Does Twitter Fly?

American Airlines Twitter Team Performance Analysis

Objectives

During the last few years, American Airlines has allocated a great amount of resources to a Twitter Team that focusses on dealing with customer complaints and maintaining a strong bond with their customers. The main aim of the team is to make sure that customers have a positive association towards American Airlines.

At this point in time, American Airlines has come to evaluate the performance of the Twitter Team. Consultants from ASK YARD have been asked to do a performance analysis of the American Airlines Twitter Team.

Methodology

Data extraction

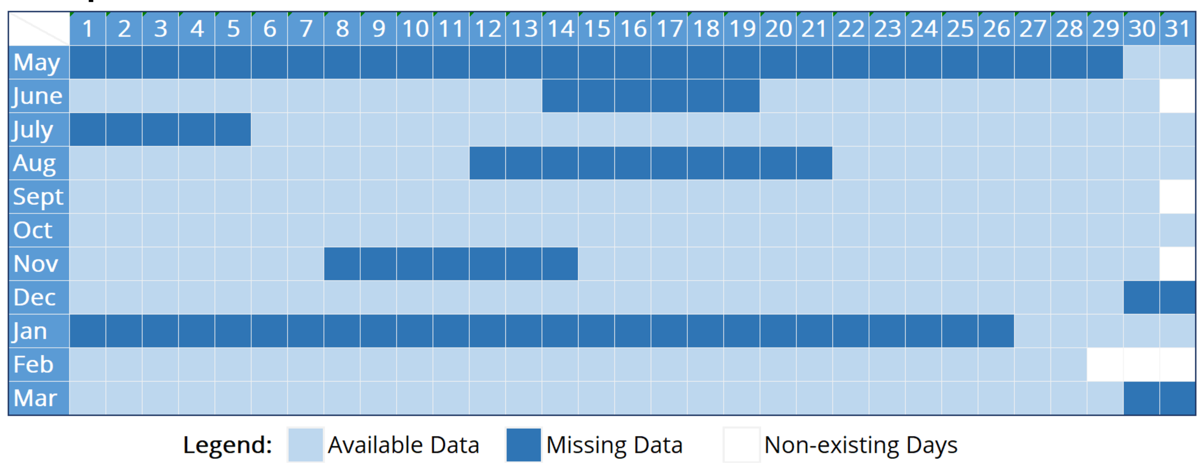
The raw data consists of a main zip file and 350 sub zip files containing tweets in JSON format. Using a Python script, a subset of characteristics of all tweets is extracted, without unzipping the individual files, thus saving valuable time and space. The approximately 5.5 million tweets are stored in a SQLite database. Through SQLite queries in Python, the necessary data for further analysis is extracted.

Complexity = O(n)

Database

- Tweet ID
- User ID
- Timestamp
- Tweet that is replied to
- User that is replied to
- Who is mentioned in tweet
- Language
- Text

Gaps in data



The majority of data in June only contains the received tweets by any airline and not the tweets the airlines sent themselves.

Conversation

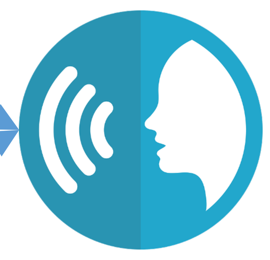
A conversation starts with a tweet that mentions @AmericanAir, to which **American Airlines** responds. The conversation length is defined as the length of the chain of replies to the initial tweet, **including** the initial tweet and only following the original user.

Sentiment Analysis

English

96% Tweets of American Airlines

74% Tweets of complete dataset

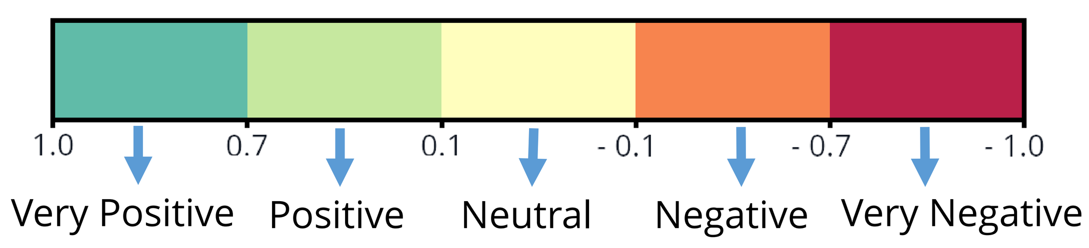


Sentiment analysis is only applied to conversations that start with an English Tweet. The sentiment of each conversation is analyzed using V.A.D.E.R., which is a valence-based approach. Advantages of V.A.D.E.R. are[1]:

V.A.D.E.R.

