Aalto University
School of Business

Twitter data mining: social media as a rising customer service channel

Bachelor's Thesis
Sirong Huang
10.05.2016
Business Technology

Table of Content

List of Tables

List of Figures

1. Introduction

Over the years, Internet has drastically changed the way consumers interact with companies. Social media, especially, has evolved to become more than just marketing platforms, but also where consumers seek conversation and demand service from companies. According to a survey conducted by McGraw Hill in 2013, among 23,200 US respondents 67% of them have used a company's social media site for servicing, compared with only 33% for marketing (J.D. Power and Associates, 2013). 33% of users even prefer to contact brands using social media rather than the telephone (Nielsen, 2012). With rising demand in social customer service, however, the phenomenon remains greatly unexplored by the academic world.

This study is an attempt to examine the current status of social media customer service through quantitative data analysis of over 30,000 tweets collected in a 15 days time period, monitoring all conversations on Tweeter between five leading consumer electronics brands in USA and their customers.

The result indicates that leading consumer electronics brands are providing high quality customer service on social media with an average response time of 30 minutes. The study also found that Twitter customers are more active and influential on social media than average. Customers expect fast problem resolution by the first reply from customer support; otherwise, they grow dissatisfied as the interactions prolong. However, solving difficult problems with longer sessions does bring the satisfaction level back up. From text analysis, most problematic products and services that demand extensive customer service are identified, as well as frequently asked questions and desired features. Overall, this research provides an initial framework to study the rising phenomenon of social customer service, and from the findings, valuable information about social customer service process is revealed.

The data was collected through Twitter API using Python programming language. The data analysis consists of 3 tasks: summary statistics, sentiment analysis and exploratory text mining. Natural language processing and supervised statistical machine learning methods are used to perform text analysis tasks.

The remaining chapters are organized in such manner: chapter 2 reviews literature on social media and data mining. Then the data collection and data analysis methods are explained in chapter 3. In chapter 4, findings of the study are presented, followed by conclusions in Chapter 5, and then the paper ends with research limitations and future research topics in Chapter 6 and 7 respectively.

2. Literature review

Social media provides large scale, real time, first-hand data into various areas. Twitter especially, with an average of 500 million tweets being sent per day, provides a valuable information source to companies and researchers alike. Twitter sentiment analysis and opinion mining have great empirical value to companies wishing to monitor customer's opinions to their products and services (Bifet, & Frank, 2010), and also have applications in various fields in social science and economics. For instance, twitter sentiment was used to predict stock market performance and scored an impressive accuracy of 87.6% in predicting the daily up and downs of DJIA (Bollen, Mao, & Zeng, 2011). German federal election result was also predicted by monitoring political sentiment on Twitter (Tumasjan, Sprenger, Sandner, & Welpe, 2010).

Customer service quality and customer satisfaction are widely recognized as key influence factors in purchase behavior and customer relationship management (Taylor & Baker, 1994). There are varying performance measurement models both from the academic world and from industry practice, due to the nature of service quality being elusive and subjective to be measured with a unified standard. The service quality measurement framework used in this paper is a highly influential model called SERVQUAL, developed in 1988 (Parasuraman, Zeithaml, & Berry, 1988). SERVQUAL evaluates service quality by comparing the gap between customer's expectation on service quality and that of the actual service received. Five determinants that influence the size of the gap are: reliability, assurance, tangibles, empathy and responsiveness (RATER). For the purpose of this study, we simplify SERVQUAL model to only measure customer's satisfaction level after the service, without considering customer's expectations before the service. In summary, we focus on examining the responsiveness, reliability and empathy aspects of twitter customer service, and then evaluate customer's satisfaction level for the service.

3. Research methods

3.1 Data collection

The twitter accounts monitored are @AppleSupport, @SamsungSupport, @LGUSA-Mobile, @Moto_Support (Motorola) and @HTCUSA. Together, these 5 brands consist of 88.6% of US mobile device market share (ComScore, 2015), and provide a rich dataset for analyzing the current practices of customer service on Twitter especially for mobile device industry

The data is collected through Twitter API, which is an application programming interface Twitter provides for developers to interact with Twitter data to build software applications. The data streaming script is written in Python programming language, using a filter that returns all the tweets sent to and from the 5 target brand's Twitter support account. One thing to note is that during the data collection process, a few server downtime and computer crashes happened, which may cause incomplete conversation records. This could potentially affect the accuracy. However since the downtime is very short compared to the overall data volume, this would not significantly affect the data quality.

The data collection starts from 20 April 2016, and runs continuously until 4 May 2016, collecting a total of 32,312 Tweets.

3.2 Data analysis

The data analysis consists of three tasks: summary statistics, sentiment analysis and text mining. Summary statistics provides an overlook of the dataset and measures key performance indicators such as response time and response rate. However, in order to reveal deeper insight on the data, sentiment analysis and text mining are performed to evaluate customer's sentiment during and after initial contact with the company, and to discover common topics, behavior patterns, and consumer intentions from the text.

