Research in Industrial Projects for Students

Sponsor

Twitter

Final Report

#Conversations

Student Members

Mary Kemp(Project Manager), *Occidental College*,

marymariekemp@gmail.com

Alexandra Howes, *University of Cambridge*

Liyu Xia, *University of Chicago*

Afzal Hossain, *New York City College of Technology*

Academic Mentor

Brian Kim, kimbrianj@ucla.edu

Sponsoring Mentors

Roja Bandari, rbandari@twitter.com

Date: August 30, 2015

This project was jointly supported by Twitter and NSF Grant DMS-0931852.

Abstract

Twitter is a real-time social networking platform, often used by individuals to express issues or feedback to an organization. The project aims to study the effectiveness of these conversations, and this mid-term report presents a summary of our work so far. We were provided with a dataset of 40 million Tweets on which we aim to perform statistical analysis in order to explore the nature and structure of conversations between organizations and users on Twitter, and how successfully such interactions are resolved. As a preliminary analysis we began by studying 500,000 Tweets from the airline industry, which gave an overview of conversation structures and lengths, and how the use of Twitter varies between organizations and the generators (software packages) they use. Based on these results, we are currently working to extend the analysis to the larger dataset, comparing conversation features between industries. Research into determining how successfully customers' issues are resolved on Twitter using natural language processes such as feature selection and sentiment analysis is also ongoing, and is showing positive results for the future. The results of the project should help Twitter to develop features and capabilities to improve the way customer service issues are dealt with on Twitter.

Acknowledgments

We would first like to thank IPAM for hosting the RIPS program. Next we thank Twitter and the NSF Grant DMS-0931852 for sponsoring our research. Special thanks go to Brian Kim our Academic Mentor, Roja Bandari our industry sponsor liaison, and Fiona Pigott our honorary sponsor mentor for their support and mentor-ship.

Additionally, we would like to thank Prof. Kilho Shin for inventing and sharing with us the Super CWC/LCC methods for fast feature selection. We would also like to thank Prof. Takako Hashimoto for providing guidance on how to use Super CWC/LCC and giving suggestions on how to manage Twitter data in general. Thanks also go to Allison Hegel for providing suggestions for using Pattern for sentiment analysis and polarity testing.

List of Figures

Chapter 1

Introduction

1.1 Sponsoring Organization: Twitter

Twitter is an online social networking service in which users communicate through 140-character *Tweets*. Since it operates in real-time, it is frequently used by individuals to ask questions, provide feedback, or to get help from organizations. Customers may contact organizations on Twitter to complain about a canceled flight or to find out about delays during transit; the use of Twitter in this way as a customer service tool is becoming increasingly popular (Section 1.4).

1.2 Twitter Terminology

A person may choose to reply to a Tweet on Twitter, creating a conversation. However since it is possible to reply to any Tweet within a conversation, conversations can become complex and non-linear, forming a tree-like structure. Throughout this report, *conversation* will refer to any self-contained Twitter interaction, and thus a single Tweet on Twitter (a Tweet which has not been replied to, and is not a reply to another Tweet) is still termed a *conversation*. Due to the way in which our data was pulled, all single Tweets necessarily contain a *mention* (see below) of another Twitter user. The size of a conversation will refer to the total number of Tweets in the conversation, and we will specify when we are only looking at conversations consisting of more than one Tweet.

Usernames on Twitter are most commonly referred to as *handles*, and are always preceded immediately by the @ symbol. The handle is how a person/ organization is identified.

On Twitter, the @ symbol is used to call out usernames in Tweets by writing *@username* in the Tweet body; this is called *mentioning*. People use @username to send Tweets, messages or links to another user's profile. All replies necessarily contain a mention of the person to whom they're replying, and thus all Tweets within a conversation (with the possible exception of the first Tweet) will contain at least one mention.

In addition to their main Twitter account, organizations commonly have a Twitter account dedicated specifically to handling customer care issues—this is called the *care handle*. An organization may have more than one care handle, and therefore there may be several Twitter accounts associated with a single company.

As another way of enhancing interactions and engaging customers on Twitter, organizations frequently employ the use of *generators*. Generators are software platforms designed to improve the way organizations interact with customers on Twitter, with particular emphasis on resolving customer service issues and maintaining positive customer relationships. Organizations may use more than one generator to interact on Twitter; examples of this are seen in the Airlines study (Chapter 3).

Finally we will adopt the convention throughout the report of referring to individuals as *users*, in order to distinguish them from organizations.

1.3 Project Outline

Twitter is becoming increasingly popular (Section 1.4) as a way for users to contact organizations, and thus a way for organizations to conduct customer service. An example of such an interaction could be a user posting a Tweet on Twitter, complaining about delays to a flight and mentioning a particular airline. The airline would then see that it had been mentioned and may choose to respond to the Tweet, possibly using a care handle. The response could be an apology, or guidance as to further action the user should take. Our project aims to study the form of these conversations, their structure, and their properties. We will also analyze cases where users are not responded to by organizations.

Furthermore we will consider what constitutes a successful interaction between a user and an organization on Twitter, and how conversations are most commonly resolved. Particular emphasis will be placed on understanding conversations in which an issue or problem was raised by a user, and how this was dealt with by an organization using Twitter. Comparisons between organizations, generators, and industries will be drawn. We will also explore consider how factors of a conversation influence the final resolution.

The project will enable Twitter to better understand how users and organizations use Twitter, and the nature and structure of conversations that result from these interactions. The way in which these conversations are resolved will be important to Twitter in understanding its role as a customer service tool. It will also be useful to outline how Twitter usage varies between generators, and to what extent the particular generator used affects the nature and success of user-organization interactions. The most and least successful generators and organizations will be identified, and comparisons of Twitter use between industries will also be made. Ultimately these results could help Twitter to improve the service for both users and organizations by the development of features and capabilities that encourage companies to continue using Twitter to perform customer service.

1.4 Related Work

Previous work on this subject was led by McKinsey & Company ([**?**]) whom Twitter hired to conduct a customer service volumes study in order to determine the best practices for companies to perform customer service through Twitter. The research looked at 1% of Tweets over a two year time span and found that users were increasingly using Twitter to contact companies, and that companies' responses to these Tweets were falling behind. It goes on to note the rise in the use of company care handles, as well as the pros and cons of having them. Finally it discusses different levels of company engagement on Twitter, from answering only Tweets in which they are mentioned, to searching for keywords and creating opportunities for customer engagement.

1.5 Approach

Due to the large size, the data was uploaded to a database (MonogDB) since indexing would allow the data to be queried and accessed quickly. However, initial exploration (Chapter 2) showed that brand care handles were missing from the largest dataset, and hence the majority of the project focuses on analyzing a 5 million Tweet subset of the data (from an initial set of 40 million). Since the data was in the form of lists of raw Tweets, it was then necessary to reconstruct conversations which was done by looking at the *in reply to* field in each Tweet. Information about these conversations were also calculated and stored at this stage, such as time length of the conversation, and number of brands involved. Due to many conversations starting without a mention, many conversations were found to have the first Tweet missing from the dataset; over half of these were recovered using the Twitter API.

Initial descriptive statistics were performed on a 500,000-Tweet dataset involving airlines (Chapter 3). This gave preliminary descriptions about the structure and nature of conversations, as well as exploring how conversations vary between organizations and the generators they use. Leading on from this, the remaining 5 million Tweets were explored (Chapter 4), with particular emphasis on drawing comparisons between industries. In particular, the activity of Twitter industries, and the proportion of Tweets from customers to which they reply, were considered.

Next we aimed to concentrate our analysis on customer service interactions between users and companies from the Travel and Telecomm industries on Twitter (Chapter 5). This first required customer service related conversations to be selected by extracting keywords from the first customer Tweets in the conversations, and manually reading through conversations. Second, these customer service related conversations were categorized based on how they end, into *resolved*, *not resolved*, *moved off platform*, *continued on platform* and *other*. This was primarily done by extracting keywords from the last customer and brand Tweets in the conversations, but other refinement rules were also implemented specific to different industries. Once conversations had been categorized, data visualization explored how conversation properties vary by the way conversations are closed on Twitter, and also looked into the industries, organizations and generators which were the most successful at resolving customer's issues.

We also performed initial logistic regression analysis to study the factors that influence the conversation categories. More details can be found in Chapter 6.

1.6 Report Structure

The report will begin by discussing initial observations about the data, how conversations were constructed from the raw Tweets and the process of obtaining Tweets which were initially missing from our dataset (Chapter 2).

The next section (Chapter 3) will present the findings of a statistical analysis performed on 500,000 Tweets from the Airline industry, which includes a discussion of the structure of conversations, and the difference in Twitter usage between organizations and generators.

Next a study on a larger data set of 4.9 million Tweets (Chapter 4) will contrast conversations across industries, drawing comparisons to the Airlines results.

Identification and categorization of customer service Tweets based on how they are concluded, followed by visualization of conversation categories, will be presented in Chapter 5.

Finally, future directions will be considered (Chapter 6), with an emphasis on Natural Language Processing and regression analysis.

At the end of the report can be found the Appendixes—in particular **??** contains code and file documentation from the project—as well as the Selected Bibliography Including Cited Works.

Chapter 2

Data Preparation and Building Conversations

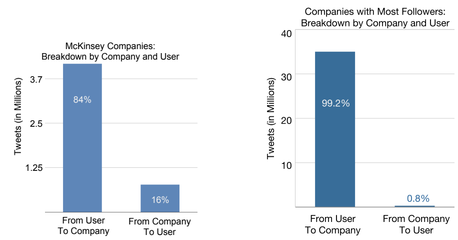
Twitter provided a data set of 40 million Tweets generated using Gnips Historical PowerTrack. Using keyword matching, this process pulled all the Tweets available between May 1st 2015 at 00:00:00 UTC and June 1st 2015 at 00:00:00 UTC. Given a list of company handles, Gnip located and pulled any Tweet from a user that mentioned one of the company handles, or any Tweet posted by one of the company handles mentioning a user. Since it was believed that the number of conversations beginning without a mention was negligible, those tweets were not retrieved.

The data was separated into two sets corresponding to two sets of companies. The first set of companies were pulled for the McKinsey study (REFERENCE) while the second set of companies were pulled for being verified users that had the most followers within each industry. These separated lists and the information provided about each presented particular difficulties in analyzing the set as a whole and will be discussed in the following sections. Ultimately we decided to focus on the McKinsey set because it was not only more manageable, but contained more customer service related conversations.

2.1 Data Description

Performing counts on the data provides an interesting look into the nature of interactions between users and brands on Twitter and allowed us to determine that combining the two company sets' tweets would prove to be a hindrance to our ultimate goal of studying customer service interactions. The entire dataset consisted of 40,214,447 Tweets. These were split in two groups one of 4.9 million tweets that were to and from the McKinsey companies and the other of 35.3 million tweets that were to and from company handles with the most followers. The percentage of to versus from tweets in both of these sets are starking and provides significant insight on the way these lists of companies were pulled. As can be seen in Fig. 2.1 the companies with the most followers respond at much lower rates than those chosen by the McKinsey Company. It is also worth noting that the total number of tweets from the McKinsey companies was almost three times the size of the number from the companies with the most followers.

Upon further investigation, we found that roughly 58% (23 million) of all tweets to and from the companies with the most followers were retweets where a only 37% of tweets to and from the McKinsey companies were retweets.



(a) (b)

Figure 2.1: Proportion of Tweets to and From Brands

While the information provided to us about the McKinsey companies included care handles, we had no information about the care handles in the companies with the most followers and only few were included. It became clear that the lack of care handles in the set of companies with the most followers would ultimately damage our ability to properly study customer service mediated through Twitter. For example, we could determine that Netflix (which was in our dataset) had a very low response rate without ever considering its care handle NetflixCares (which was not in our dataset) which may have a very good response rate. The uncertainty of this dataset's care handles and its disproportionate amount of tweets to and from led to the decision to remove all 35 million in order to focus our research on customer service related conversations.

The rest of this paper will focus on the McKinsey companies consisting of 4.9 million Tweets in addition to 164,000 more Tweets that were obtained to complete conversations (see Section 2.3.2.

2.2 Management and Processing

Given that our data set was originally 40 million Tweets, we decided to use MongoDB (an open-source database developed by MongoDB, Inc) in order to store and manipulate the data more efficiently. Using indexing, accessing and querying data in MongoDB can be done quickly, allowing information to be extracted for analysis.

We found that filtering out tweets that were not in English was also necessary because we intended to look at Tweet bodies to determine resolution later on.

2.3 Building Conversations

Since our goal is to study the interactions being had on Twitter, it was necessary for us to reconstruct these interactions (or conversations) from the raw Tweets.

On Twitter, users can reply to a Tweet by clicking a reply button and using a user mention. In our data, a Tweet that replied to another Tweet will contain an inReplyTo field, containing the unique ID of the Tweet being replied to. As a user can choose to reply to any Tweet in a conversation, conversations are often complex and non-linear, and form a tree-like structure as seen in Fig. 2.2.

In order to obtain all of the information provided by a conversation, we use a tree building technique to group Tweets and organize them by their depth in the conversation (referring to the number of Tweets a single Tweet in a conversation is away from the root or first Tweet). After these Tweets are grouped and ordered, we obtain company level

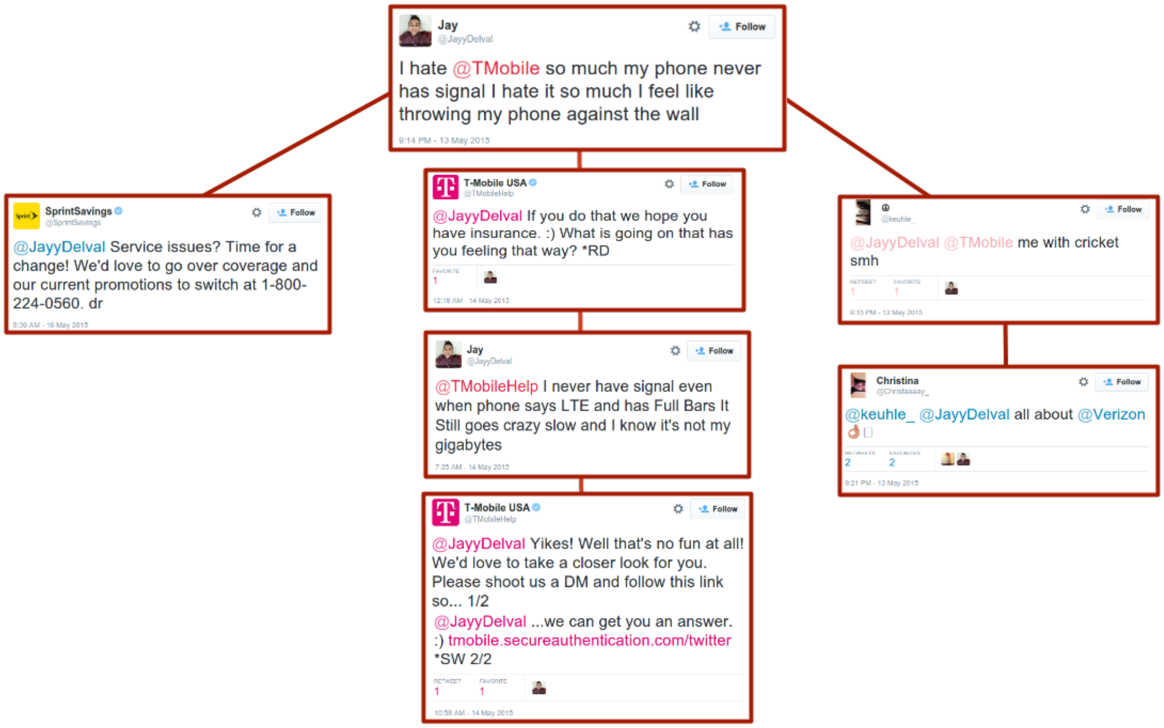


Figure 2.2: Conversation Tree

information such as brands involved, first/last user/brand Tweet ids, hours to first response, and size and depth of conversation. For more information on how these conversations were built see **??**(REFRENCE CONVO BUILDER IN APPENDIX)

2.3.1 Conversation Properties

There are 5 conversational attributes that should be noted from Fig. 2.2:

1. *The tree structure of the conversation*: The first Tweet (from JayyDelval) at the top of the tree is at what we will call *depth 0*, and is responded to by three different users (at depth 1) beginning three new branches of the conversation.

2. *Roles of brand and care handle*: In this case the user Tweets at the brand handle, TMobile, however, the response comes from the care handle, TMobileHelp, while the TMobile handle never responds at any point in the conversation. This shows a clear delineation of roles of these two handles and represents the relationship between the two.

3. *Response from brand without prompt*: We see that the first Tweet in the conversation never mentions the wireless brand Sprint. However, the handle SprintSavings identifies an opportunity to capitalize on the TMobile customer's dissatisfaction with TMobile and adds themselves to the conversation. This is a good example of how some brands proactively search and reply to users.

4. *Multiple brands involved*: Here we can see that not only does SprintSavings respond but yet another brand is mentioned, Verizon. This shows that different brands may converse within the same conversation tree.

5. *Lack of response from brand*: Although Verizon is mentioned, it chooses not to respond to the thread. This represents one of many cases where a mention from a user does not get a reply.