- Recognizes and includes slang into lexicon
- Considers syntax and context
- Recognizes capitalization
- Handles changes in a sentence sentiment intensity when it contains word: 'but'
- Considers modifying words that are present in front of a sentiment term

The V.A.D.E.R. analysis takes O(nm), in which n is the number of tweets and m is the length of the tweet. The accuracy of the sentiment analysis is 85% (based on a test set from the University of Michigan)[2]. According to Gilbert and Hutto (2014) V.A.D.E.R. outperforms most classifiers and algorithms and is therefore a very good method for analyzing tweet sentiment[3].



Sentiment change =
[First customer Tweet] – [Average of remaining customer tweets]

Results - Exploratory Analysis

Amount of Sent and Received Tweets @ AmericanAir

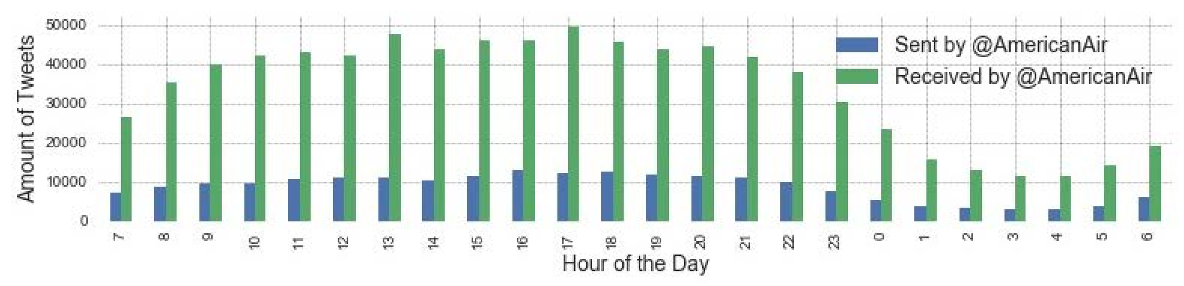


Figure 1: Distribution of American Airlines tweets per hour of the day (ET) (Including non-English tweets)

Conversations & Response Rates per Airline

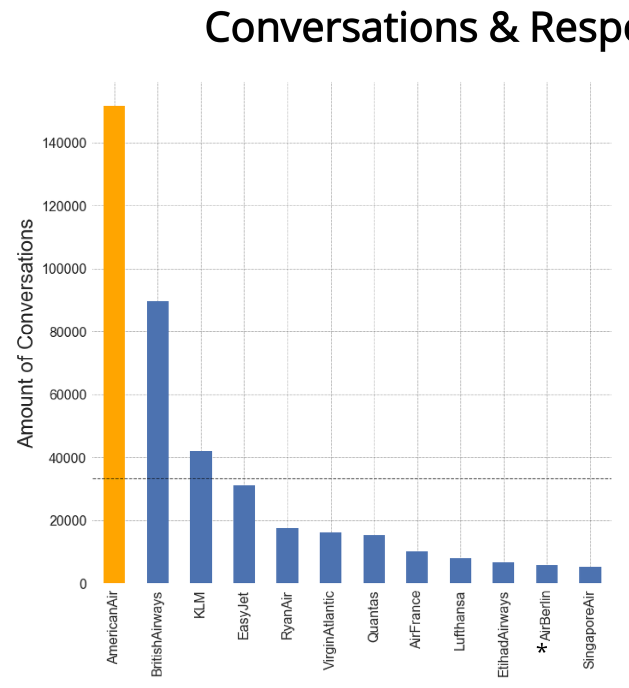


Figure 2: Comparison of conversations (including conversations with non-English tweets)

