data Analysis for WestJet and Canada Jetlines</li><li>Sentiment Analysis – emotional valence analysis, negative/positive tweets</li><li>Network Analysis – graphs, micro/macro level analysis</li><li>Academic paper and external resource collection</li><li>Report on "Analysis" and "conclusion"</li></ul> |
| Saeede Sadat Asadi Kakhki | <ul><li>Collected data and Exploratory Data Analysis for AirCanada and Porter Airlines</li><li>Association rule Analysis, Influential user analysis</li><li>Academic paper and external resource collection</li><li>Report on "Previous Work / Literature Review", "Analysis" and "References"</li></ul> |
| Bernardo Najlis | <ul><li>Collected data and Exploratory Data Analysis for Sunwing Airlines and NewLeaf Travel</li><li>Topic Modelling for all airlines</li><li>Academic paper and external resource collection</li><li>Report on "Problem Statement and Dataset Selection", "Analysis", "Data Collection and Preprocessing" and "Conclusion"</li></ul> |

**Problem Statement and Dataset Selection**

The focus of the project was to perform social media analytics on the Canadian commercial airlines industry with the intent of finding insights about social media usage in the field, how the selected different companies use social media channels to communicate with their passengers and the type of conversations they have. The intent is to identify patterns or behaviours in the chatter to make potential recommendations that can improve customer engagement.

The airlines selected for the dataset collection consists of:

- **Air Canada** (Flag Carrier and Largest airline in Canada)

- **Westjet** (Second largest Canadian airline, with national reach)

- **Porter** (Regional Canadian Airline, flying exclusively from Toronto, Ontario)

- **Sunwing Airlines** (Low-cost airline, with destinations to US, Mexico and the Caribbean)

- **Jetlines** (Ultra low-cost airline based in Vancouver, BC)

- **NewLeaf** (Canadian Virtual Airline selling tickets for Flair Airlines)

**Previous work / Literature review**

Airlines are a constant topic of social discussion with the largest airlines receiving hundreds of comments every day from customers on topics ranging from the look of a new aircraft to complaints about delays.(1)

A well-defined policy or framework for adopting social media needs to be in place to mitigate risks and avoid any potential legal liability. Thoughtful use of social media and the development of a coherent social media policy are essential. Key tenants of policy making are: Listen, Regulate, Empower, Analyze.(2)

As a result, airlines can make advantage of data analytics since it is about making better decisions.(3) They can apply different data analytics techniques on their social media data, like time distribution, topic popularity, passenger geographical distribution, event tracking, sentiment analysis which compare sentiment of the tweets of different airlines or even for each airline, topic modeling to help them make better decisions.(4,5,6,7)

On the other hand, passengers' positive or negative reviews of airlines will directly affect other passengers' purchase decisions.(8) Passengers can also take advantage of these plat forms. They can make the best choice and get a clear picture of what to expect from each airline such as services and products E.g.: meals & drinks offered, in-flight entertainment, crew service standards etc. and also share their ideas by uploading pictures, videos and status updates about their personnel experiences on Facebook, Twitter etc.(9)

**Data Collection & Preprocessing**

The data collection process focused solely on Twitter: the team gathered observations for the six airlines analyzed using the *'twitteR'* package in R. The team chose to use Twitter search API because of ease of access and convenience to collect data. The collection criteria followed these rules:

- Collect as many tweets as possible (up to 3,200 maximum allowed by the search API)
- All tweets and retweets from and to the official airline twitter account
- All tweets that mentioned the airline's name (capital/lowercase) or use it with a hashtag

Data was collected for all airlines on three dates during the project: March 9, March 11 and March 18. The additional collection done in March 18 was to capture potential observations correlated with a big snow storm in Canada and North America on the week of March 13. The tweets were collected, saved and shared with team members using CSV files. As some of the criteria led to obtaining the same samples multiple times, a simple process for exact match row de-duplication was run. Additional cleanup was done with date/time variables (using the *'lubridate'* R package) adding metadata columns to identify replies from retweets and tweets coming from the official airline account. Here is a summary with the number of observations per airline; the total corpus analyzed is composed of 44,400 tweets.

| Airline | Number of records | % of Corpus |
|---|---|---|
| Aircanada | 25,528 | 57.50% |
| Porter | 2,176 | 4.90% |
| Westjet | 9,762 | 21.99% |
| Jetline | 712 | 1.60% |
| Newleaftravel | 4,785 | 10.78% |
| Sunwing | 1,437 | 3.24% |