6. *Common word usage in final brand/care Tweet*: Finally, we see in the final TMobileHelp Tweets that very common word usage occurs, namely *DM* and *link*. These are examples of words we plan to look for when searching for resolution and success of interactions.

These observations shed light on the complexity of each conversation and allow us to better understand how interactions between users and brands can be characterized. The frequency of such incidences are provided in Section 2.3.3. It is worth noting items 2 and 4 when building conversations; care handle tweets as well as tweets throughtout the industry should be included in your dataset if the goal is obtaining complete conversations.

2.3.2 Missing Tweets

Building conversations on the 4.9 million Tweets created 598,856 conversations of size greater than one, 46% of which were incomplete (the first Tweet was not in our dataset. We found due to the lack of conversations, it would be imperative for us to figure out why these tweets were not included and try to obtain them.

There are several reasons why Tweets could be missing from conversations: the Tweets were posted before May 1st, the root Tweets of the conversation may not have mentioned a user or company, the Tweets could not have been in English, or the Tweets may have been deleted or made private by the user. Since there is little that could be done about the second two, we concentrated on investigating the first two reasons.

Initially we conjectured that the only broken conversations were due to Tweets that were completing conversations which began before May 1st (i.e. before the data pulling began), however we were somewhat skeptical about this due to the high number of broken conversations. We therefore investigated the cases where companies responded to Tweets about themselves in which they were not mentioned, and where users responded to brands who did not mention them. An example of this is shown in Fig. 2.3 and is discussed below.

Figure 2.3: Missing Tweet Example

We see in Fig. 2.3 that a user is tweeting about State Farm and Geico without mentioning either company with the @ symbol. Therefore this Tweet would not be in our original dataset. This means that after the Tweets are grouped into conversations we would only get the Tweets in the blue box, whilst the first Tweet in the conversation tree, from user Matt Bartlett, would be represented as an empty Tweet. This would affect our ability to calculate time lengths of conversations as well as response times from organizations to users. It would also impair our ability to study the text in the Tweet bodies of each Tweet, since we would not know how the conversation was initiated.

In an effort to obtain the most complete conversations, we collected the Tweet IDs of all Tweets missing from the dataset and used the Twitter API (application programming interface) to pull them back into the conversations. Out of the 273,752 missing Tweets, 187,780 (69%) were obtained from the API and after filtering out non-English Tweets this number fell to 164,729 (60%). 9% of these recovered Tweets were from companies, 7% (about twice the amount we expected) were Tweets that were posted before May 1st (i.e. had been outside the data pulling time period), and 22.7% were replies to further missing Tweets. We found that 13% of all conversations of size greater than one were started by users and were responded to by a company without a mention. Note that this number does not include cases as in Fig. 2.2, where the first tweet mentions a brand but a brand that is not mentioned responds, so this number maybe much higher.

With the missing Tweets, the final number of Tweets we consider is 5.1 million.

2.3.3 Conversations Description

Our 5.1 million Tweets produced 3.7 million conversations, 84% of which contain only one Tweet. Due to the percentage of Tweets from users versus from organizations, this percentage is not surprising. Out of the 16% of conversations that have more than one Tweet, about 74% contain a Tweet written by a brand. If we also account for the number of conversations that are incomplete, we have about 350,000 complete conversations that contain an interaction between a user and a brand.

Furthermore, we investigated the frequency of different conversation properites. We found that out of all conversations of size greater than one (587,722 conversations):

- 21% have greater size than depth, ie. has a tree structure.
- 13% involve both care and brand handle.
- Over 13% will have a response from a brand without a mention at that brand.
- 2.8% will contain more than one brand.

We also found that:

- 25% of conversations that contain a response from a brand with a care handle, involve both brand and care handle, which implies that 25% of the time, a company's care handle picks up a tweet mentioning its brand handle.
- Approximately 74-99% of all user tweets mentioning a brand will not have a response from a brand.
- The key term 'dm' or 'direct message' appear in 12% of all conversations of size greater than 1.

2.4 Key Observations

Many observations were made when processing and managing the data that should be kept in mind for future research in this area.

1. *Care handles matter*: One of the main lessons we learned is that if we are to study the succes of a company's ability and methods of conducting customer service on Twitter, we must not only know if the company uses a care handle, but we must treat that handle as part of the organization and obtain the tweets from that handle.

2. *Retweets may not be useful*: As can be seen in the set of Tweets for companies with the most followers, more than half of the tweets pulled were retweets. These tweets tend to not be related to customer service and often do not provoke a brand response. Therefore further study would allow the removal of these tweets.

3. *Conversations are not always linear*: Due to their tree like structures, many conversations are complicated to analyze. It is important for us to focus on depth rather than number of tweets within a conversation to have a better idea of how many tweets are actually within one string of a conversation.

4. *Missing Tweets*: We also found that the way the data was pulled caused us to have almost half of our conversations to be incomplete. It will be important in future research to have a more complete pull of Tweets within conversations, by also pulling Tweet by in_reply_to ids. To put this in perspective, out of all conversations of size greater than one, we were missing 46%. By obtaining these missing tweets were were able to complete 47% of those incomplete conversations. Therefore, to have complete data, it will be necessary to complete these conversations.

5. *Industry Labels need to be consistent*: This wasn't directly discussed in this section, but a lot of work needed to be done on creating useful lists of the companies and industries involved. One particular issue we faced was that the company information we recieved for the two sets of companies used different sets of industry labelings. Given that many of our comparisons were cross industry it was crucial for these lists to be consistent.

Chapter 3

Airlines Industry Study

The airline industry is comprised of 500,000 Tweets—just over 1% of the size of the total data set. Statistical analysis on this industry would thus avoid complications associated with large data, whilst still giving a reliable insight into the nature of conversations in the total data set. The analysis was intended as a preliminary investigation into the types and structures of interactions between organizations and individuals on Twitter, and to some extent the effectiveness of these interactions. Differences between organizations within the airline industry are identified, as well as comparisons of generators they use. Furthermore, the results of the study will guide our analysis of the entire data set by providing insight into which trends appear to be the most useful in our overall analysis. Tools and programs created in the airline study will make analysis of the total data set faster and more efficient.

3.1 Descriptive Analysis

Approximately 13% of the 250,000 conversations in the airline industry have the first Tweet missing from the data set, due to the way the data was pulled and grouped. These conversations are excluded in all of the analysis below, primarily because the study is only a preliminary insight and thus dealing with them thoroughly, or attempting to recover them, was not deemed necessary.

3.1.1 Conversation Sizes

Descriptive analysis began with finding the distribution of features of conversations, such as the number of Tweets within a conversation (Fig. 3.1) and the length (in hours) of conversations (Fig. 3.2). Fig. 3.1 shows that almost half of all conversations are only one Tweet long — further investigations showed that 98% of these cases are Tweets by a user, mentioning an airline, to which the airline did not respond; these no-responses are investigated further in Section 3.1.2. Just over 15% of the remaining conversations are 2 Tweets, and further analysis showed that 86% of these conversations are initiated by a user and closed by an airline. These 2-Tweet conversations may represent a user's issue being resolved quickly and effectively, be of a conversational nature, such as a user thanking the airline for good service, or Tweeting that they are flying off on an exotic holiday, or may represent a failure of an airline to keep the user engaged in the conversation. The remaining fraction (14%) of 2 Tweet conversations were user-user, with no airline involvement. The figure is right-skewed, with a rapidly decaying tail for long Tweets.

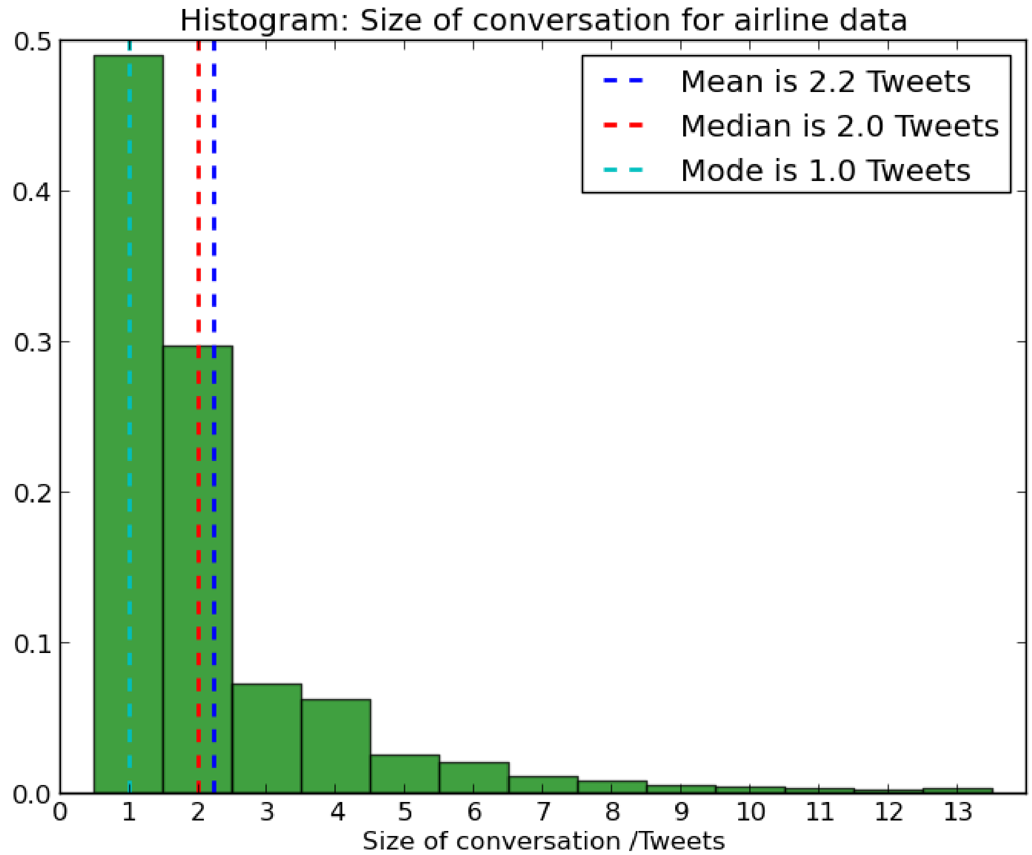


Figure 3.1: Airline Industry: lengths of conversations

Figure Fig. 3.2 excludes single Tweet conversations (since they do not include an interaction) and the histogram shows the expected shape; right-skewed with a peak at a short time (2.5 minutes). The median conversation was 40 minutes long, but the mean was much higher at over 7 hours—this is not surprising given that the longest conversation was found to be 57 days. The tail of the plot inspired further investigation which revealed that conversations extending over approximately 10 hours typically involved many individual users and were generally not of a customer service nature (Appendix). A typical type of such a long conversation would be an airline Tweeting 'Comment below for a free trip to Las Vegas!' which would inspire a flood of responses, or a Tweet from a celebrity at an airline prompting many unique users to respond.

3.1.2 Organizations and Generators: Times to First Response

Generators are software platforms used by organizations to interact on social media sites, such as Twitter. They are intended to enhance interactions with users, with an emphasis on resolving customer service issues quickly and effectively and maintaining positive customer relationships. Across the 10 organizations in the Airline industry study, 16 distinct generators are used.

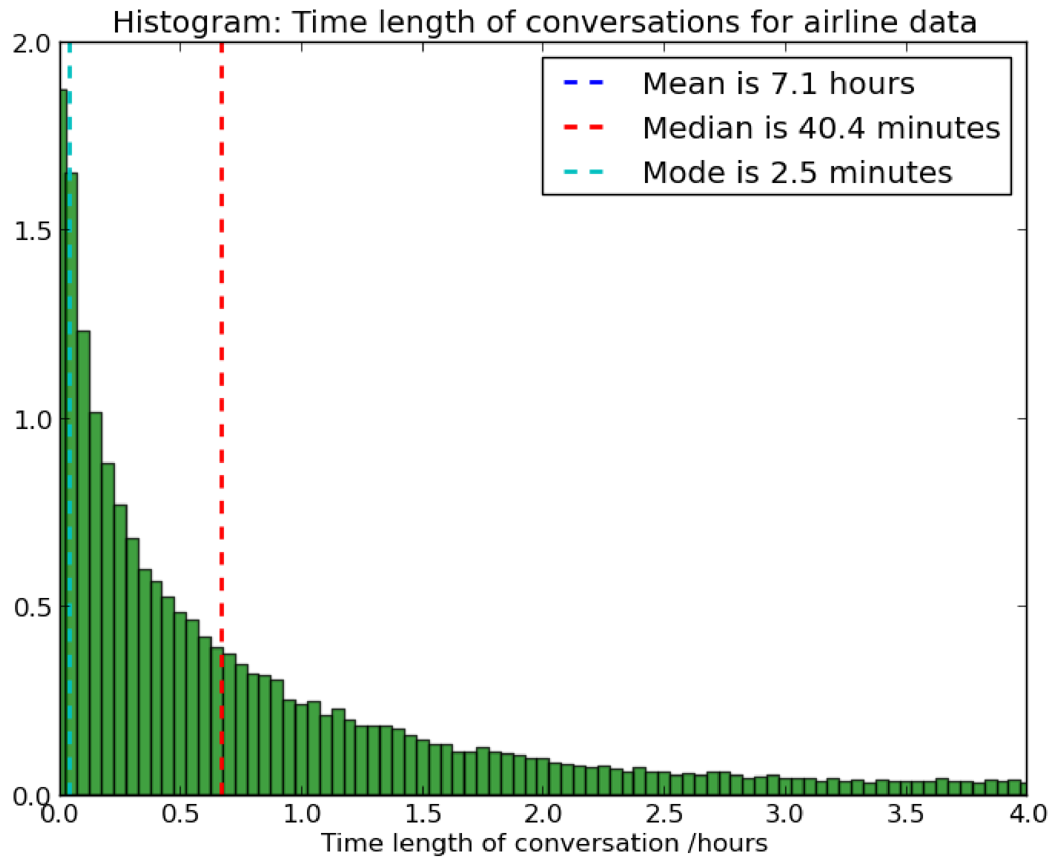


Figure 3.2: Airline Industry: time lengths of conversations

Analysis found that each organization uses one generator vastly more than any other; these 8 primary generators are displayed in red in Fig. 3.3. The other generators are used minimally by each airline. It is worth noting in this figure that Sparkcentral.com is used by Alaskaair, Deltaassist, and Jetblue, whilst every other organization uses a different generator. On the y axis, Fig. 3.3 shows the fractional usage of a generator by an organization, summed over all organizations. The plot is normalized to integrate to unity, and so the y axis can be interpreted as a measure of a generator's popularity. The most popular generator is Sparkcentral.com, whilst the three non-primary generators (Twitter for iPhone, Sprinklr and Adobe Social) have (as expected) the lowest values. Note that despite not being a primary generator for any of the 10 airlines, Sprinklr still shows a fairly high popularity; this is due to the fact that it is used a small amount by many of the organizations.

With this in mind, we can begin to compare how Twitter performance varies between organizations and generators. An interesting comparison to be made is the time it takes an organization to respond to an initial Tweet posted by a user—the hours to first response. The plot in Fig. 3.4 gives the hours to first response summed over all organizations and generators, and shows the expected shape — right-skewed with a peak at low time (1.5 minutes). (Note that some Tweets by users were never responded to, and these are not included in the data). Since the hours to first response can be interpreted as a preliminary indication of an organizations success in interacting with users on Twitter, it is informative to compare the performance across organizations (Figure Fig. 3.5) and generators (Fig. 3.6). These density plots show that out of all of the organizations Alaskaair, Jetblue and Deltaassist performed the best, with high peaks at low times and a rapidly decaying tails. A quick look at Fig. 3.3 shows that these organizations correspond to Sparkcentral.com, the highest performing generator in Fig. 3.6. Virginamerica, which uses @VirginAmerica as its primary generator, also showed a good response rate, whilst Americanair using SNAP100 showed consistently slow responses.

However, it is also informative to consider Tweets from users to airlines which weren't responded to; such situations make up over half of the total data set. Figure Fig. 3.7 shows the fraction of conversations in which a user mentioned an airline, and the airline responded — this is always less than 60%, and the average is 30%. (Note that although a similar comparison can be made for generators by studying Fig. 3.3, the results are not as meaningful since users mention airlines, not generators). **??** shows that the best responders are Deltaassist and Americanair, with Alaskaair and Jetblue following close behind. These last two, plus Deltaassist, all use Sparkcentral.com, the fastest generator to respond in Fig. 3.6. Conversely, although Figure Fig. 3.6 shows that Americanair is one of the slowest at replying, Figure Fig. 3.7 suggests that it is one of the most reliable responders. The figure also shows Flyfrontier to have a response fraction of zero — it did not reply to any of the 120 Tweets in which it was mentioned. Delta also shows a poor response fraction, however this is due to having a service handle (Deltaassist) which responds to its Tweets.

3.2 Conclusions and Future Direction

The airline study has demonstrated the nature of conversations between airlines and users; almost half the time users' Tweets at airlines are not responded to at all, and the remaining conversations tend to be short—on average 2 Tweets. The most common conversation time is consistent with this finding, at a brief 2.5 minutes, however the mean of over 7 hours indicates that very long conversations are also being held. These long conversations tend to be held by an airline and many unique users and are generally not of a customer service nature. Generator usage by organizations was also presented, and the differences in how Twitter usage varies between organizations and generators was analyzed. Studies into how quickly organizations respond to Tweets by users on Twitter gave a median of 20 minutes, and vast differences in this response time were observed between both companies and generators. The reliability of an organization to respond to a user's Tweet identified that the most consistent responders tended also to have been the fastest, but that this was not always the case.