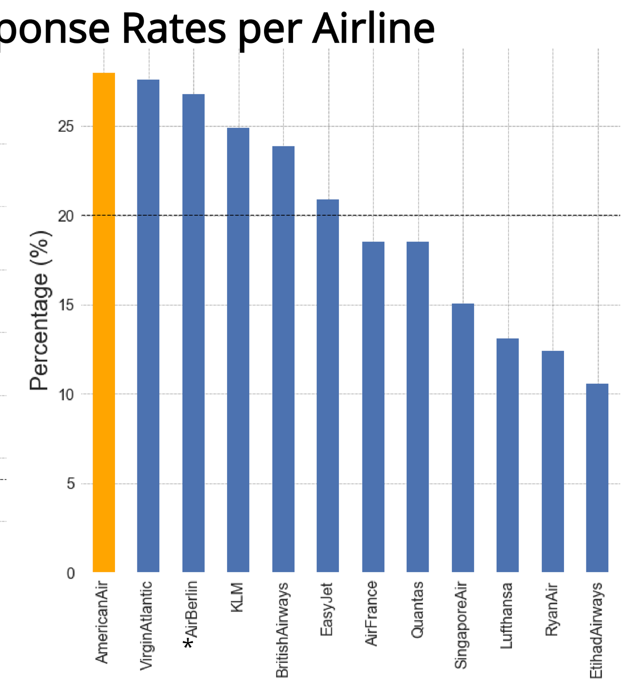


Figure 3: Comparison of response rates (including conversations with non-English tweets)

Average Reply Time & Conversation Length per Airline

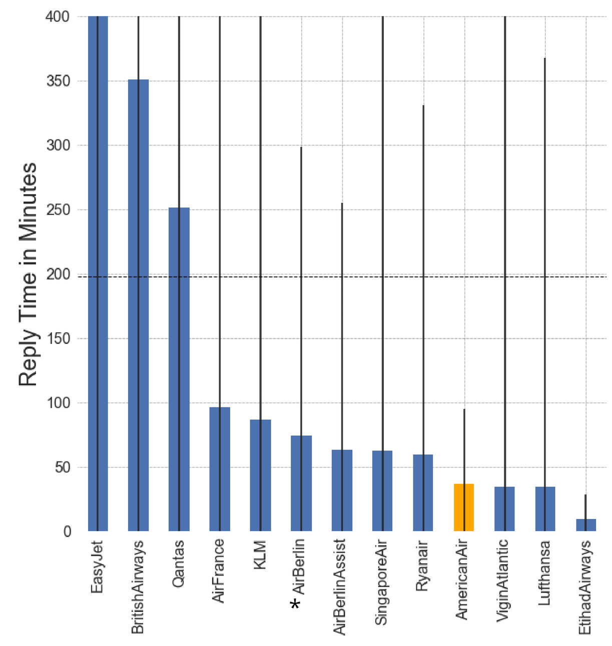
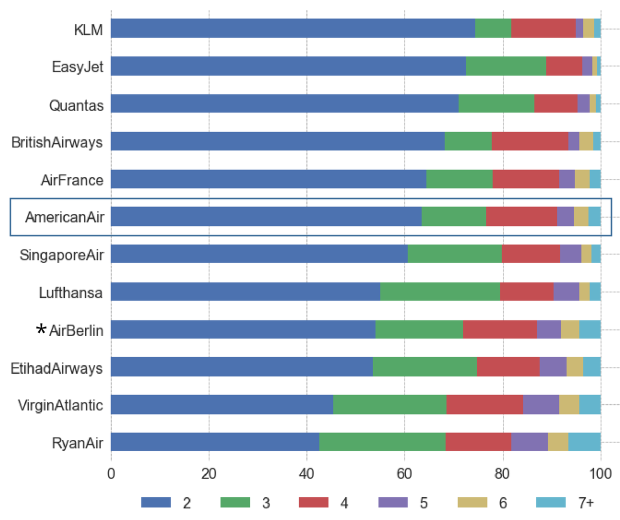