**Analysis**

**Exploratory Data Analysis**

We performed exploratory data analysis (EDA) by looking at the number of tweets in general for each airline, percentage of tweets sourced from each airline official account, and percentage of tweets replied and retweets in total data collected for each airline. Some of important findings are that Newleaf Travel is leading in using its Twitter official account having 90% of tweets in our collected data from its official account followed by Westjet with only 15% of tweets and finally only 2% of tweets are from Sunwing official twitter account. The percentage of replied tweets is following the same distribution as how 6 Canadian Airlines have used their official accounts. Analyzing top most used terms, we found that for all airlines, the most used term is their own brand. We also identified the top most mentioned destinations; Toronto appears on three of the six airlines, NewLeaf doesn't have any destination and SunWing has mentions to "Tenerife", "Beach", "Resort" and "Vacation" destinations as expected.

**Time-based Analysis**

We also did time-based analysis over the tweeting behaviour for each airline to see how the distribution is different for each airline over the period of our analysis. To make it even more clear, we looked at the tweeting behaviour distribution over the week days hoping to see any measurable difference across all 6 Canadian Airlines. For NewLeaf Travel tweeting almost always starts in afternoon throughout the week and follow the same distribution for all days. Another general observation across all airlines is that on Mondays, Wednesdays, and Thursdays they are experiencing the most number of tweets.

**Sentiment analysis**

We performed sentiment score analysis of the tweets and for each airline we focused on the very positive (>2) and very negative (<-2) score tweets. We have summarized the result in the appendix (Sentiment Analysis- Score) with the score defined as positive count divided by total count. As you can see Porter

is leading with highest score followed by Newleaf, Jetlines, Westjet, AirCanada, and finally Sunwing with the lowest score.

Then we identified the most frequent terms in the top positive and negative tweets per airline: across all of them there are references to delays as one of the most negative terms. A very clear outlier shows up for Air Canada in a tweet making references to a terrorist threat.

By plotting the emotional valence of tweets across time, we can see how the sentiment varies through the days analyzed, with Air Canada and Westjet having the most variability and the highest number of tweets overall, and NewLeaf being the most consistent and staying always in positive values. The latter can have a correlation with how much of the tweets of NewLeaf are generated from the official account, allowing to have a much better control of the sentiment of the message associated with it.

We used LDA (Latent Dirichlet Allocation) to perform topic modelling and find groups of terms that are clustered (mentioned) together in tweets. If we look at all airlines tweets as a joint dataset we could not identify independent topics, but interesting findings appear for specific airlines (Air Canada, JetLines, SunWing) where specific topics have patterns that are independent of the rest.

**Network Analysis – Macro Level**

We found the macro level measurements for these six airlines to get a better view about their structure and be able to compare them. Density values show that Newleaftravel has the most connected network and AirCanada has the least one. It's reasonable to have this result since the ratio of number of nodes to edges is highest for AirCanada. Furthermore, AirCanada has the widest network, it is interpretable by taking into account it's centralization value. Comparing reciprocity value can somehow give a proper view about responsivity of airlines to their consumers. Zero value for reciprocity show there is no two-way conversation among nodes. As a result, Porter, Newleaftravel, Jetline and Sunwing are not

responsive. Modularity values in average are high for these airlines. It shows these networks consist of different conversations. Please refer to the appendix for macro level measurements.

**Influential user analysis (organization/personals/bots):**

We searched for three most influential users for all of six airlines, according to in-degree and out-degree centrality measures. Official accounts play an important role in both indegree and outdegree and for all the airlines, these official accounts appear in the three most influential accounts, except Westjet, AirCanada and Jetlines in in-degree measure and Sunwing in out-degree measure. We detect if these accounts belong to an individual, organization or even a bot. we observed many aviation industry students tweeting/retweeting. There are cases when some customers who had tried contacting airline by phone about their complaints used twitter as medium to take airline attention and other people retweeted their complain. We also observed that excluding airlines official accounts the top influential users are individuals and not organization. Please refer to the table below for summary of the influential users analysis (excluding their official account).

| Airline | Density | Diameter | Reciprocity | Centralization | Modularity |
|---------|---------|----------|-------------|----------------|------------|
| AirCanada | 0.000141809 | 8 | 0.007968127 | 0.1509167 | 0.8468935 |
| Porter | 0.00474389 | 3 | 0 | 0.09467456 | 0.8195001 |
| Jetlines | 0.01603146 | 2 | 0 | 0.1175746 | 0.7178711 |
| Sunwing | 0.004352031 | 3 | 0 | 0.06551517 | 0.9156307 |
| WestJet | 0.000400569 | 4 | 0.009920635 | 0.08149005 | 0.909061 |
| NewLeaf | 0.0119863 | 3 | 0 | 0.2624421 | 0.5748299 |

**Conclusions**

Airlines can greatly benefit by using social media as their services are always used by individual people; engaging in direct conversations with their customers is a great way to improve their service by listening to what their passengers are saying about them.