Following on from this study, a useful direction would be to investigate the types of conversations being held between organizations and users, and the fraction of these which deal with customer service issues. It would be interesting to re-perform the above analysis on only Tweets of a customer service nature, and to compare the results with those presented above. Furthermore, it would be interesting to study how the speed and reliability of responses by organizations on Twitter relates to the resolution of the issue or customer satisfaction — whether a quick response correlates with a better overall outcome.

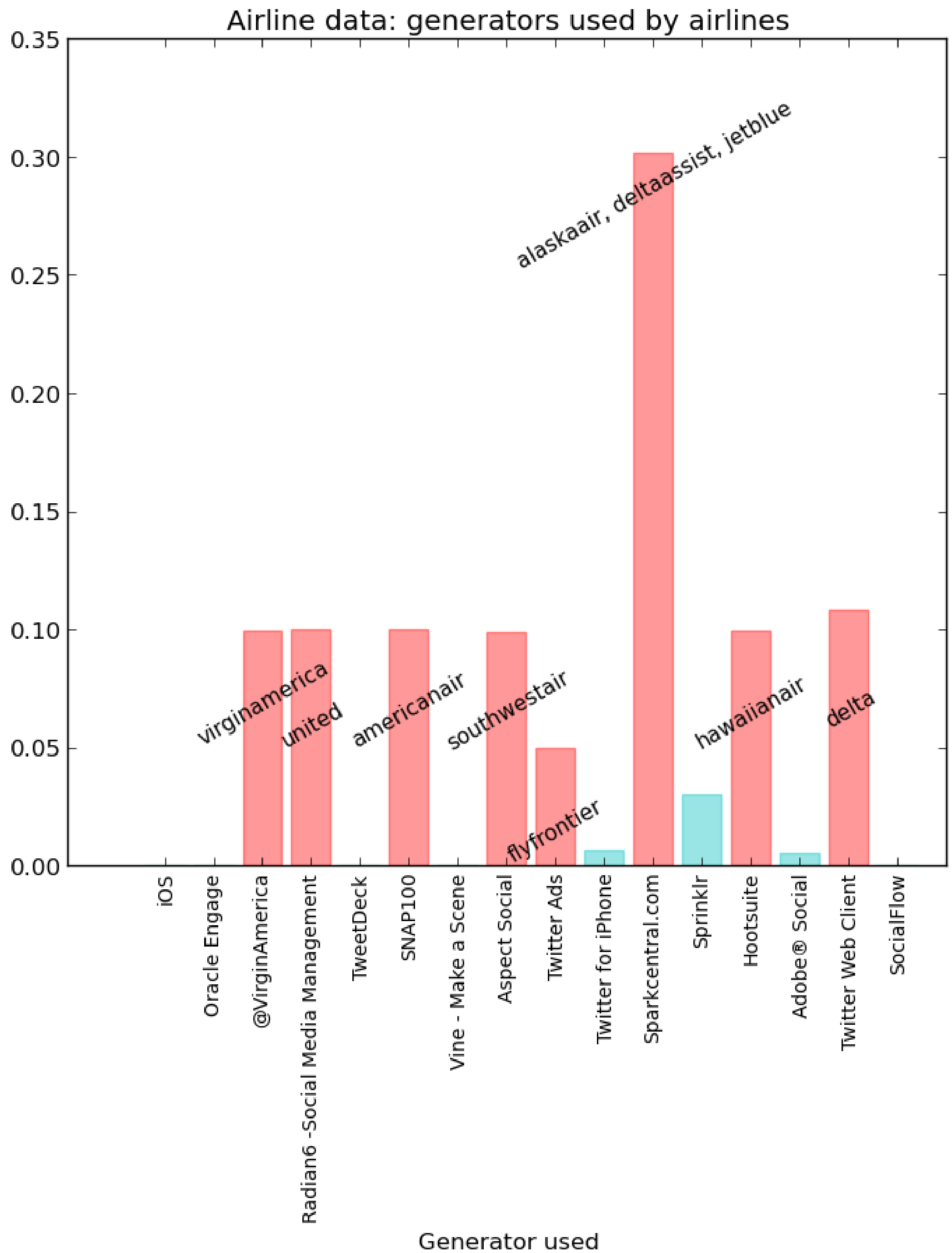


Figure 3.3: Airline Industry: generators used by airlines. Primary generators are shown in red, with the corresponding organizations that use them.

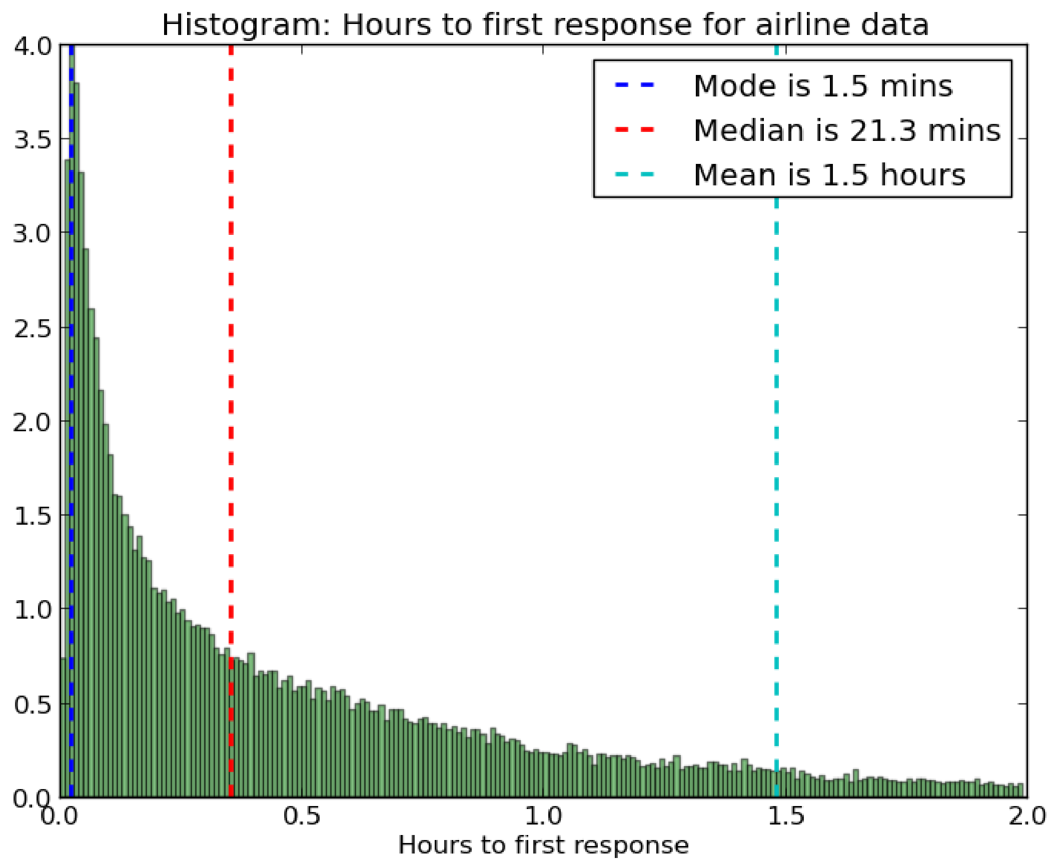


Figure 3.4: Airline Industry: hours to first response from airline to user.

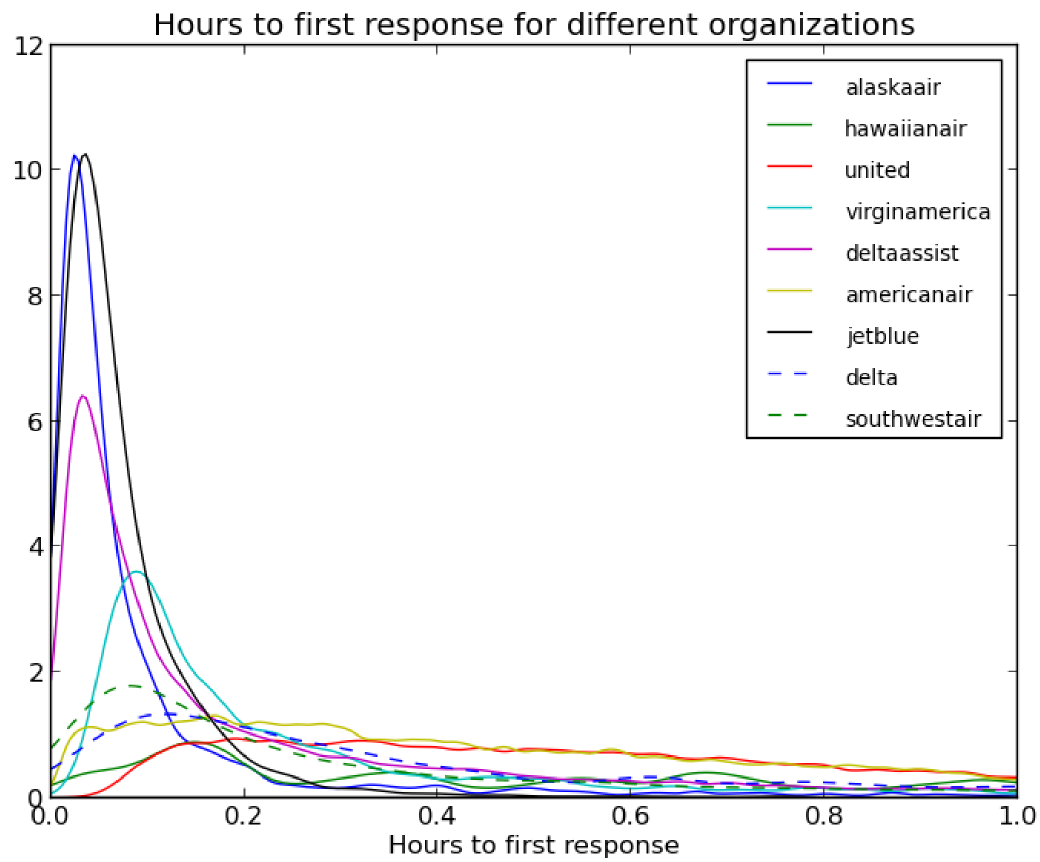


Figure 3.5: Airline Industry: hours to first response from airline to user for different airlines.

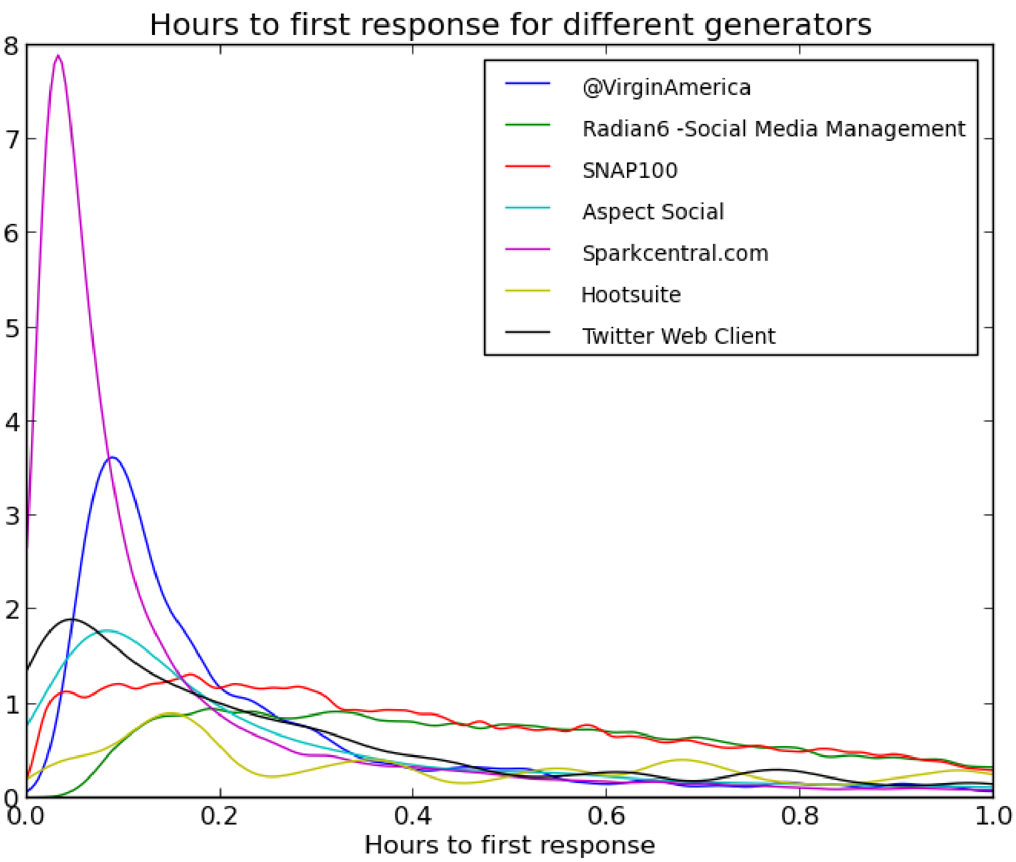


Figure 3.6: Airline Industry: hours to first response from airline to user for different generators.

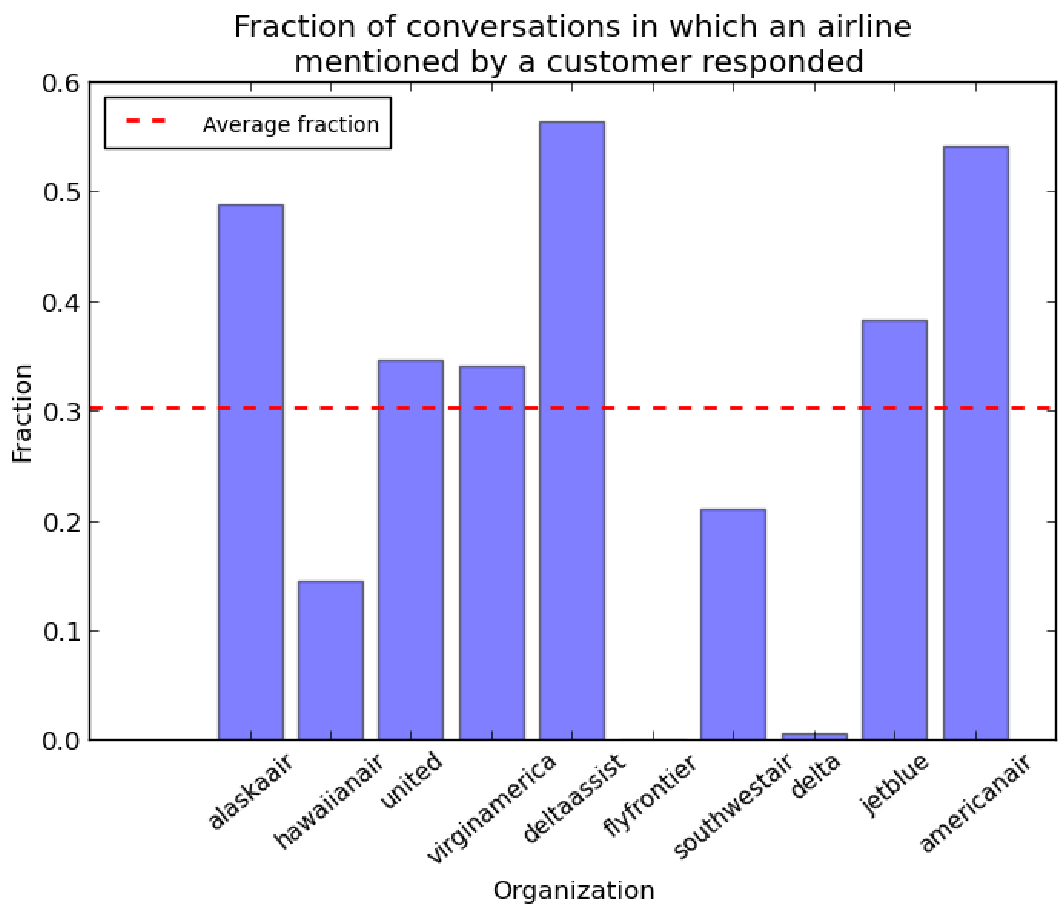


Figure 3.7: Airline Industry: fraction of conversations in which an airline mentioned by a user responded.

Chapter 4

Initial Organizations Study

The McKinsey list of companies consistsed of 366 companies and 117 care handles with a total of 483 handles. These companies ranged 19 industries using Twitter's labelings. If Twitter did not have a label either they were non verified users or private, we hand labeled the industry deciding by our knowledge of the company and where it best fit. To see the hand labeled companies see (APPENDIX)

From the 5.1 million Tweets to and from this company we obtained 3.7 million conversations (for more statistics on the conversations as a whole see Section 2.3.3. Allowing our findings from the Airline study to guide us we explored properties of these conversations as they vary by industry in order to better understand industry-wise use of Twitter.

4.1 Conversation Sizes

Although for most of the report we have referred to conversation size as the number of Tweets in a conversation, the size of a Twitter conversation can be interpreted in three different ways: number of Tweets within the conversation, depth of the conversation, and time length of the conversation (Section 2.3); these are all discussed below.