Figure 4: Comparison of reply time (incl. standard deviation)



* In all graphs AirBerlin includes AirBerlinAssist

Figure 5: Comparison of conversation length

Results - Sentiment Analysis

Initial Tweet Sentiment

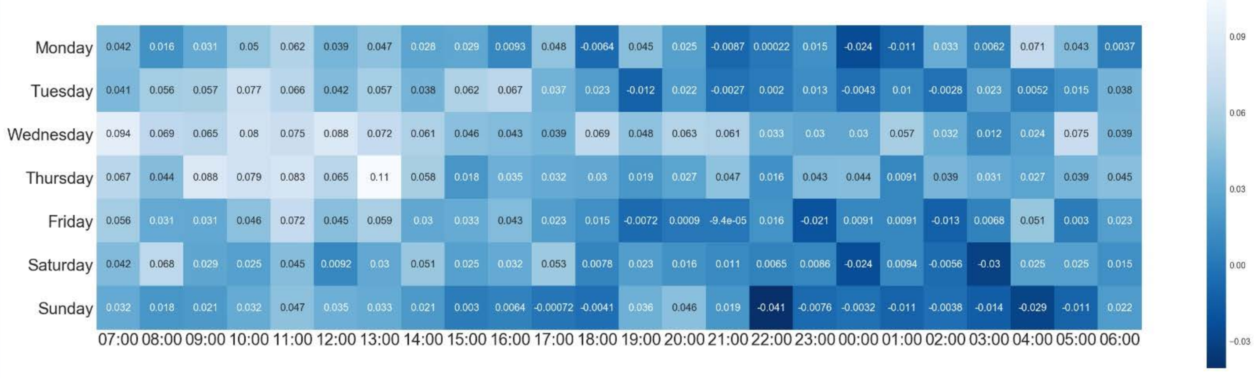


Figure 6: Initial tweet sentiment per hour of the day (ET) and day of the week

Overall Change in Sentiment

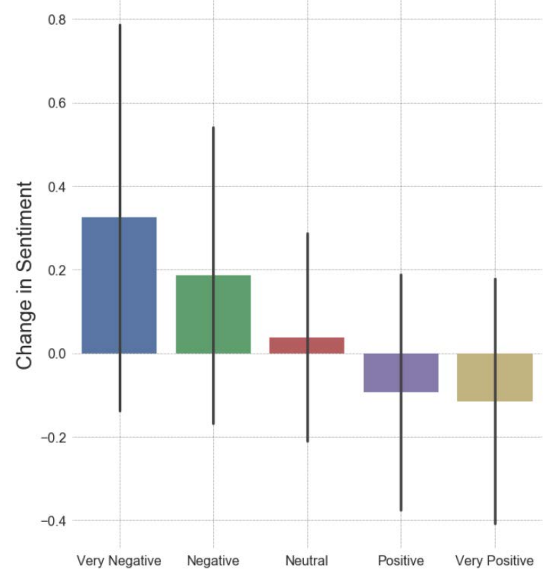


Figure 7: Change of sentiment per category (incl. standard deviation)

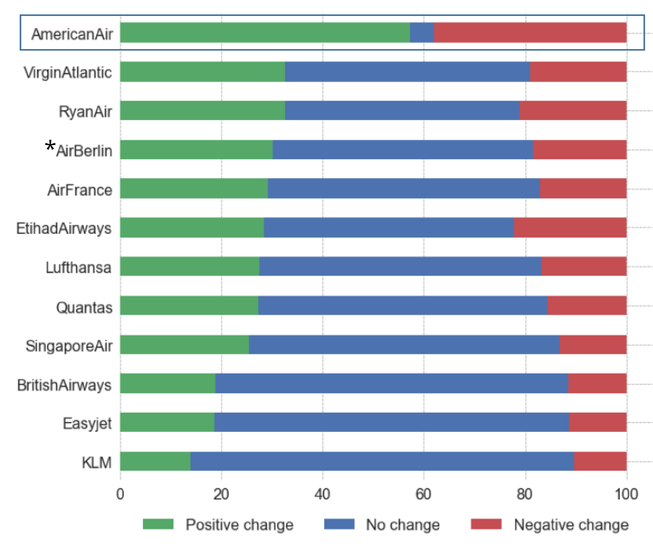


Figure 8: Comparison of relative change in sentiment over all conversations (excluding 2-tweet conversations)