Knowing what are the most positive and negative terms associated with an airline brand can help decision makers understand what their brand is doing right and wrong: in the case of airlines the most common complaint is about delays.

Controlling the twitter conversation (like NewLeaf does) allows an airline to manage the conversation and steer its emotional valence. On the other extreme, the SunWing twitter account has only 14 tweets and looks like abandoned, leaving an opportunity for the conversation to be taken over by competitors.

Positive tweets generated organically by the right person can generate a lot of positive influence if a high enough number of their followers echo the message.

**Lessons Learned – Team**

We learned how different types of analysis can be brought together in order to make conclusions and recommendations to a company based on their social media participation.

**Lessons Learned – Aliasghar Arabi**

In general, I learned about how to use combination of search terms to make sure I am collecting right and comprehensive data for my analysis. I learned how to use dictionary of positive and negative words and perform sentiment analysis to find most frequently common positive and negative terms for each airline. I learned how to use different famous packages for network analysis such as igraph, network, and sna to create network and calculate the micro and macro measures. I learned about very cool package for network visualization called vizNetwork by which I created interactive graph for indegree and outdegree where you can zoom in/out and find the influential users easily. Lastly, I realized how much information

business can use by analysis the social media to listen to their clients and find opportunities to improve their services.

**Lessons Learned - Saeede Sadat Asadi Kakhki**

In the first step, extracting data taught me that do the data extraction frequently in a period of time is so helpful for the search criteria with huge amount of tweets, like AirCanada airline, to extract more and more data to increase the accuracy of our conclusions. Furthermore, I understand that look for the influencers can be so informative for organizations specifically for organizations that are completely connected with individuals. They can understand their positive and negative points. Also analyse the social media network in micro and macro level can give good view about the network structure and how nodes are connected to each other and find the similarities and differences among different networks.

**Lessons Learned – Bernardo Najlis**

I learned how social media analytics can help understand the conversations around a company or a group of companies in an industry. Understanding all aspects of a company mentions (quantitative, qualitative, text analytics, sentiment and network analysis) is a must to increase customer engagement and isolate potential damages to the brand. Specifically related to tools, I learned how to use the ggthemes package to create ggplot2 charts in the theme of "The Economist" (and other themes) and also how to use the LDA package to do topic modelling analysis to identify clusters of terms mentioned in tweets.

**References**

1. Social Media Analytics for American Airlines [online]. Available:

   https://www.talkwalker.com/industry-research/airlines/american-airlines-social-media-analytics

2. Social Media for Airline Industry [online]. Available FTP: niit-tech.com Directory: sites/default

   Files: files/Social%20Media%20in%20Airline%20Industry(ISG)_0.pdf

3. Leveraging Customer Analytics: The Airline Industry [online]. Available:

   http://knowledge.wharton.upenn.edu/article/leveraging-customer-analytics-the-airline-industry/

4. Twitter US Airline Sentiment [online]. Available:

   https://www.kaggle.com/crowdflower/twitter-airline-sentiment

5. Twitter Topic Modeling Using R [online]. Available:

   https://gist.github.com/bryangoodrich/7b5ef683ce8db592669e

6. Sien Chen, Yinghua Huang, Wengqiang Huang, "Big Data Analytics on Aviation Social Media:

   The Case of China Southern Airlines on Sina Weibo", 2016 IEEE Second International

   Conference on Big Data Computing Service and Application.

   doi:10.1109/BigDataService.2016.51

7. Pouti Isidore, Moll More Borja, "Airline companies in social media: a case study of HOP",

   Halmstad, 2013

8. Dolan, Rebecca and Shahbaznezhad, Hamidreza, Social Media Strategy in Services: The Effect

   of Airline Brand Posts on Engagement (February 7, 2017). Available at SSRN:

   https://ssrn.com/abstract=2913352

9. Impact Of Social Networking In The Airline Industry Marketing Essay [online]. Available:

   https://www.ukessays.com/essays/marketing/impact-of-social-networking-in-the-airline-

   industry-marketing-essay.php

10. Xiaotong Duan,TianshuJi,WanyiQian, "Twitter US Airline Recommendation Prediction".

    Available: http://cs229.stanford.edu/proj2016spr/report/042.pdf

11. List of Canadian Airlines [online]. Available:

    https://en.wikipedia.org/wiki/List_of_airlines_of_Canada

12. Text Mining with R – Twitter Data Analysis [online]. Available:

    http://www.rdatamining.com/docs/text-mining-with-r

13. Finding and Removing Duplicate Records [online]. Available:

    http://www.cookbookr.com/Manipulating_data/Finding_and_removing_duplicate_records/