Number of Tweets in the conversation is plotted in Fig. 4.1 and shows a similar distribution to the results from the Airlines study (Fig. 3.1), however the mean number of Tweets is slightly lower at 1.7 rather than 2.2 for the airlines. Fig. 4.1 shows that almost three quarters of conversations are composed of only one Tweet; of this proportion, 99% represent an individual Tweeting at a brand and not being responding too, whilst the remaining 1% of conversations are a Tweet by a brand. Further time investigations showed that the percentage of all Tweets from users mentioning a brand are never responded to may be as low as 74%, however it is important to bear in mind that not all of these are of a customer service nature Chapter 5. Approximately 15% of remaining conversations are of length two Tweets.

Depth of conversation, shown in Fig. 4.2, gives a better idea of the length of an actual interaction between an organization and a user within a conversation (Section 1.2) because of the tree structure of conversations.

Finally, the length in time of the conversation is plotted in **??**.

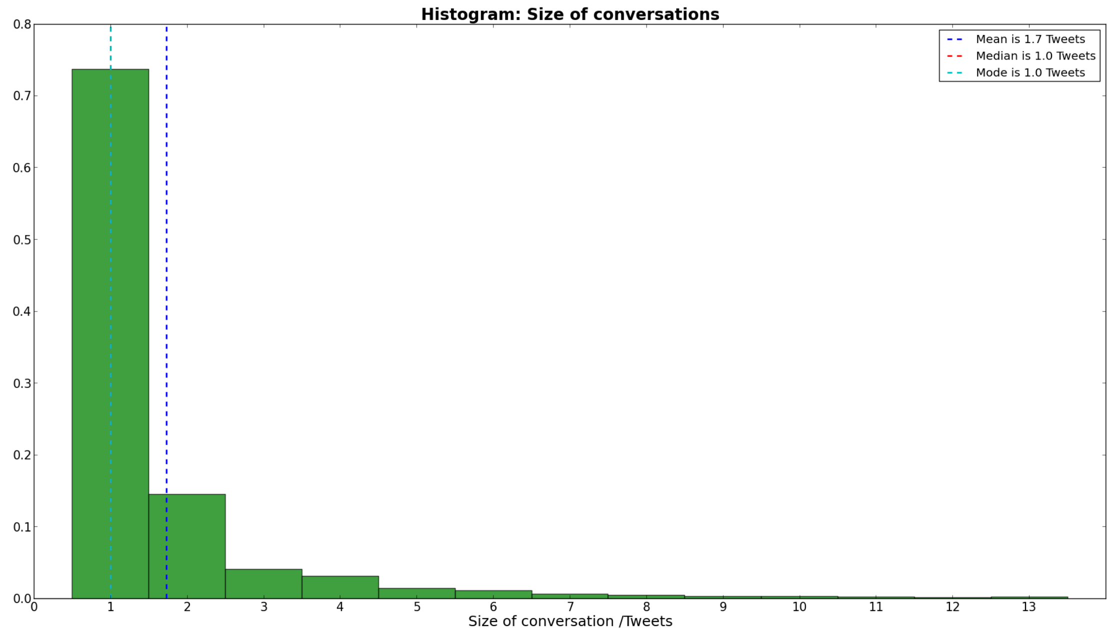


Figure 4.1: Number of Tweets in conversations

4.2 Organization Twitter Activity

Fig. 4.3 shows the total number of conversations in which an organization from a particular industry is responding, averaged by the number of organizations within that industry; this can be interpreted as the Twitter activity of the typical organization from an industry. The large differences in the number of conversations between certain industries show that certain industries, such as Travel, are using Twitter to interact with customers much more than, for example, News. These can mostly be rationalized; it would be expected that industries such as Travel and Telecomm receive many more complaint or enquiry (inquiry? What's the difference?) Tweets about their services from customers, than industries such as News and Entertainment. The latter industries may be expected to be using Twitter for reasons other than to interact with customers, such as to release news broadcasts or advertise their products. They may also receive a significantly higher amount of 'chatty' Tweets and retweets which will dilute any number of customer service Tweets that they may receive.

Travel and Telecommunication are identified from Fig. 4.3 as the most active industries—that is, the average company within those industries was involved in the highest number of conversations. These industries were thus selected for further study into the types of conversations in which they were engaging (chapter*).

4.3 Responding to Users

The way in which organizations and industries react to customer Tweets on Twitter can be explored in a number of different ways. As a preliminary overview of the relationship

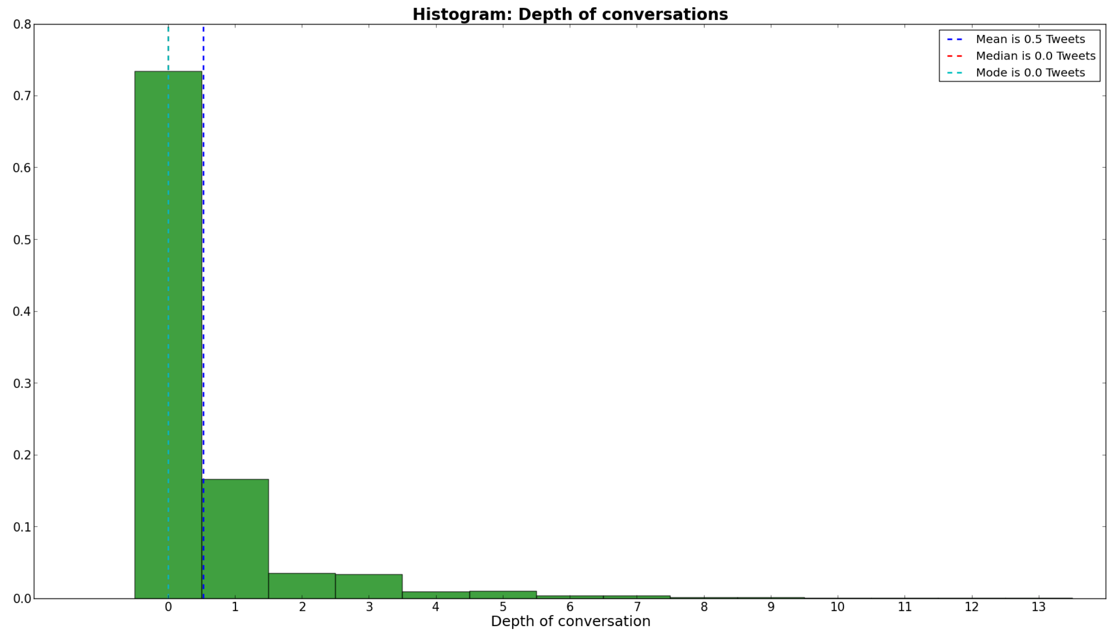


Figure 4.2: Depth of conversations

between user and organization Tweet activity, Fig. 4.4 shows the proportion of all Tweets in the dataset which are to and from an organization, for the 5 most active industries (defined by the number of conversations in which an organization from that industry is Tweeting). Clearly users are Tweeting at organizations much more than organizations are Tweeting at users, suggesting that not all Tweets by users at companies receive a response. From **??**, there does not seem to be any general relationship between the number of Tweets to and from industries- i.e. a high volume of Tweets at an organization does not necessarily imply that the organization is an especially active Twitter user. For example, over a quarter of all Tweets are directed from customers at the Retail industry, but it is the Telecomm industry, which receives only 15% of all Tweets, which Tweets the most.

4.3.0.1 Response Proportions

X% of conversations in the initial list are initiated by customers; this high proportion prompted analysis of how companies are responding to these interactions towards them. Fig. 4.5 shows the proportion of times a Tweet by a customer mentioning an organization is responded to by that organization (termed the *response proportion*), compared across industries.

Fig. 4.5 shows an average proportion of just under 10%, significantly lower than the 30% result from the Airlines study (Chapter 3 Fig. 3.7). However, the airlines dominate the Travel Industry which shows the highest response proportion of all industries, and so this difference is not surprising. Comparing the response proportion of the industries to Fig. 4.3 shows that in general, industries which Tweet the least tend to have the lowest response proportion. Entertainment, News, and Government and Politics are the worst responders in Fig. 4.5, and also engage in very few conversations. On the other hand, Travel and

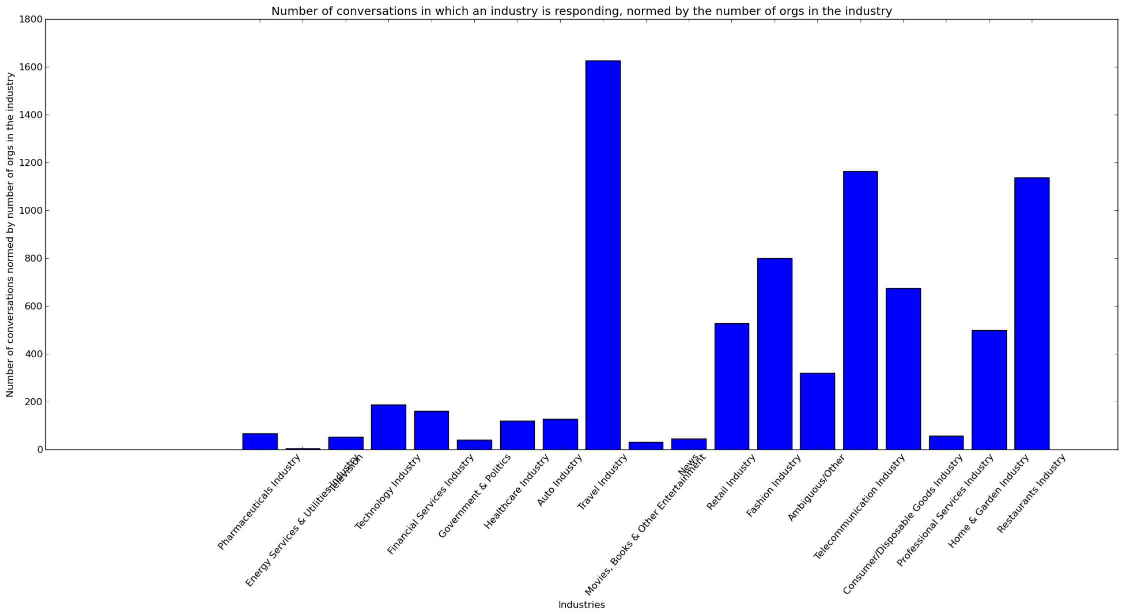


Figure 4.3: Industry Tweet amount per organization

Telecomm have arisen once more as the best responders across the industries. Again these results can be rationalized; one would expect Tweets towards the Travel and Telecomm industries to be of a more urgent customer service nature (e.g. delayed flights and service problems).

4.3.0.2 Hours to First Response

Concentrating now only on the Tweets to which an organization within an industry does respond, Fig. 4.6 gives the *hours to first response*—the time taken by an organization to respond to the first Tweet by a user. Whilst for some industries the plot shows the familiar right skewed distribution from the Airlines study (Fig. 4.6), for many industries there is a more level distribution of times.

Fig. 4.7 shows the hours to first response against the proportion of total conversations in which an industry is responding; there is no apparent relationship between these two variables.

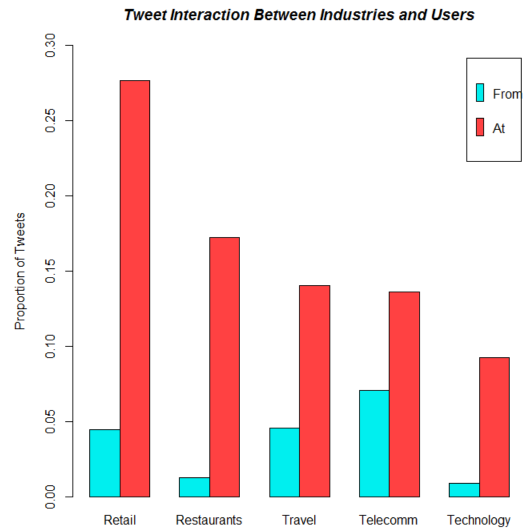


Figure 4.4

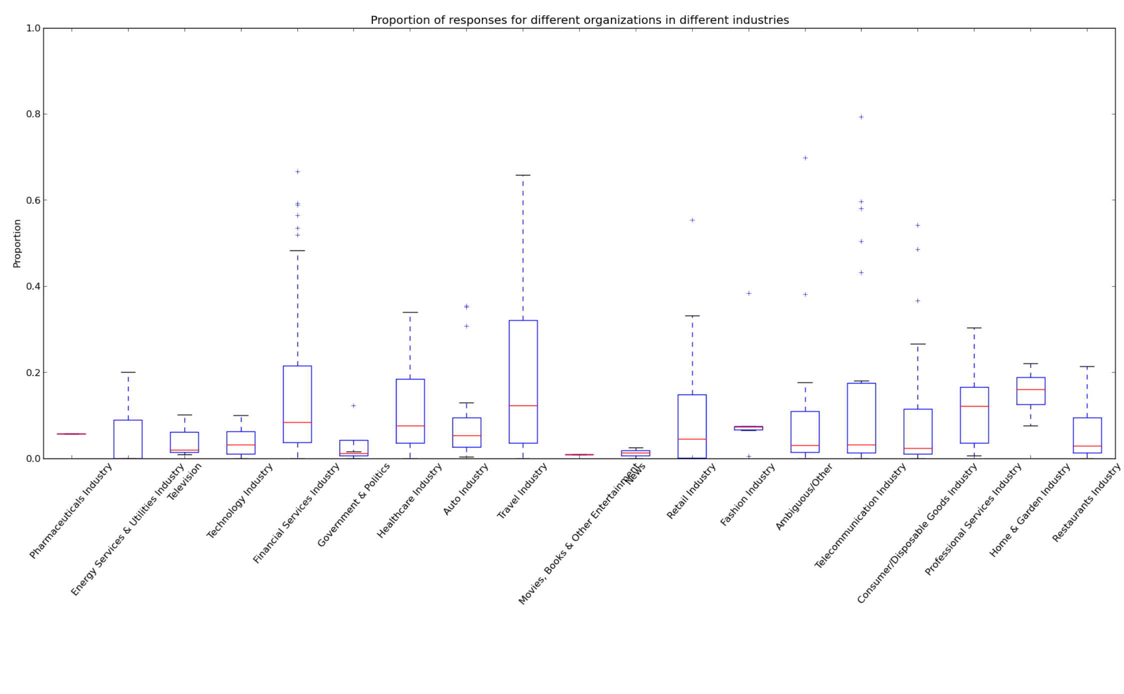


Figure 4.5: Proportion of responses when mentioned

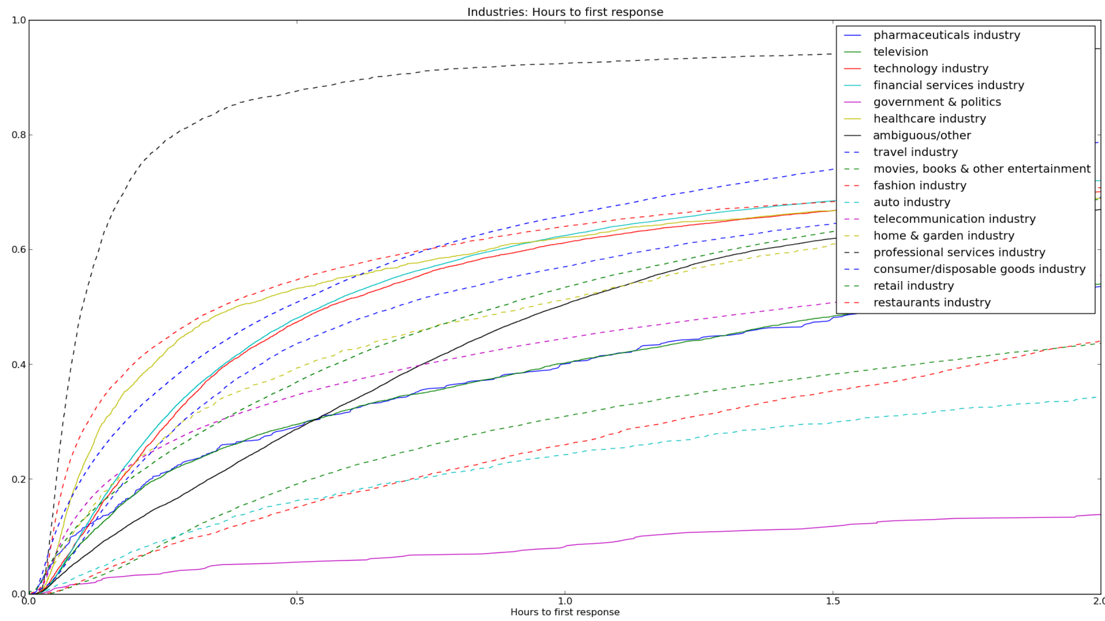
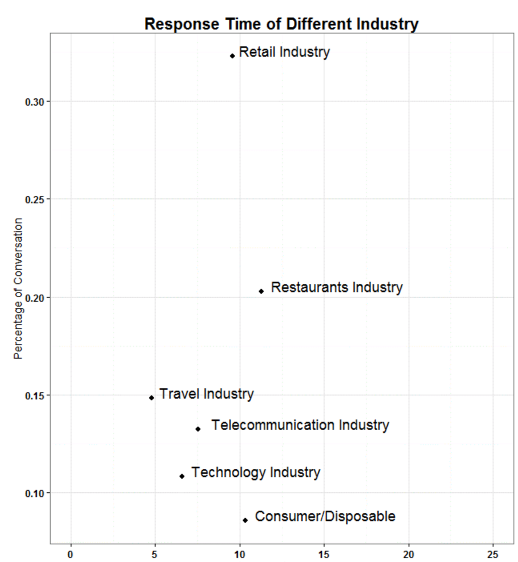


Figure 4.6: Hours to first response for industries



f

Figure 4.7

Chapter 5

Identifying and Categorizing Customer Service Conversations

One of our main purposes is to define resolution for the conversations happening between individual users and organizations. However, due to the wide variety of conversations happening on Twitter, not every conversation requires any form of resolution in the first place. Therefore, we have a three-step problem:

- Identify the customer service related conversations that need some sort of resolution among all the conversations
- Categorize all the customer service related conversations
- Visualize different categories and explore which factors influence a conversation's likeliness to be resolved versus unresolved

The first two steps are achieved primarily through extracting keywords from different parts of the conversations. We identify customer service related conversations by finding the keywords of first customer Tweets that signal questions or complaints from users; we categorize those conversations mainly based on the keywords extracted from last customer Tweets and last brand Tweets. The detailed description of how we generically perform keyword extraction once we have a specific corpus aggregated is outlined in the (Section 5.1) section below. The definitions of each category, how conversations are categorized are found in, and our exploration and visualization of these categories can be found in Section 5.3. Finally, an initial logistic regression analysis on resolved versus unresolved is included in future directions (Chapter 6).