Change in Sentiment per Conversation Length

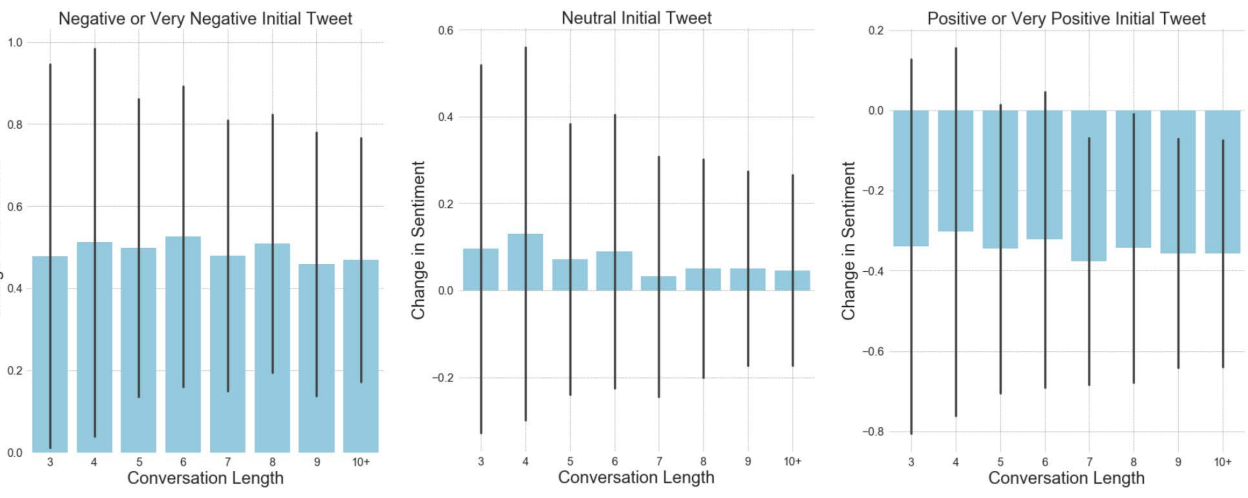


Figure 9: Change of sentiment based on the sentiment category of initial tweet (incl. standard deviation)

Sentiment per Airline

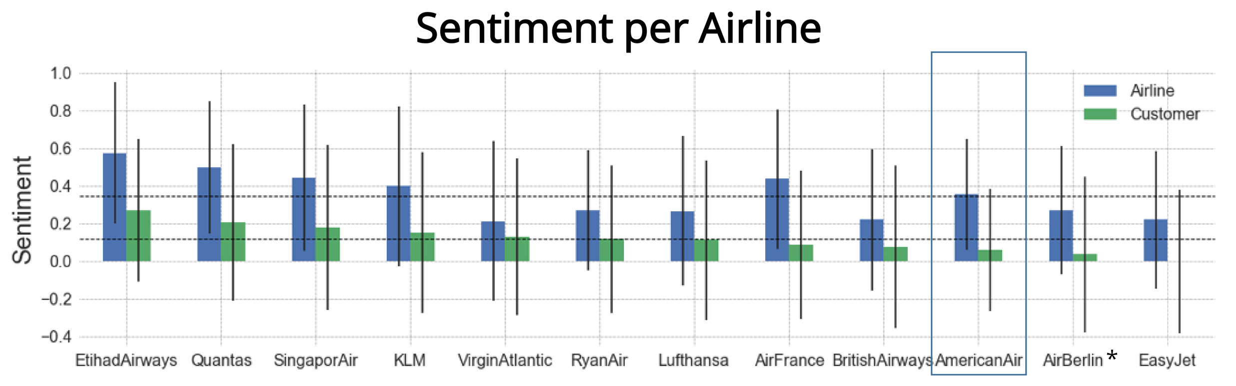


Figure 10: Comparison of average customer and average airline sentiment (incl. standard deviation and averages for Airline and Customer sentiment)

Discussion & Conclusions

Figure 1 shows that there is a significant drop in the amount of tweets and responses during the night. This coincides with a decrease in flights during that period. In addition, figure 6 shows that during this period customers start the conversation less happy than during the day. Moreover, this effect is stronger during the weekend than during the week. Furthermore, figure 1 shows that the American Twitter Team is active during the night, so the less satisfied customers are addressed.