14. Links to Indegree and Outdegree Network Visualizations

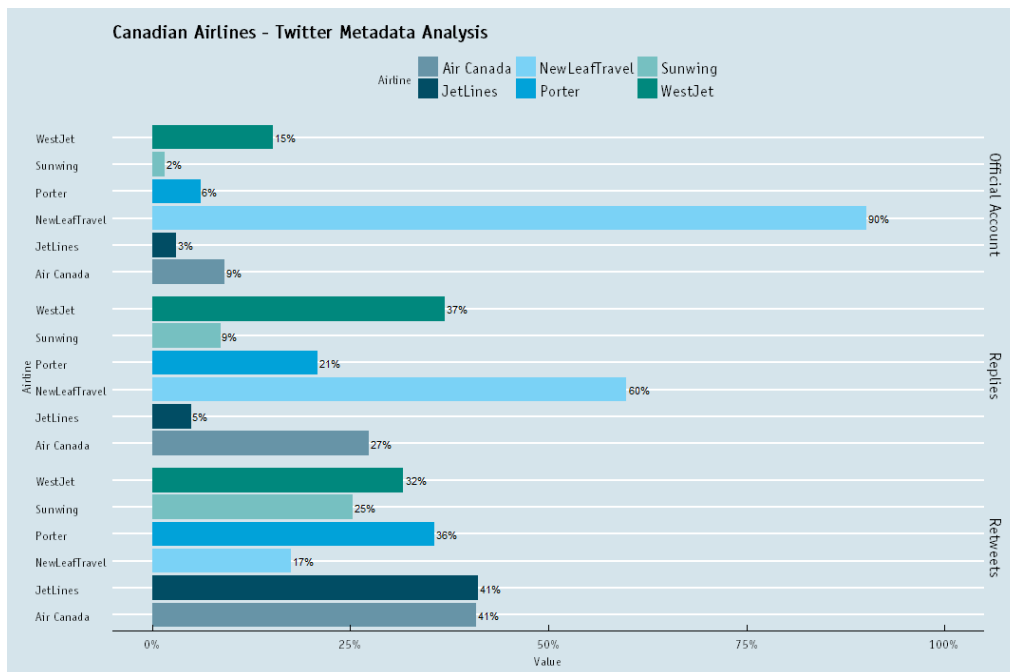| Airline | Indegree Network | Outdegree Network |
| --- | --- | --- |
| Air Canada | http://rpubs.com/ds8006/a | http://rpubs.com/ds8006/ao |
| Westjet | http://rpubs.com/ds8006/w | http://rpubs.com/ds8006/wo |
| Porter | http://rpubs.com/ds8006/p | http://rpubs.com/ds8006/po |
| JetLines | http://rpubs.com/ds8006/j | http://rpubs.com/ds8006/jo |
| Sunwing Airlines | http://rpubs.com/ds8006/s | http://rpubs.com/ds8006/so |
| NewLeaf | http://rpubs.com/ds8006/n | http://rpubs.com/ds8006/no |

**Appendices**

1.  **Influencer Users Analysis**

| Airlines | Influencers | Centrality | Organization | Reason |
|---|---|---|---|---|
| Porter | cookieitup | Indegree | organization | Handmade cookiemaker, porterairline provides these cookies during its travels for its passangres and it attracts passangers. |
| | herbert_dan1996 | Indegree | individual | Student who studeis in Georgian College Aviation Management who is interested in airports, airlines and flights. |
| | mattmcmuscles | Outdegree | Individual | This user post a positive tweet that was retweeted so many times regarding Porter animation. |
| | ouasport | Outdegree | Organization | Ontario University Athletics. Porter is one of the sponsor of the games. |
| Westjet | winnipegwatch | Indegree | bot | A bot which retweets any tweet from Winnipeg or tweets mentioning Winnipeg.There are some flights from or to this city by Westjet |
| | avgeekretweeter | Indegree | bot | It retweets relevant #avgeek tweets. #avgeek is comes from Aviation Geek, which points to fans of aviation. |
| | bbbsc | Indegree | organization | It's a charity that help Canada's youth. Wesjet company helps 10 charities. |
| | ilove_aviation | Outdegree | individual | An individual who posts about avizration.It posted that Westjet almost crashes into the water and so many people retweets this tweet. |
| | airlivenet | Outdegree | Organization | Live airspace and services.It posted that westjet almost crashes into the water and so many people retweets this tweet. |
| | karynpugliese | Outdegree | individual | Lost his bag in a flight and cann't contact with westjet with the numbers they provided so he use twitter to contact them which retweets several times. |
| Air Canada | eddieashton_ed | Indegree | individual | A person that some of his posts have negative meaning about |