5.1 Keyword Extraction

Once we have aggregated the desired corpus, usually in the format of a list of Tweet bodies that we are interested in, we perform the following preprocessing on the Tweet bodies.

- Remove the usernames
- Remove the URLs
- Remove the words containing numbers

- Remove non-ascii characters
- Remove the redundant white spaces

After we have cleaned the raw Tweet bodies by performing the preprocessing described above, we generate 3 sparse matrices of term frequencies for 1-word, 2-word, and 3-word phrases respectively, using the scikit-learn package in Python (NEED TO REFERENCE APPENDIX). The rows represent the Tweets, and the columns represent the phrases. Each entry in each matrix is the number of times the phrase appears in the corresponding Tweet. Note that for *unigram*, the calculation of 1-word phrase term frequencies, we remove the standard stop words for English, with the detailed list of stop words documented in scikit-learn. But we do not remove the stop words for *bigram* (2-word phrases) and *trigram* (3-word phrases), since stop words can be important constituents of potential keywords as well. For example, 'thank you' is likely to be a very frequent phrase, but 'you' is a stop word in scikit-learn. We also only consider phrases that appear more than once in the entire corpus.

After we generate the sparse matrices of term frequencies, we sum up each column to obtain the total number of times a particular phrase appears in the corpus. Thus we obtain a vector of total term frequencies for each phrase. We sort the vector and choose the top 100 most frequent 1-word, 2-word, and 3-word phrases respectively.

Now we delve into applying keyword extraction to identifying and categorizing customer service related conversations.

5.2 Identifying Customer Service Conversations

5.2.1 Method

We find the signal keywords for customer service related conversations by extracting keywords and phrases from first customer Tweets, since most of the time the concerns or complaints, if any, are voiced in the first customer Tweet.

We ask the following questions about each first customer Tweet in a conversation:

- Is it replied to by at least one brand?
- Is it a retweet?

While extracting keywords from the first customer Tweets, it is natural to assume that each industry has very different customer service topics, and the keywords would be different. Thus within a certain industry, we aggregate all the first customer Tweets that are replied to by at least one brand and are not a retweet. Note that a conversation started by a customer retweet is very likely to be about a trending issue, instead of being customer service related. More details about first customer Tweets that are not replied to by any brand are included in the discussion.

We obtain the top 100 most frequent 1-word, 2-word, and 3-word phrases respectively using the procedure described in the keyword extraction section (Section 5.1). For each of those phrases, we sample 10 conversations from MongoDB and read through the first customer Tweets of those conversations, and manually identify all the phrases that often correspond to concerns or complaints, i.e. customer service related. An example of a phrase that usually signals a need for resolution in the Travel Industry is 'can you help'. See the example Tweets below (Fig. 5.1).

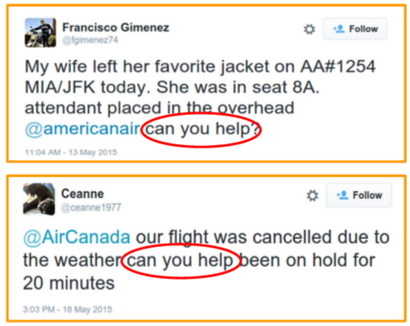


Figure 5.1: Examples of customer service related Tweets

We also identify phrases that, instead of being customer service related, often reflect trending issues or automatically generated Tweets. These Tweets usually share very similar format and wording, and are not customer service related. An example of a phrase signaling a trending issue for the Travel Industry is 'unitedfortahera' (see the sample Tweets in Fig. 5.2). In May 2015 a muslim women called Tahera was publicly discriminated against on a flight operated by United Airlines (Fig. 5.3). This became a trending topic on Twitter and dilutes the conversations with United with atypical conversations that are not customer service related, thus should be removed. Examples of automatically generated Tweets are in Fig. 5.4. In this case the signaling phrase is 'chance to win'. These cases also clutter our conversations and need to be removed.

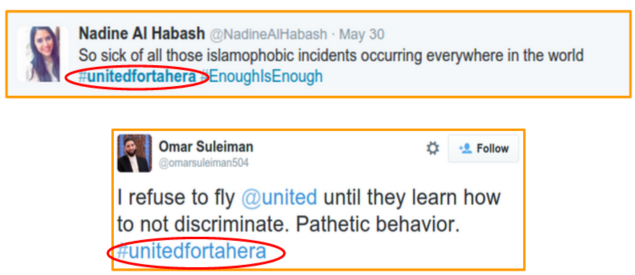


Figure 5.2: Examples of trending issue related Tweets

Once trending and automated keywords are identified we label them as 'bad phrases'. Then we perform a refinement process by calculating term frequencies again as described previously but now we filter out the first customer Tweets that contain the bad phrases. We iterate this refinement process until we do not see any bad phrases in our list of most frequent phrases. The primary reason for this refinement process is to identify the conversations that are clearly not customer service related, while we search for signaling keywords for customer service related conversations. Moreover, removing those bad phrases would leave more space for customer service related keywords in the list of top 100 we look at. In the end, we obtain a list of keywords that signal a need for resolution, as well as a list of bad phrases that indicate trending issues or automatically generated Tweets.

5.2.2 Discussion

Due to time constraints, we only performed this process for the Travel and Telecommunication Industries. These processes need to be done separately due to the differences in issues that need to be resolved. For example, 'cancelled' and 'delayed' are very frequent and informative for the Travel industry which primarily includes airline companies, while 'outage' and 'signal' are very frequent in the Telecommunication Industry which primarily includes phone and internet companies. This justifies why we extract keywords industry-wise. For complete lists of keywords for the two industries, see (NEED TO REFERENCE APPENDIX).

Our method of keyword extraction and manual categorization also allowed us to find one more condition for our customer service tweets. We found that there were a significant number of conversations that contained a shout out to a brand with a response from the brand often saying 'thank you' or 'happy to hear'. In an effort to remove these cases, we decided to add the final condition that if a conversation was of size 2 and the last tweet was by a brand with happy keywords then it is not of a customer service nature.

Through this method we finalized the rules for a conversation to be of a customer service nature. A conversation will be considered a customer service conversation if it satisfies all of the following:

1. The conversation begins with a customer tweet.
2. The conversation does not begin with a retweet.
3. The first Tweet of the conversation contains a customer service signal keyword.
4. The first Tweet of the conversation does not contain a trending or automatically generated keyword.
5. The conversation is not a shout out.

How we tested for items 3-5 and for a complete list of keywords can be found in **??**(APPENDIX).

(THIS IS FOR THE APPENDIX)

5.2.2.1 Not Replied to Tweets

We also performed keyword extraction for the first customer Tweets that do not receive a reply from any brand. We found that these tweets primarily consisted of retweets and automatically generated Tweets. Once we filtered out these Tweets we began again the process of identifying which Tweets signaled need for customer service. Our results showed that the Tweets that did not receive a reply but should have contained a subset of the keywords that we identified as signals of a need for resolution in our list of keywords for the replied first customer Tweets. For example, 'customer service' and 'delayed' are keywords for both replied and not replied groups of first customer Tweets. This suggests that by looking only at the Tweets that were replied to, we were able to obtain a thorough list of keywords that signaled resolution. It also implies that there are not specific issues that are being ignored by companies, but only an overwhelming number of Tweets where only a fraction of each issue is being addressed.

Upon further investigation, we were able to calculate the proportion of Tweets containing each keyword that receive a reply versus those that don't. In Travel industry 38% of the first customer Tweets that are not retweets are replied to by some brand. On the other hand, out of the all the first customer Tweets that contain the keyword 'delayed', about 60% are replied to by a brand. 56% of the first customer Tweets containing the keyword 'cancelled' are replied to by a brand. Similarly, in Telecomm Industry about 30% of the

Table 5.1: Customer Service Keywords and Rules: Telecomm

Customer service signal words	internet, phone, signal, time, help, going, cable, new, getting, working, need, data, wifi, area, tv, day, pay, today, work, having, days, fix, guys, outage, watch, month, trying, network, issues, hours, mobile, know, problem, broadband, use, week, media, calls, box, account, connection, issue, paying, slow, problems, home, house, bad, times, speed, app, customers, won, isn, switch, number, says, iphone, minutes, think, update, fixed, hour, customer service, my phone, my internet, is there, on my, to get, is the, to be, trying to, can you, can you, has been, you guys, this is, going on, no service, in my, service is, have to, with the, need to, going to, internet is, why is, the phone, no signal, want to, my bill, can get, do you, are you, not working, will be, outage in, of my, the internet, your service, to watch, to my, with my, is down, how do, what is, and it, is not, my service, ve been, have no, no internet, on hold, what the, right now, service in, for my, you re, what going, able to, the last, to do, to pay, my account, to fix, have been, pay for, out of, there an, up with, to call, all day, paying for, is this, tell me, to have, an hour, down in, my cable, is going, an outage, at all, up on, what going on, on the phone, is there an, in my area, going on with, is going on, on my phone, there an outage, an outage in, why is my, in the area, internet has been, can you help, sort it out, on hold for, has been down, internet is down, what is going, is there problem, for the past, trying to get, for the last, hung up on, is not working, can you please, be able to, going to be, trying to watch, do you have, pay my bill, been down for, is there any, no signal in, been on hold, why does my, the middle of, to watch the, in the middle, have no service, to switch to, you tell me, my internet has, why is it, to get my, my phone is, you guys are, to pay my, no service in, to cancel my, been trying to, can you tell, on and off, have to pay, any idea when, seems to be, why do you, internet outage in, over an hour, having problems with, order the fight, there problem with, it would be, no signal for, in and out
Trending and Automated Keywords	nfl, just voted for, chance to win, proud to support, download it from, speedtest result
Shout Out conditions	Conversation Size = 2 last Tweet by brand
Shout Out keywords	wonderful, glad, happy, great, fantastic
Shout out veto keywords	not wonderful, not glad, not happy, not great, not fantastic

Table 5.2: Customer Service Keywords and Rules: Travel

Customer service signal words	?, time, delayed, customer service, help, hours, check in, hour, gate, delay, bag, waiting, luggage, trying, worst, boarding, cancelled, minutes, seats, upgrade, phone, booked, bags, free, wait, lost, want, book, late, ticket, online, delay, baggage, stuck, week, can you, trying to, need, is the, do you, on hold, is there, is it, how do, what is, how can, what going on, you tell me, would like to, be able to, is going on, it possible to, on the runway, going on with, to change my, have to pay, how do get, hung up on, what is going, what up with, you help me, still waiting for
Trending and Automated Keywords	tahera, shame, discrim, coke, opened, islam, wifi at, autistic, kicked off, way to go, internetest, my favorite rock, shreveport, chance to win, missed his flight, in las vegas, animal hunting trophies, daily is out
Shout Out conditions	Conversation Size = 2 last Tweet by brand
Shout Out keywords	wonderful, glad, happy, great, fantastic
Shout out veto keywords	not wonderful, not glad, not happy, not great, not fantastic

first customer Tweets that are not retweets are replied to by some brand. For the keywords, 75% of the first customer Tweets containing 'internet outage' are replied to, and 61% of the first customer Tweets containing 'can you help' are replied to. This indicates that there is still much room from improvement for both Travel and Telecomm Industries and supports the findings made by the McKinsey Co. Section 1.4. However, this method does suggest a better outlook on the response rates of companies by filtering those conversations that do not require a response as the response rates tend to be higher for those Tweets that signal customer service need. For more details about the response rates of keywords for Travel and Telecommunication Industries, the reader can refer to the .csv files (NEED TO REFERENCE APPENDIX).

5.3 Categorizing Customer Service Conversations

Now that customer service related conversations can be identified, we can establish rules to categorize these conversations into the 5 categories defined in the next section.

5.3.1 Defining Customer Service Categories

We categorized conversations in 5 distinct categories:

Continued on platform: The conversation is moved to a direct message (Fig. 5.5).

Moved off platform: The brand asks to continue the conversation via phone, email, online forms, or directs the user to a physical entity, such as a store or gate agent

Table 5.3: Category Rules: Direct Message

	Travel	Telecomm
Refinement Rules	None	None
Keywords for any Tweet	dm, dm/follow, follow/dm, /dm	dm, dm/follow, follow/dm, /dm

Table 5.4: Category Rules: Moved off platform

	Travel	Telecomm
Refinement Rules	Any brand Tweet has URL or phone number.	Any brand Tweet has URL or phone number.
Keywords for any brand tweet	email, link, e-mail,contact, call, links, ll forward, ll pass, ll be sure, see an agent	email, link, e-mail, contact, call, links, here if you, your zip code, for letting us, us we can, local office, to see if, us know how, so we can

(Fig. 5.6).

Resolved: The brand addresses the user's issue and the user expresses a positive response or a user asks a question and receives an answer from the brand (Fig. 5.7).

Unresolved: Questions remain unanswered or the user expresses dissatisfaction with their interaction with the brand (Fig. 5.8).

Other: This category consists of conversations that do not fit into the first four categories. This consists of replies from brands that do not actively seek to resolve an issue such as apologies, and generic answers that reach out to customers but do not provide resolution.

These particular categories are interesting to us because we want to get insight on how often conversations are being taken off public Tweets and in what forms as well as how successful companies are when they choose to continue publicly.
(THIS IS FOR THE APPENDIX!) To categorize Travel we used the following rules

Other Categorized by: Conversation makes it through filter but does not get categorized in moved off, continued on, resolved, or not resolved.

5.3.2 Method

Once we have extracted the conversations that signal a need for customer service, the next step is to find the keywords that signal that the conversation belongs in each category. To do so using keyword extraction **??**, we pull the keywords for the last brand and customer tweets for the two cases: 1. *The last tweet is by a brand.* 2. *The last tweet is by a customer.* For each keyword or phrase, we sample 10 conversations from MongoDB to reference as we manually categorize the keywords. By reading the conversations, we get a better understanding of these conversations and develop 'refinement rules' for more accurate

Table 5.5: Category Rules: Resolved

	Travel	Telecomm
Refinement Rules	Conversation is of size 2 and the first customer tweet contains a question mark and a brand responds.	Conversation is of size 2 and the first customer tweet contains a question mark and a brand responds.
Keywords for the last brand tweet	my pleasure, refunding, refund, wonderful, fantastic, great, happy, glad, re welcome, enjoy your, for flying, enjoy the, look forward to, we look forward, were able to, enjoy your flight, air traffic control, :), :-), ;), ;-)	my pleasure, at this time, let us know, if you need, glad to hear, we can help, thank you for, in your area, to look into, you need anything, me know if, please let me, to hear it, there is anything, you need any, be happy to, keep us posted, if you ever, out to us, you ever need, we can assist, happy to help, be in touch, here for you, you have any, assist you with, able to get, in the future, you need us, :), :-), :-), ;), ;-)
Keywords for the last customer tweet	wonderful, thank, thanks, thx, thnks, help, for the quick, the quick response, for the info, for the update, forward to it, looking forward to, the best, great, fantastic, happy, glad, :), :-), ;), ;-)	wonderful, thank, thanks, thx, thnks, help, for the reply, seems to be, for your help, you so much, you for your, let me know, you for the, great, fantastic, happy, glad, you very much, back to me, taken care of, glad to hear, have to wait, getting back to, so much for, the rest of, :), :-)
Veto Keywords	not wonderful, no thank, not thank, no thanks, not thanks, no thx, not thx, not thnks, no thnks, not help, no help, t help, not grea, not fantastic, not happy, unhappy, not glad	not wonderful, no thank, not thank, no thanks, not thanks, no thx, not thx, not thnks, no thnks, not help, no help, t help, not grea, not fantastic, not happy, unhappy, not glad

Table 5.6: Category Rules: Unresolved

	Travel	Telecomm
Refinement Rules	Last tweet contains a question mark.	Last tweet contains a question mark.
Keywords for the last customer tweet	unhappy, frustrating, frustrated, angry, disappointing, disappointed, sorry, but, sorry but, trying to, can you, on hold, fuck, waiting for, at the gate, on the plane, sitting on the, you need to, can you please, over an hour, there is no, to speak to, trying to get, was supposed to, to find out, worst, fail, annoying, hate, never, :(, :-(	useless, in my area, this is the, you guys are, you need to, it would be, not good enough, to fix it, have to wait, just wanted to, to talk to, in the middle, off and on, there is no, to do that, it off and, fuck, unhappy, frustrating, frustrated, angry, disappointing, disappointed, 'sorry, but', sorry but, trying to, can you, on hold, waiting for, can you please, over an hour, there is no, to speak to, trying to get, was supposed to, to find out, worst, fail, annoying, hate, never, :(, :-(

categorization. For example, conversations of length 2 with a question mark in the first customer Tweet and a brand replies are very likely to be in the resolved category since the brand is likely to be answering the customer's question. Meanwhile, if the last Tweet in the conversation has a question mark, then the conversation is very likely to be unresolved because either the customer or brand's question went unanswered.