Figure 2 and 3 indicate that American Airlines has by far the most conversations compared to competitors and also the highest response rate. Although the gaps in the data may make the response rates seem lower than they actually are, it is clear that the response rate can be improved.

Figure 4 shows that American Airlines has one of the fastest and most steady reply times, as they have one of the lowest means and lowest standard variations. In addition, figure 5 shows that all airlines have a majority of 2-tweet conversations, followed by 3- and 4-tweet conversations. This means that the majority of conversations (2-tweet) will give no information on the change of sentiment during the sentiment analysis, as there is only one customer tweet.

The overall change in sentiment in figure 7 indicates that American Airlines is changing the sentiment of customers who start out as positive or very positive in a negative way, while increasing customer sentiment when customers start out as negative, very negative or neutral. However, all of these average changes are much smaller than their standard deviations. This means that the change is not very stable.

Figure 8 compares the first sentiment to the average sentiment of the remaining tweets and determines if this is higher (positive change), lower (negative change) or neutral (no change). American Airlines has the most positive changes, but also the most negative changes. In general, the more positive change, the more negative changes as well.

Figure 9 shows that the change in sentiment does not change when the conversation length increases, regardless of whether the change is negative or positive or what the category of the initial tweet is. The standard deviation is high, but consistently so.

Finally, figure 10 shows that the average customer of American Airlines is more negative compared to competitors, but that the Twitter Team is overall more positive than the average competitor. This and the results of figure 8 shows that the American Airlines Twitter Team is able to make a difference in customer sentiment.

Recommendations

The American Airlines Twitter Team is performing well compared to competitors, as American Airlines processes more conversations than any other airline, their response rate is the highest, and their reply time is great as well. However, the following is recommended:

- Since American Airlines already has a positive impact on customers with negative sentiments. It is important to focus on the positive tweets. American Airlines should critically revise the current script for answering tweets and build upon the success stories. Develop user profiles to determine which script to use for each customer and develop training for the Twitter Team in order to make more positive changes in sentiment happen.
- Customer tweets are more negative during the night than during the day. Therefore, the night shift of the Twitter Team should consist of especially well-trained staff that know how to deal with upset customers well.
- American Airlines has a great response time, but still only a small amount of tweets is replied to (about 30%). After improvements have been made to increase the sentiment change, the activities of the Twitter Team should be expanded, focusing on answering more tweets and not so much on reply time.

Evaluation & Improvements

Although this project has been handled with the utmost care, there is always room for improvement.

- V.A.D.E.R. has limitations when customers are negative in a polite way. Future research could focus on a classifier specifically for airline tweet sentiment analysis.
- The way a conversation has been defined limits the possibility of researching aspects such as multiple participants in a conversation.
- The gaps in the data probably have caused conversations to be incomplete, affecting the real change in sentiment.
- The initial tweet is compared to the average of the remaining tweets in order to compensate for any misclassification by V.A.D.E.R. in the final tweet. As a result, some changes might be wrongfully classified as negative and vice versa. Only those conversations that started positive or neutral, went very negative and ended positive are wrongly classified in this situation.
- Possible outliers such as Christmas or the Black Lives Matter event have not been removed, which may have clouded the overall results of the analyses.

[1] http://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html
[2] https://www.kaggle.com/c/si650winter11
[3] http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf

Technische Universiteit Eindhoven University of Technology
Where innovation starts

J.C.M. Arts | A. Castella | Y.T. Hu | S. Papandroudis | J. Phelipa | K.M.A.A. Schrouff | D.A. Warzyńska

Tilburg University
Understanding Society