| | | | | |
|---|---|---|---|---|
| | | | | aircanada, he might have bad experience with it. |
| | duncanbureau | Indegree | individual | One of the aircanada employees who post positive tweets about Aircanada |
| | vsmith1660 | Indegree | individual | Aircanada Manager Commercial Latinam and Caribbean |
| | collectingly | Outdegree | individual | A numismatist with 30k followers who complaints about aircanada in different aspect |
| | lanaparrilla | Outdegree | individual | Artist. She had a flight with aircanada and an ambassador fo aircanada help her in her fiasco flight and she thanks her. Several retweets |
| | franktmcweety | Outdegree | individual | A person who tweets aircanada CEO prefers Trumps polocies to Tredeaus and so many retweet it. |
| | stephenlautens | Outdegree | individual | This person post new regarding possible aviation terrorism attack which mention the name of airlines and so many people retweets it. |
| Jetlines | mjmorabito | Indegree | individual | Retweet tweets related to jetline. Jetline is close to go public. Mention low cost of Jetlines. |
| | nowyvr | Indegree | organization | Publish news related to Vancover. |
| | dhirajnarang1 | Indegree | individual | Retweet tweets related to jetline. Mention low cost of Jetlines |
| | marketonemedia | Outdegree | orhanozation | Media and marketing company who post a tweet regarding the low cost of jetlines among the canadian airlines |
| | bivnews | Outdegree | organization | West Coast business news and intelligence. Post news regarding low cost Jetline airlines is close to go public and start trading in TSX |
| NewLeaf | winnipegwatch | Indegree | bot | A bot which retweets any tweet from Winnipeg or tweets mentioning Winnipeg.There are some flights from or to this city by Westjet |
| | yhmdirops | Indegree | organization | Hamilton international airport that posts news regarding flights and travel. |

| | | | | |
|---|---|---|---|---|
| | lisaleclair43 | Outdegree | individual | Losts her luggage in Newleaftravel and post tweets regarding this issue |
| | the_mark_hill | Outdegree | individual | Good experience with newleaftravel flight and share it with others |
| Sunwing | consumerpi | Indegree | organization | Consumer protection. Post complaints of consumers |
| | avgeekretweeter | Indegree | bot | It retweets relevant #avgeek tweets. #avgeek is comes from Aviation Geek, which points to fans of aviation. |
| | presidenciard | Outdegree | organization | Official twitter account of the Presidency of the Dominican Republic. Post tweets related to the investment of Sunwing |
| | cbcnb | Outdegree | organization | CBC New Brunswick. Post about the event related to Sunwing that the flights was out of power and people claim about it. |

2. **Exploratory and Metadata Analysis – Tweets from Official Account, Retweets and Replies per Airline**



Canadian Airlines – Twitter Metadata Analysis

Official Account:
- WestJet: 15%
- Sunwing: 2%
- Porter: 6%
- NewLeafTravel: 90%
- JetLines: 3%
- Air Canada: 9%

Replies:
- WestJet: 37%
- Sunwing: 9%
- Porter: 21%
- NewLeafTravel: 60%
- JetLines: 5%
- Air Canada: 27%

Retweets:
- WestJet: 32%
- Sunwing: 25%
- Porter: 36%
- NewLeafTravel: 17%
- JetLines: 41%
- Air Canada: 41%

### 3. Sentiment Analysis- Score

| group | pos.count | neg.count | all.count | score |
|---|---|---|---|---|
| AirCanada | 1258 | 1122 | 2380 | 53 |
| WestJet | 866 | 465 | 1331 | 65 |
| Newleaf | 851 | 207 | 1058 | 80 |
| SunWing | 35 | 69 | 104 | 34 |
| Porter | 160 | 11 | 171 | 94 |
| Jetlines | 3 | 1 | 4 | 75 |

### 4. Sentiment Analysis – Emotional Valence per Airline