Since each industry tends to use Twitter differently and resolve different issues, we found it necessary to analyze each industry separately. Due to time constraints, we only categorize the customer service related conversations for the Travel and Telecommunication Industries. The complete list of keywords and rules for each category by industry can be found in the **??**(APPENDIX).

Once we develop these rules, the conversations can then be categorized. All conversations within an industry will be filtered for a customer service signal (Section 5.2 PERHAPS APPENDIX?) and then categorized. Each conversation goes throufh the flow chart in Fig. 5.9. There is an inherent hierarchy in the order that conversations are categorized, which also allows for no conversation to be labeled more than once. The order was determined by matter of importance to our research (move off and dm are filtered first because they are of interest to Twitter), and which was most likely to be correct if it belonged to more than one category. For further information on how these conversation are categorized, see **??** (CITE CODE LIST THINGY)

5.4 Data Visualization

5.4.1 Comparing the Travel and Telecomm Industries

Having identified the customer service conversations and categorized these conversations into the five categories discussed previously, analysis on the categorized conversation data commenced. This began with Fig. 5.10 which shows the total number of conversations in each category for the Travel and Telecomm industries. Since the Travel industry is larger than the Telecomm industry by 12,000 Tweets, it is more informative to study Fig. 5.11 which gives the conversation category counts as a proportion of the total number of conversations.

Fig. 5.11 shows that the proportion of resolved conversations in the Travel industry is almost twice the proportion in the Telecomm industry, and the proportion of Not resolved conversations is slightly lower. It can thus be concluded that in terms of resolving customer service issues on Twitter, the Travel industry is stronger than the Telecomm industry. This result is not surprising; it is likely that the travel-related issues which customers raise on Twitter are generally of a more urgent nature than those related to telecommunication.

5.4.2 Travel Industry: Conversation Categories Visualization

5.4.2.1 Conversation Category Proportions

Efforts will now be concentrated on exploring the variation in properties of different conversation categories in the Travel industry, which is composed of 50,000 conversations. Following on from Section 5.4.1, **??** shows the proportion of conversations of different categories for the travel industry. The largest category is resolved conversations—over a quarter of customer service issues raised by customers are resolved within the conversation itself. Approximately 45% of conversations continue elsewhere, either remaining on Twitter through direct messages (Continued on platform) or moving off Twitter (Moved off platform). Less than 10% of conversations are not resolved, and show no indication of being continued somewhere else.

The proportion of conversations of different categories sheds light on the way that Twitter is being used. Thus an interesting direction to explore would be how the proportion of conversations of a particular category vary between the organizations within the travel industry. Before exploring this further, Fig. 5.12 shows the number of conversations in which a particular organization from the Travel industry is responding. Note that all conversations involving Tweets from Travel industry organizations were run through the categorization process (**??**), and since it is possible for a Travel organization to be interacting in the same conversation as an organization from a different industry, Fig. 5.12 actually contains some non-Travel industry companies as well. Fig. 5.12 can be interpreted as representing the activity of particular organizations, and clearly this varies enormously between different organizations. Therefore, for the sake of simplicity, in the following discussion we choose to concentrate upon the most active organizations, making an arbitrarily chosen conversation cut-off at 2000 conversations per organization. Fig. 5.12 shows that 7 organizations pass this cut-off, which are all incidentally airlines.

The types of conversations in which these 7 airlines are responding, as a proportion of the total number of conversations in which they are responding, are shown in Fig. 5.13. The proportion of Resolved conversations is at least double Not resolved for all organizations, suggesting that these 7 airlines are doing well at resolving customer service conversations. Comparing the heights of Continued on platform and Moved off platform does not show any general favor for either; the relative proportions vary between organizations, and are on average the same.

In a sense, it is possible to view in which an organization invites the customer to continue the conversation elsewhere (either on or off Twitter) as being successful, or at the very least, having some sort of resolution. Drawing from this idea, **??** shows the proportion of conversation categories again, but grouping Resolved, Continued on platform and Moved off platform into the single category, Resolved or continued elsewhere. Then, finding the ration between the new category 'Resolved or continued elsewhere' and 'Not resolved' gives a measure of how well an organization is using Twitter to resolve customer service issues- in this sense, Easyjet is doing the best, whilst Jetblue is doing least well.

Generators are software platforms developed specifically for organizations to interact with users and conduct customer service on social networking platforms, such as Twitter. Previous studies (**??**) have indicated that the generator used by an organization affects the nature and success of the conversation on Twitter, and thus a natural extension to the above discussion is to consider how conversations vary between generators used. First, however, Fig. 5.14 shows the number of conversations in which a particular generator is used— clearly there is a lot of variation between the 'activity' of generators. As with the organizations, we choose to look at generators which are involved in at least 1000 conversations, leaving 8 generators to study further.

Fig. 5.15 shows the proportion of conversations of each category for the most active 8 generators. This plot shows very similar properties to Fig. 5.13; the number of 'Resolved' conversations is at least double the number of 'Not resolved' for all generators, and there is a lot of variation in the relative proportions of 'Continued on platform' and 'Moved off platform'. The similarities between the two Figures is not surprising, since in general particular organizations tend to favor a particular generator well above all others, and generally these 'primary generators' are different for different organizations. For the airlines, the primary generators are discussed in Section 3.1.2. As a measure of how well different generators are used to resolved customer service issues on Twitter, the ration between the sum of the categories 'Resolved', 'Continued on platform' and 'Moved off platform' (see Fig. 5.15) to 'Not resolved' was found. Lexer Exchange was found to have to be the best generator in this sense, with the highest ratio, whilst SNAP100 had a higher proportion of 'Resolved' and a higher proportion 'Not resolved' conversations, giving it the lowest ratio of all.

5.4.2.2 Conversation Sizes, Lengths and Depths

Next it is interesting to investigate how conversation sizes, lengths and depths vary between conversation categories; these are shown in Fig. 5.16, Fig. 5.17 and Fig. 5.18 respectively. Besides from reflecting the familiar properties of conversations— a small median and mode, and a right-skewed shape— these plots do not show notable differences between the conversation categories. This itself is interesting— one cannot conclude, for example, that a long conversation implies a resolution, nor does a small number of Tweets within a conversation indicate a lack of resolution.

In order to investigate whether any relationship between conversation category and conversation is hidden by variations in organizations' use of Twitter, Fig. 5.19 shows the median conversation length (in hours) for each conversation category, for the 7 most active organizations in the Travel industry. Once again, there is no clear pattern in the relative lengths of conversations of different categories.

A natural next step is to consider the proportion of Tweets in a conversation contributed by organizations or users. Fig. 5.20 shows a CDF of the proportion of Tweets in a conversation that are contributed by an organization. The conversations which end in a resolution tend to contain a higher proportion of Tweets from organizations, and this is in line with our intuition; if an organization is more active within a conversation, it suggests that more attention is being paid to the user's concern, and therefore the issue is more likely to be resolved.

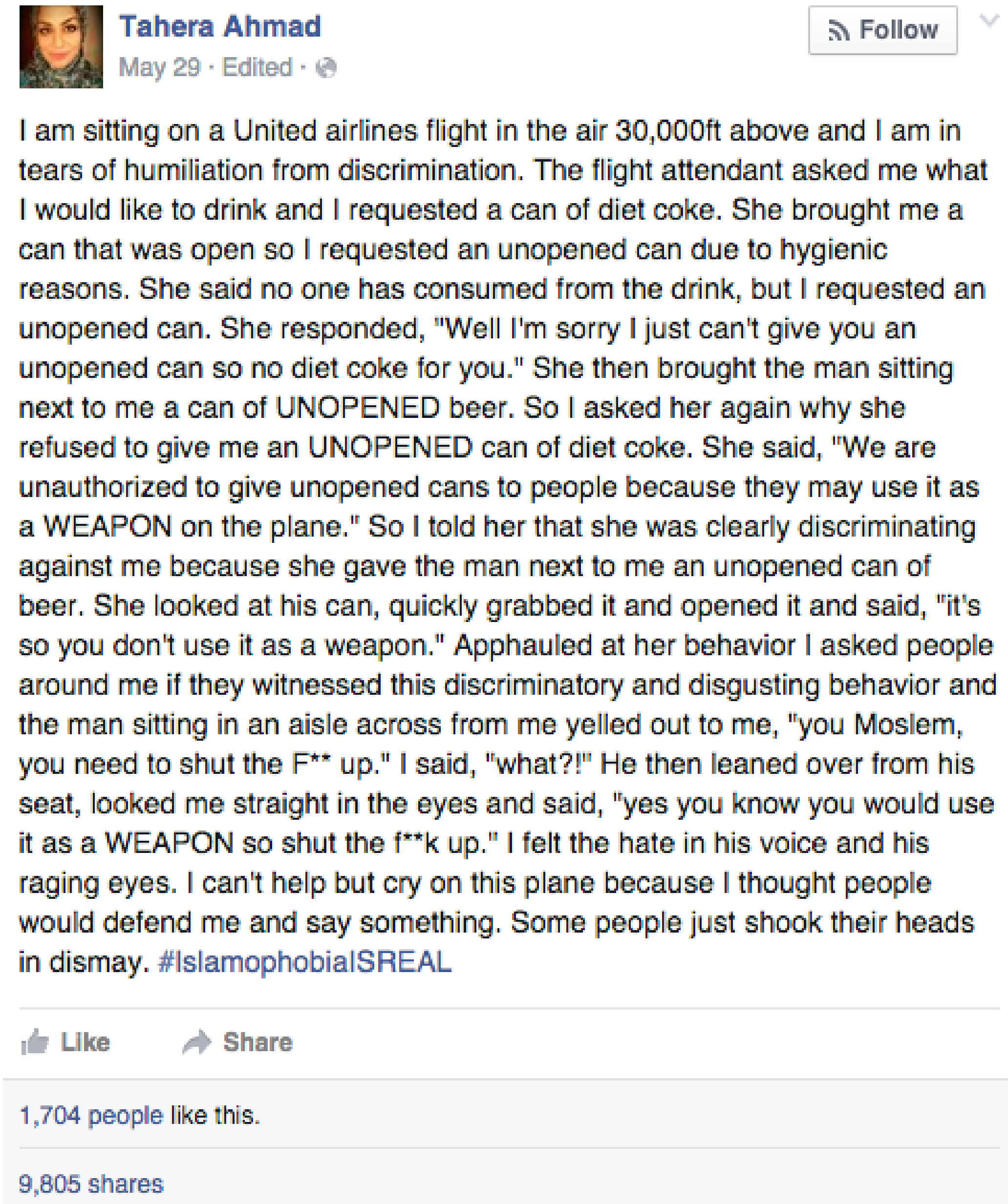


Figure 5.3: Tahera Event in May 2015

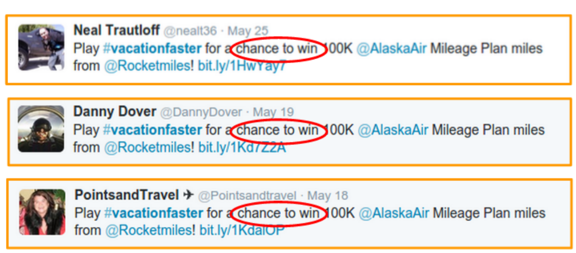


Figure 5.4: Examples of automatically generated Tweets

Figure 5.5: Example of a Direct Message conversation

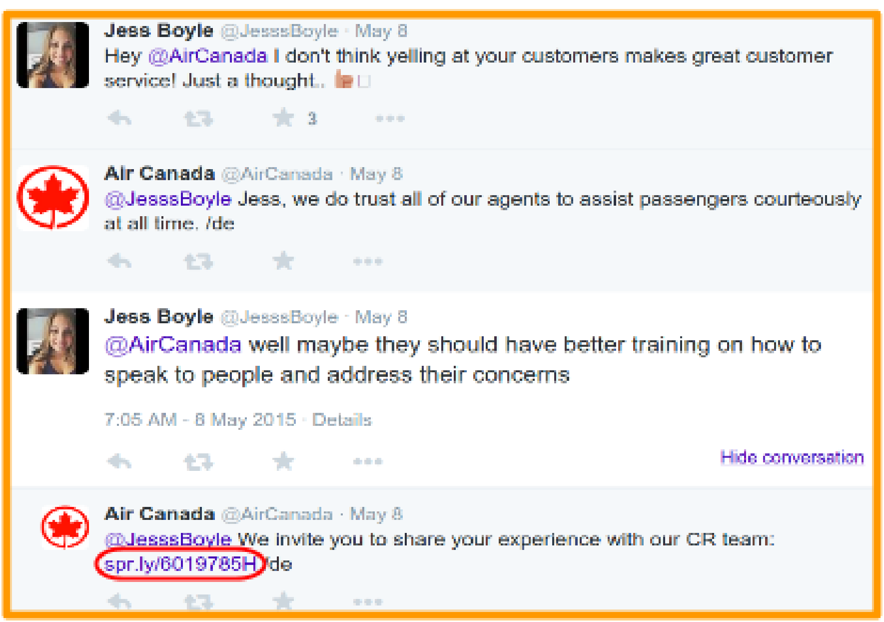


Figure 5.6: Example of a Moved off Platform conversation

Figure 5.7: Example of a Resolved conversation

Figure 5.8: Example of an Unresolved conversation

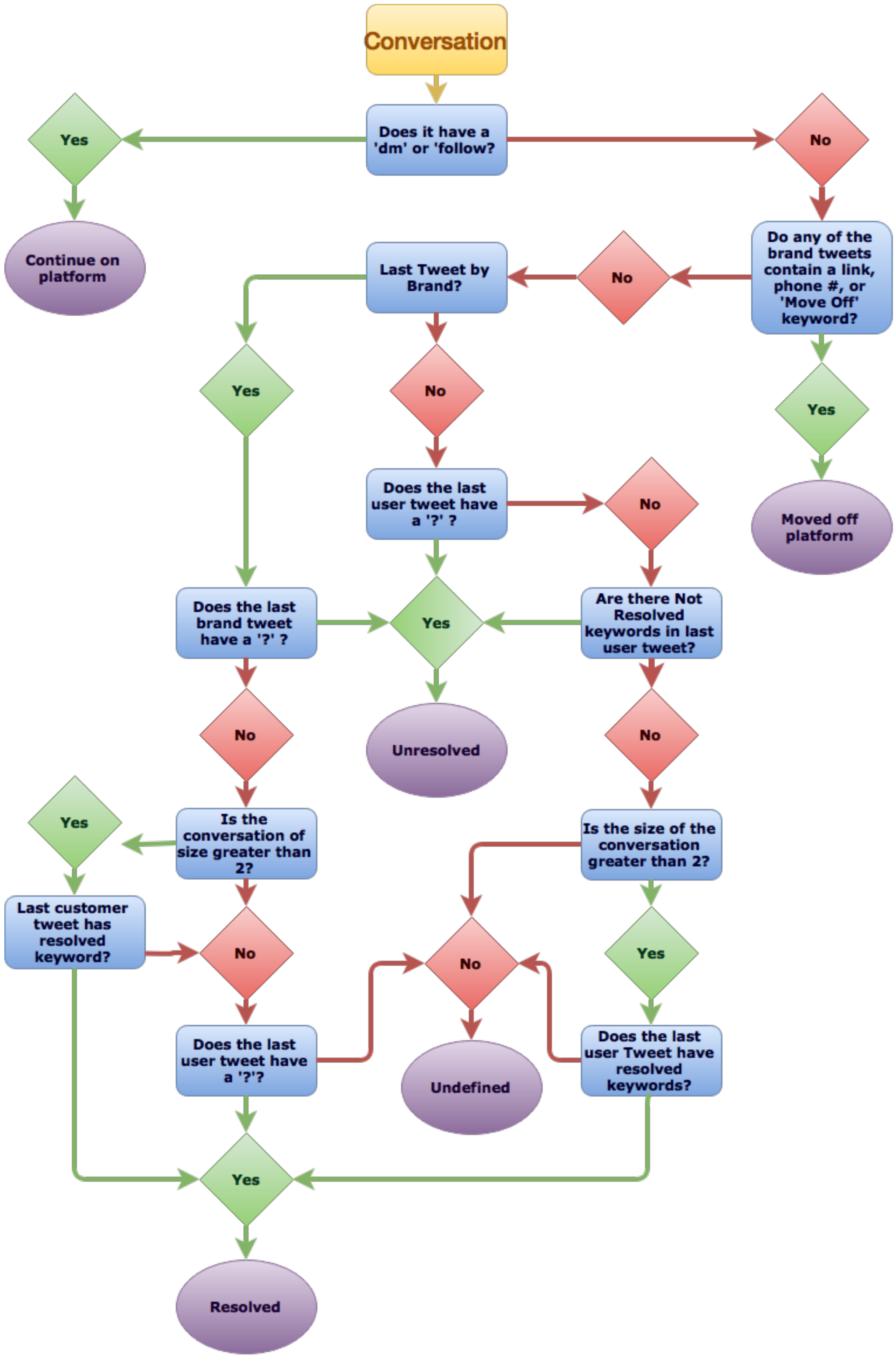


Figure 5.9: Flow Chart for Categorization

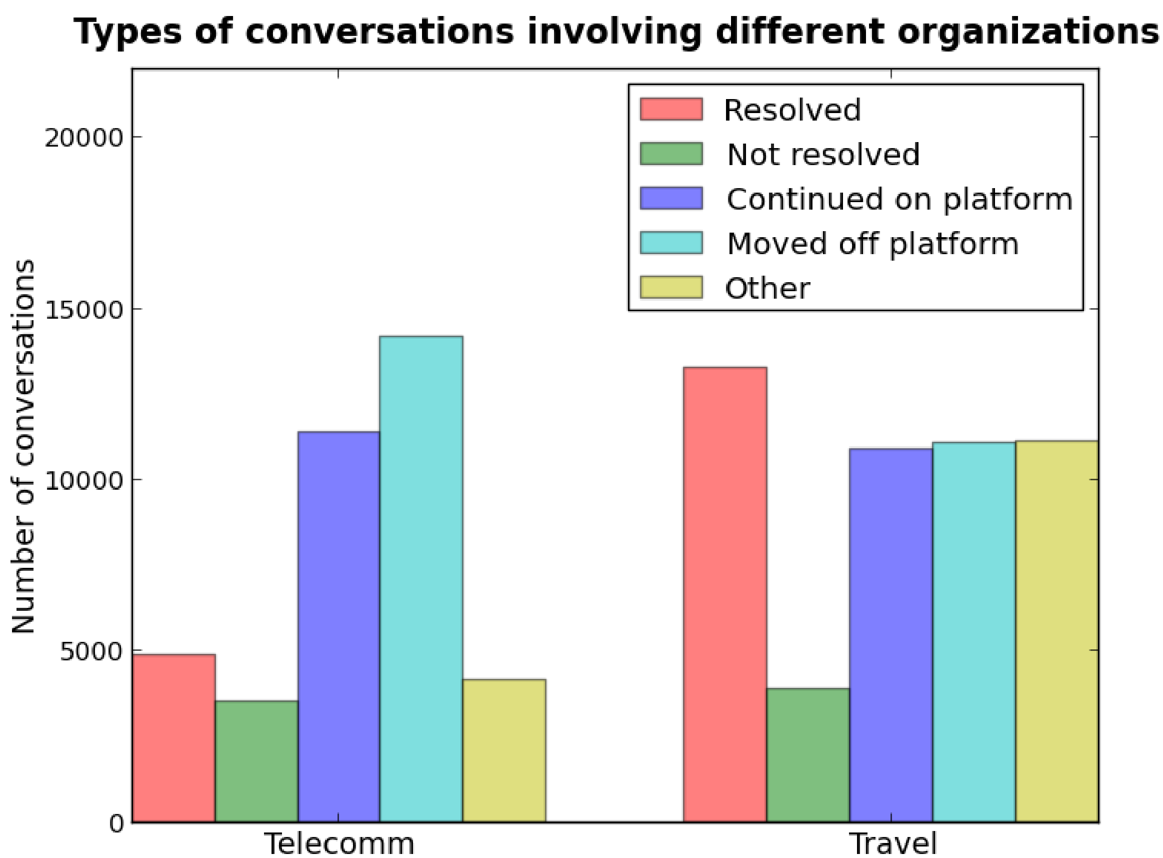


Figure 5.10: Categories: Basic category count

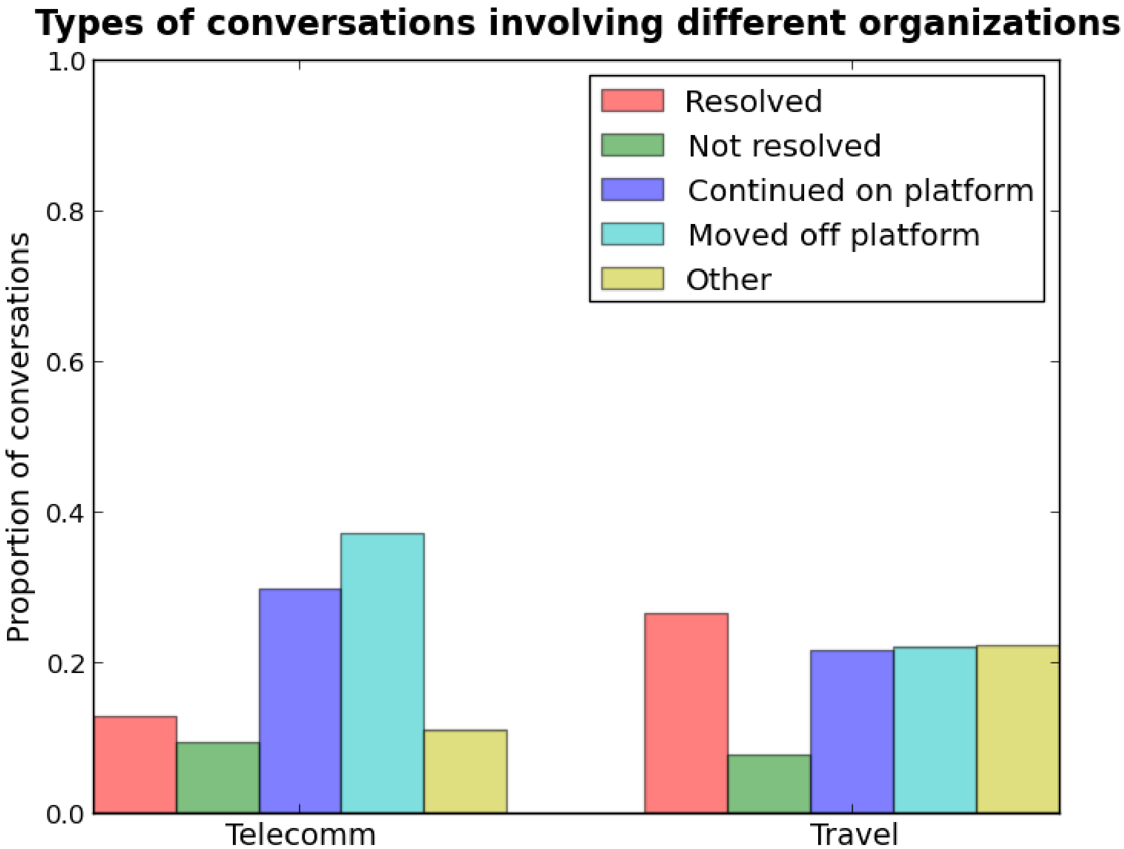


Figure 5.11: Categories: category proportions

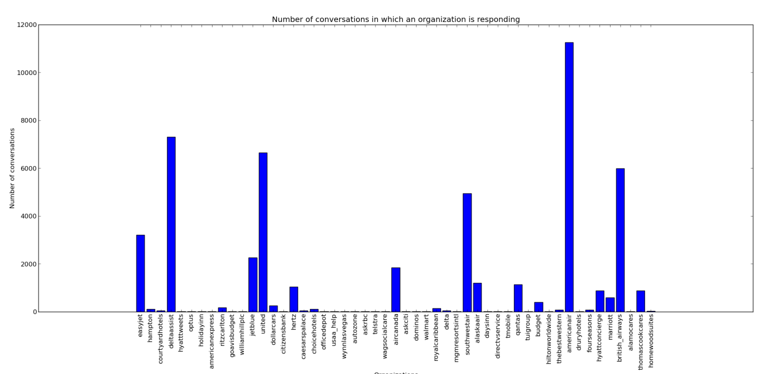


Figure 5.12: Category convo count

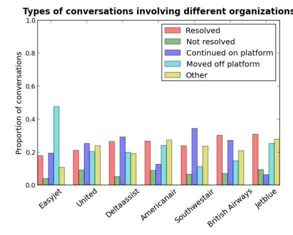


Figure 5.13: category count proportion orgs

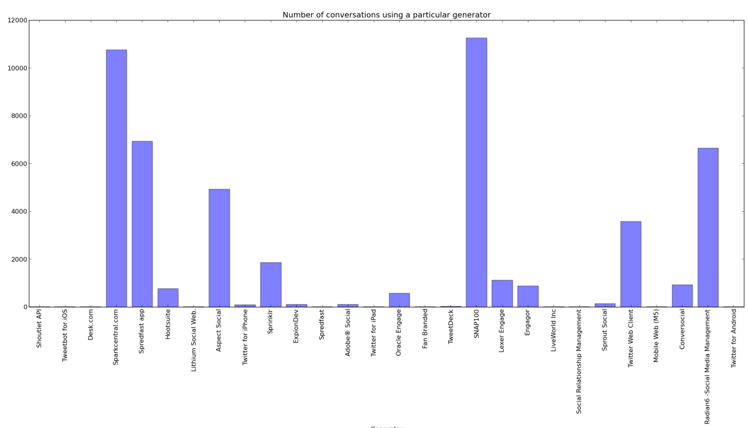


Figure 5.14: category generator orgs used convos

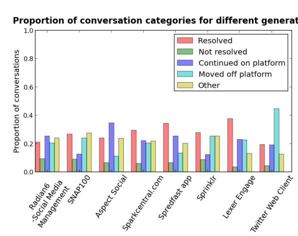


Figure 5.15: category conversation generators 1000

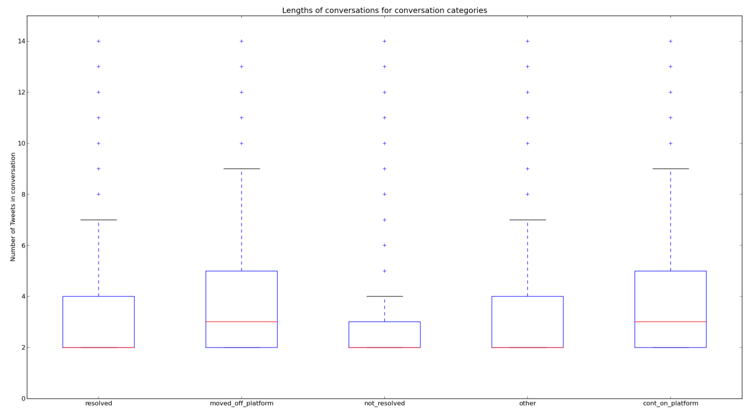


Figure 5.16

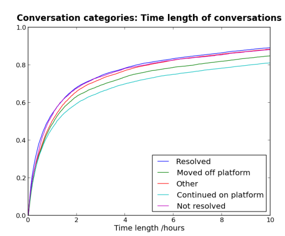


Figure 5.17

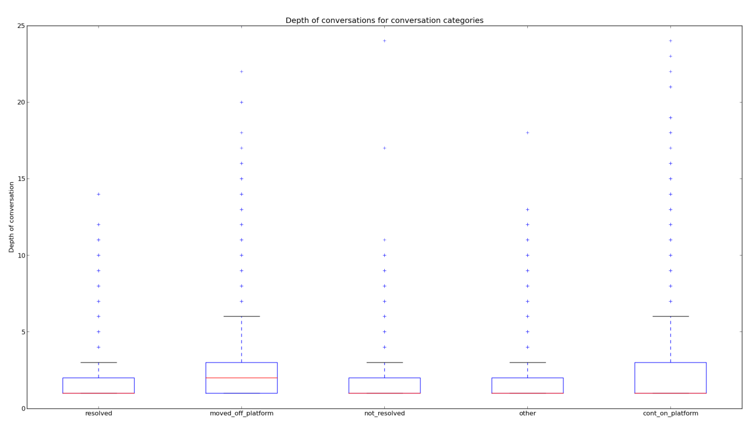


Figure 5.18

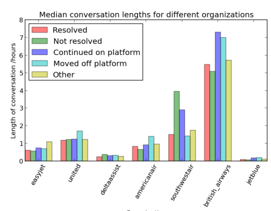

Median conversation lengths for different organizations
Resolved
Not resolved
Continued on platform
Moved off platform
Other
Length of conversation /hours
0
1
2
3
4
5
6
7
8
easyjet
united
deltaassist
americanair
southwestair
british_airways
jetblue
Organization

Figure 5.19

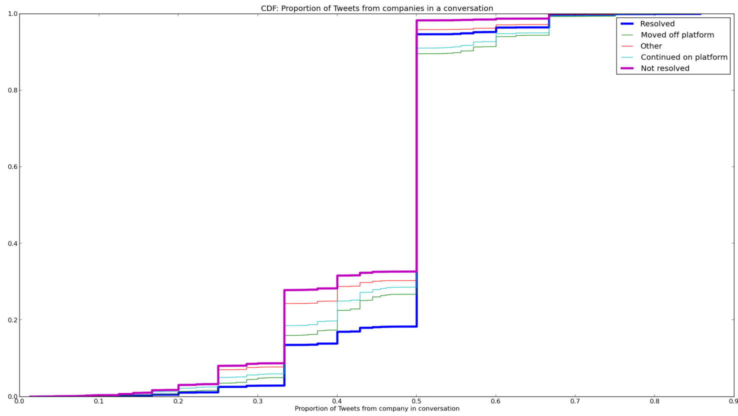


Figure 5.20

Finally, Fig. 5.21 shows the 'hours to first response'— the time between a user's Tweet and the first Tweet by an organization in the conversation. The Figure shows that there is very little variation between the conversation categories, and thus any sort of preconception that a quick response by a brand leads to a resolution, is misplaced. However, it could be that automated responses by organizations, which tend to be very quick, are shifting the hours to first response to smaller values, particularly for the 'Not resolved' category. This could be an interesting line of further inquiry.

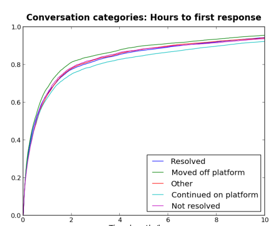


Figure 5.21

5.4.3 Travel and Telecomm: Common Generators

Since organizations generally use only one generator (the 'primary generator') for the majority of conversations, and the 'primary generator' tends to be different for different organizations, it can be challenging to separate how generators and organizations affect conversations. Fig. 5.22 does make some attempt at this, however, comparing the proportion of resolved conversations between the Travel and Telecomm industries, for generators common to both industries. (Note that a cut-off of 100 conversations for each generator in each industry, was enforced.) The dashed lines for each industry show the total proportion of resolved conversations for the industry, to which the heights of each bin should be compared.

The Figure suggests that the generator used does affect the outcome of the conversation; for 6 out of 8 of the generators the average resolved proportions are either both above, or both below, the proportion for each industry, indicating that for both industries the resolved proportion is always better than, or worse than the total (respectively). Taking the average over the industries, the generator with the highest proportion of resolved conversations is Sprout Social, followed by Sparkcentral.com.

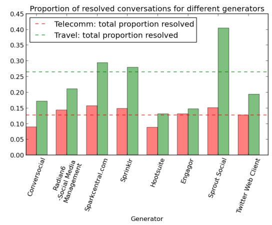


Figure 5.22

5.5 Conclusion

Developing a process of categorizing conversations and studying the results led to significant insights into conversations. We found immediately that before we could even look at how conversations are resolved, it was important to know which conversations needed resolving. This led to intensive research on term and keyword frequencies of first customer Tweets to determine which keywords and conditions signaled that a user was looking for customer service from a brand. With customer service related Tweets identified, we could then use a similar method to obtain keywords and conditions to begin categorizing our conversations. Finally, with our conversations categorized we were able to perform descriptive statistics on each set within each industry to determine if there were any obvious factors that led to a conversation being resolved.

Although conversations of different categories were not found to vary significantly in length, depth or size, it was found that *resolved* conversations tended to have a higher proportion of Tweets from organizations than *not resolved* conversations. Furthermore, comparisons between the Travel and Telecomm industries suggested that the Travel industry is most successful at resolving conversations. This industry was analyzed by organization and generator, which identified Easyjet as the most successful organization at dealing with customer service issues on Twitter, and Lexer Engage as the best generator.

We also found that although companies and generators tend to have a greater number of either moving the conversation into a direct message or moving it off the platform, it does not appear that they choose one over the other. Which suggests, that companies may have different methods for different types of conversations.

Furthermore, we see that Telecomm typically sends their conversations off the public platform more often then Travel, which may indicate why Travel has twice the proportion of resolved conversations as Telecomm.

Finally we see interesting results as resolved conversations tend to have more Tweets in the conversation than other categories, and that resolved conversations tend to vary by industry, company, and generator.

Although this process of categorization worked well for Travel and Telecommunication, further analysis on other industries will require strict attention to the types of conversations being conducted. For example, in the Restaurants Industry, conversations tend to be extremely 'chatty', so much so, that initial attempts to identify the customer service natured conversations failed due to too many crossover keywords between Tweets that actually had a complaint and Tweets that just wanted to shout out to a company. Furthermore, the responses from these companies are frequent and short, with a company often responding to Tweets without need of resolution just to say hi. Travel and Telecomm worked well for these methods since many of the conversations being had were serious and customer service related. When exploring new industries perhaps, more sophisticated techniques in keyword extraction(**??** REFERENCE FUTURE DIRECTION KEYWORD EXTRACTION) will be needed.

Chapter 6

Future Directions

6.1 Better NLP

NLP is a highly interdisciplinary field concerned with the interactions between computers and human (natural) languages. There are many softwares and toolboxes online for NLP and more specific tasks, including stemming (the process of reducing inflected (or sometimes derived) words to their word stem), polarity testing (detecting positive/negative/neutral attitude). Here we focus on sentiment analysis and keyword extraction, the two main parts that we have relied on in our research.

6.2 Sentiment Analysis

We primarily perform sentiment analysis to prepare for regression (Section 6.3). We also believe that sentiment analysis is important for reflecting on the nature of conversations. For sentiment analysis, we use the rather rudimentary sentiment analysis package from Pattern (NEED TO REFERENCE). This package provides a function called *sentiment* (NEED TO REFERENCE). Inside Pattern, there is a lexicon of sentiment-related adjectives that are frequently present in product reviews. Each adjective is assigned a sentiment score between -1 and 1 for the positivity of the word, and another score between 0 and 1 for the subjectivity. The function *sentiment* identifies all of the sentiment-related adjectives in the texts provided. Then it averages over all the sentiment scores and subjectivity scores to obtain a positivity measure between -1 and 1. Compared to other NLP tools, Pattern and its sentiment analysis package are the simplest to use. However, there are also many limitations:

- NLP is still a fast developing field, thus the accuracy and precision people obtain for subfields such as polarity testing and sentiment analysis are still not satisfactory. Therefore, we need to be very careful about the scores obtained from Pattern.
- Due to the breadth of the topics on Twitter and the character limit of each Tweet, the content of Tweets can be very different from normal texts. Tweets include many more acronyms, hashtags and URLs, making NLP for Tweets especially hard.
- So far we have only looked at the conversations concerning one company, but from our observation of the data set, about 2% of all conversations involve more than one companies. Performing sentiment analysis in these scenarios will be trickier, since

we need to figure out which attitude is directed towards which company. These 2% of conversations will be particularly interesting for analyzing how certain companies win over the users in multi-company-involved conversations.

Currently we perform quality control of our categorization and sentiment analysis mainly by manually reading through sampled Tweets and conversations and see if the categorization is appropriate. But given the three major limitations listed above, in the future we may need some sentiment analysis tools that are specialized in handling Tweets in order to improve our current study. There are indeed many other well-known packages for sentiment analysis that are frequently used, including GATE (NEED TO CITE), RapidMiner (NEED TO CITE), and the NLTK package for Python (NEED TO CITE), which may be adopted for future developments.

It is also important to note that sentiment analysis is nowhere near perfect in terms of theoretical development. Despite the drawbacks listed above, the size of our data set may help compensate for the possible inaccuracy. By setting thresholds for the positivity score to detect polarity, we can better interpret our results. Moreover, even though Pattern cannot guarantee complete accuracy, humans are not capable of being 100% correct about the attitudes of a sentence without context, not to mention giving a scaled score on positivity and subjectivity. The accuracy needs to be high enough to capture the sentiment of the majority of the Tweets, so that we can spot meaningful relationships among variables with statistical significance.

6.2.1 Keyword Extraction

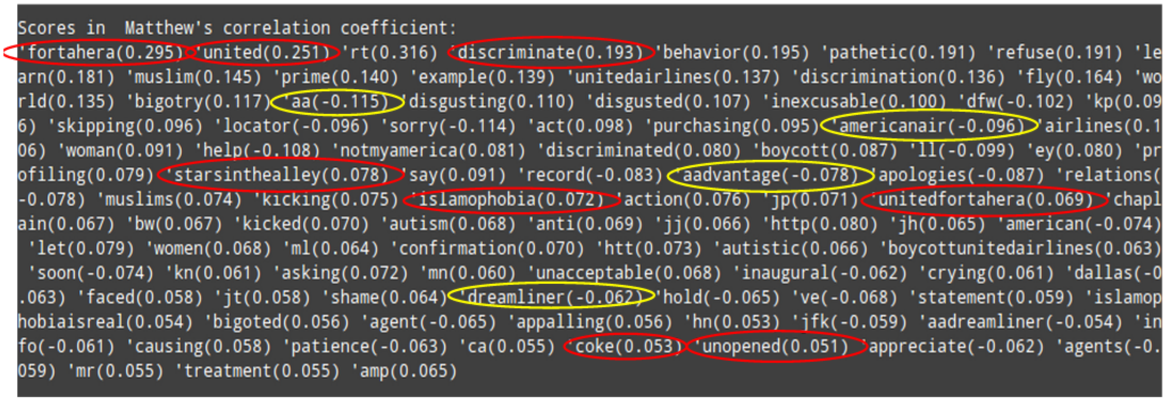
Currently we employ the most straightforward measure to extract keywords by using term frequencies of all the phrases. However, there are plenty of feature selection methods that rely on more sophisticated metrics and statistics from information theory.

One main alternative that we have experimented with is Super CWC/LCC, a consistency-based feature selection algorithms (NEED TO REFERENCE). These tools extract comparative keywords from multiple sets of text. Here we present a simple example for comparing the keywords of United Airline (UA) related Tweets and American Airline (AA) related Tweets.

First we extracted all the Tweets related to UA and AA. We then generated the sparse matrices of term frequencies for UA related and AA related Tweets respectively using the procedures described in Section 5.1. Next we inputted them into Super CWC/LCC algorithm which computed, based on many different metrics (such as symmetric uncertainty, mutual information, Bayesian risk and Matthews correlation coefficient), the feature words that are much more likely to appear in Tweets related to UA than those related to AA, and vice versa. Fig. 6.1 shows the sample output of the algorithm in the terminal.

From the output (Fig. 6.1) we have created a list of keywords related to UA and AA. Each keyword extracted also has a signed number attached to it. This number is called Matthew's Correlation Coefficient (MCC), which is a measure of the quality of binary classification. All the keywords with positive MCC are keywords for UA (some of them are circled in red), and all the keywords with negative MCC are keywords for AA (some of them are circled in yellow).

Recall that during the month of May 2015, (the time frame that our data set), a major trending issue occurred when a muslim woman, Tahera Ahmad, was publicly discriminated against on a UA flight. Fig. 5.3 shows an excerpt from Facebook that details what happened. As we can see from the previous output (Fig. 6.1), many of the words circled in red are, in



```
Scores in  Matthew's correlation coefficient:
'fortahera(0.295) 'united(0.251) 'rt(0.316) 'discriminate(0.193) 'behavior(0.195) 'pathetic(0.191) 'refuse(0.191) 'le
arn(0.181) 'muslim(0.145) 'prime(0.140) 'example(0.139) 'unitedairlines(0.137) 'discrimination(0.136) 'fly(0.164) 'wo
rld(0.135) 'bigotry(0.117) 'aa(-0.115) 'disgusting(0.110) 'disgusted(0.107) 'inexcusable(0.100) 'dfw(-0.102) 'kp(0.09
6) 'skipping(0.096) 'locator(-0.096) 'sorry(-0.114) 'act(0.098) 'purchasing(0.095) 'americanair(-0.096) 'airlines(0.1
06) 'woman(0.091) 'help(-0.108) 'notmyamerica(0.081) 'discriminated(0.080) 'boycott(0.087) 'll(-0.099) 'ey(0.080) 'pr
ofiling(0.079) 'starsinthealley(0.078) 'say(0.091) 'record(-0.083) 'aadvantage(-0.078) 'apologies(-0.087) 'relations(
-0.078) 'muslims(0.074) 'kicking(0.075) 'islamophobia(0.072) 'action(0.076) 'jp(0.071) 'unitedfortahera(0.069) 'chapl
ain(0.067) 'bw(0.067) 'kicked(0.070) 'autism(0.068) 'anti(0.069) 'jj(0.066) 'http(0.080) 'jh(0.065) 'american(-0.074)
 'let(0.079) 'women(0.068) 'ml(0.064) 'confirmation(0.070) 'htt(0.073) 'autistic(0.066) 'boycottunitedairlines(0.063)
 'soon(-0.074) 'kn(0.061) 'asking(0.072) 'mn(0.060) 'unacceptable(0.068) 'inaugural(-0.062) 'crying(0.061) 'dallas(-0
.063) 'faced(0.058) 'jt(0.058) 'shame(0.064) 'dreamliner(-0.062) 'hold(-0.065) 've(-0.068) 'statement(0.059) 'islamop
hobiaisreal(0.054) 'bigoted(0.056) 'agent(-0.065) 'appalling(0.056) 'hn(0.053) 'jfk(-0.059) 'aadreamliner(-0.054) 'in
fo(-0.061) 'causing(0.058) 'patience(-0.063) 'ca(0.055) 'coke(0.053) 'unopened(0.051) 'appreciate(-0.062) 'agents(-0.
059) 'mr(0.055) 'treatment(0.055) 'amp(0.065)
```

Figure 6.1: Command Line Output of Super CWC/LCC

fact, words and hashtags related to the Tahera event described in the excerpt (**??**). Thus this can serve as an example of Super CWC/LCC being able to capture the trending issues related to certain companies.

As Super CWC/LCC shows promising capability of extracting trending issues and key contents, we can certainly imagine performing similar keyword extraction process comparing more companies within a certain industry (instead of just two in this example), and even comparing different industries.

6.3 Regression Analysis

With the customer service related conversations categorized (as described in Chapter Chapter 5), we performed some preliminary logistic regression analysis to explore the factors that may be of importance to whether a certain conversation will be resolved or unresolved.

6.3.1 Motivation

We have seen some comparisons of statistics across different categories for Travel and Telecommunication Industries. Apart from visualizations, we have looked at whether certain variables are statistically significantly different for the resolved and unresolved categories using t-test.

For example, we looked at the sentiment scores for the last customer Tweets of the resolved and not resolved categories for the Travel Industry and saw a statistically significant difference with the average sentiment score to be 0.11 for resolved and -0.02 for not resolved ($p < 7.58e\text{-}32$). This was be expected, since we defined our resolved and not resolved categories largely based on the keywords in the last first customer Tweets. However, it is still noteworthy to confirm this expectation, since we did not explicitly use sentiment to decide the categories.

Fig. 6.2 is a cumulative density function for the sentiment scores of last customer Tweets for resolved and not resolved categories in the Travel Industry . It is clear from Fig. 6.2 that the percentage of not resolved last customer Tweets with negative sentiment scores is higher than the resolved ones, and the percentage of resolved last customer Tweets with positive sentiment scores is higher than the not resolved ones.

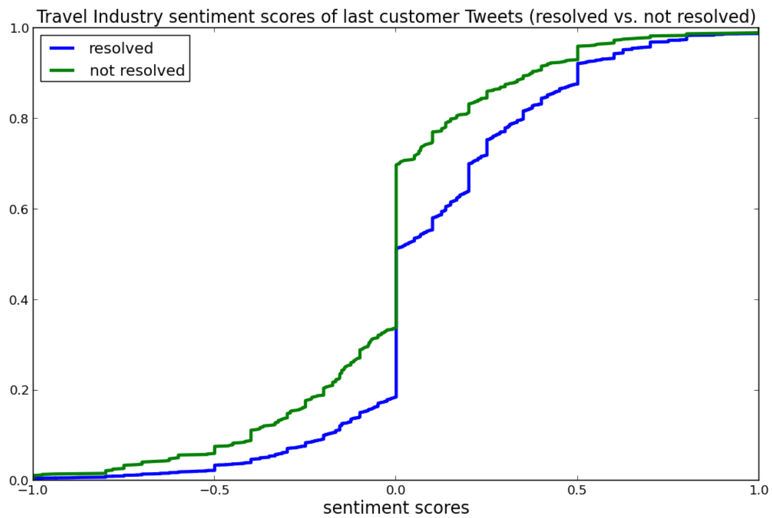


Figure 6.2: Cumulative density of sentiment (resolved vs. not resolved)

Within Travel Industry we performed similar analysis for many other variables, including hours to first response and size of conversations for resolved and not resolved conversations. Both variables are statistically significantly different. The average hours of first response is 3.56 for resolved and 2.15 for not resolved conversations ($p < 8.00e\text{-}12$). See the cumulative density in (Fig. 6.3). For size of conversation, the average is 3.70 for resolved and 4.47 for not resolved ($p < 2.36e\text{-}18$).

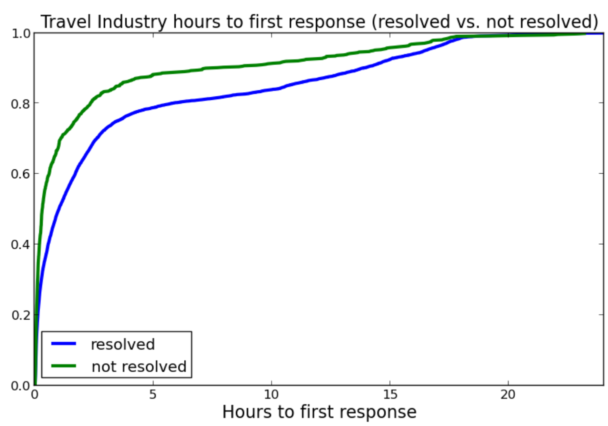


Figure 6.3: Cumulative density of hours to first response (resolved vs. not resolved)

Those significant differences motivated binary classification of the resolved and not resolved categories through including those variables in a logistic regression model, since ultimately we want to help companies improve their customer service on Twitter platform. We included primarily the variables that the companies have control over, such as the sentiment, number of words, number of characters of its own Tweets. In addition, we included hours to first response, and generators. For the logistic regression we only consider the conversations with hours to first response < 24 hours. We also only included all the resolved and not resolved conversations for both Travel and Telecommunication Industries.

Table 6.1: Summary of Logistic Regression

	Coefficients	p-values
Intercept	2.32	$< 2e-16$
sentiment of first brand Tweet	0.64	5.33e-16
number of words of first brand Tweet	0.05	3.11e-05
number of characters of first brand Tweet	-0.0089	6.63e-04
hours to first response (< 24)	0.066	$<2e-16$
size of conversations	-0.14	$<2e-16$
generator	0.42	1.58e-09

In the following example, we regressed the resolved and not resolved categories with respect to the sentiment, number of words, number of characters of first brand Tweet, hours to first response, size of conversations, plus an indicator variable called generator, which is 1 if the generator used by the brand is Sparkcentral.com, and 0 otherwise. Here is a summary of the result:

Since we largely determined the categories based on last brand Tweets, we regressed with respect to the properties of first brand Tweets instead. However, note that many of the resolved conversations are of length 2, since if a conversation is of length 2, the first customer Tweet is a question and the brand responds it categorized as resolved. Among the 17,612 observations, over 7,000 conversations are of length 2, which may explain why the coefficient for the variable size of conversations is negative. A significant result is that the indicator variable, generator, is also significant because we do not define resolution based on which generator a brand primarily uses.

Another somewhat counter-intuitive observation is that the coefficient for hours to first response is positive, which means that the model predicts that the likelihood of a conversation being resolved would increase if the hours to first response also increases. This coincides with what we see from the visualizations of different categories, as not resolved has the shortest response time. Our current hypothesis is that many quick responses are automatically generated, and the customers may not be very satisfied with those responses. Whether this is the case is subject to future studies.

We have also done some initial analysis of how well our model is at capturing the data and making predictions. However, a big problem is that the number of conversations that are resolved is almost ten times that of not resolved in our observations. Better testing techniques, such as multiplying the set of not resolved conversations or assigning different weights to the two groups, shall be employed in the future to calculate the fit of the model.

6.4 Breaking up Time Frames

An important thing that we do not explore in this study is how the number and nature of the conversations changes at different times during the day. For example, in the Travel Industry we would expect many more conversations about flight delays during rush hours than the evening. Looking at the keywords, resulting statistics and categories of conversations will be a nice refinement to our current study.

Table 6.2: Industry Mapping: McKinsey Industries to Twitter Industries

McKinsey Industry	McKinsey Industry
Travel & Hospitality	Travel Industry
Finance & Insurance	Financial Services Industry
Household Products	Home & Garden Industry
Technology Hardware	Technology Industry
Consumer Packaged Goods	Consumer/Disposable Goods Industry
Utilities	Energy Services & Utilities Industry
Automotive and Aftermarket	Auto Industry
Telecomm & Media	Telecommunication Industry
Healthcare	Healthcare Industry
Retail & Retail Industry	
Transportation & Logistics	Travel Industry
Diversified Business Services	Ambiguous/Other
Wholesale and Distribution	Ambiguous/Other

6.5 Extension to Other Industries

We have presented a framework for carrying out conversation categorization for Travel and Telecommunication Industries. However, this framework is applicable to other industries in general as well. So a natural next step is to extend this study to other industries present on Twitter.