Before sentiment analysis and text mining can be performed, the twitter text requires preprocessing. The unique linguistic characteristics of Tweets pose many challenges on natural language processing (NLP). For instance, '@', '#', 'RT', emoji, links, abbreviations, internet slang and misspellings all complicate the text processing process. In this study, NLTK library is used to tokenize, stem and filter the tweets and the final product is a neat feature set extracted form each tweet.

Sentiment analysis is a series of natural language processing, text analysis and statistical machine learning tasks performed to identify and extract subjective information in text (Pang & Lee, 2008). For the purpose of this study, we use a supervised machine learning algorithm called Naive Bayes classification to label the text as either positive or negative. The Naive Bayes classifier selects the most likely classification V_{nb} from a bag of words feature set computed using Bayes' rules, with the assumption that the features are independent from each other. Below is the formula from the book (Lewis, 1998):

The Bayes Naive classifier selects the most likely classification V_{nb} given the attribute values $a_1, a_2, \dots a_n$. This results in:

$$V_{nb} = \text{argmax}_{v_j \in V} P(v_j) \prod P(a_i|v_j) \tag{1}$$

We generally estimate $P(a_i|v_j)$ using m-estimates:

$$P(a_i|v_j) = \frac{n_c + mp}{n + m} \tag{2}$$

where:

n =	the number of training examples for which $v = v_j$
n_c =	number of examples for which $v = v_j$ and $a = a_i$
p =	a priori estimate for $P(a_i|v_j)$
m =	the equivalent sample size

The Naive Bayes classification method has wide application in machine learning and data mining for its simplicity, scalability and effectiveness (Bifet & Frank 2010). Despite its unrealistic assumption, the classifier is remarkably successful in practice, often competing with much more sophisticated techniques (Rish, 2001).

In supervised machine learning, it is common practice to use manually labeled corpus of text to train the classifier. However, with rapidly evolving text features in Twitter, it is very difficult to hand label a large enough dataset to train the classifier effectively (Go, Bhayani & Huang, 2009). The training corpus used in this study is a 10,000 tweets dataset collected by three Stanford researchers using a novel method called Emoticon (Go, 2009). Rather than manually label a small number of tweets, they fetched a large number of tweets with clear indication of emotions such as the emoticon ‘ :) ’ or ‘ :(’ as training data. The testing data is another corpus called STS-Gold containing 2,034 tweets hand-labeled by 3 PhD students (Saif, Fernandez, & Alani, 2013). Figure 1 shows the classifier’s test result and a list of most informative words. The ration on the right indicates the number of times the word appeared in negative (0) or positive (4) tweets. With 82% accuracy, the classifier we trained can be used confidently to determine the sentiment of tweets in our dataset.

Figure 1 Naive Bayes classifier training result

```
train on 10000 instances, test on 2034 instances
accuracy: 0.82005899705
Most Informative Features
                     sad = True              0 : 4      =     28.6 : 1.0
                horrible = True              0 : 4      =     18.1 : 1.0
                congrats = True              4 : 0      =     17.2 : 1.0
                    crap = True              0 : 4      =     16.1 : 1.0
           unfortunately = True              0 : 4      =     14.0 : 1.0
                   hurts = True              0 : 4      =     13.8 : 1.0
           #followfriday = True              4 : 0      =     13.3 : 1.0
               wonderful = True              4 : 0      =     10.3 : 1.0
                headache = True              0 : 4      =     10.1 : 1.0
                    argh = True              0 : 4      =      9.9 : 1.0
```

4. Findings

4.1 Overview

Table 1 provides an overview of the dataset. Among a total of 32,312 tweets collected, **27,590** tweets are selected as the main dataset for analysis, because those are “original tweets” written by users in an attempt to initiate a conversation with the company. 4,235 retweets are excluded from analysis since they are simply sharing activities that do not represent customer service interactions. Users that send more than 15 tweets to brands within 15 days are suspected of spamming. After a simple text filtering, 9 spammers with a volume of 487 spam tweets are detected and excluded from the dataset in further analysis.

An interaction is a series of tweets extracted by tracking a unique customer’s twitter ID, and pull together all records of the conversation between that unique customer and the company. Among 9,355 unique customers who contacted the company, 6,171 received response and thereby started a conversation with the company, whereas 3,184 of them received no response in our record.

Table 1 Data overview

	Total number of tweets			Total number of interactions	
	Original tweets	Retweets	Spam	Conversations	No response
	27,590	4,235	487	6,171	3,184
Total	32,312			9,355	

4.2 Length of interaction and response rate

The interactions are divided into three groups based on number of tweets included: messages with no response contains all tweets sent by customers to the company but received no reply; short conversations consist of conversations with 2 to 4 total messages between a unique customer and a company; medium length conversations record conversations with 5 to 10 overall messages; long conversations records every conversation with more than 10 overall tweets.

Table 2 Tweets grouped by length of interaction

	No response messages	**Short conversations**	**Medium length conversations**
Number of tweets in each interaction	1	2-4	<5
Number of interactions	3,184	5,155	1,016
Percentage of total interactions	34%	55%	11%

As can be seen from Table 2, there is a fairly high non-response rate of 34%. There are several possible explanations as to why those messages were not responded by the company. From the technical side, there has been a few server downtime in the data collection process, causing possible incomplete records of the conversation. However, by analyzing the tweets' text, evidence can be found that most tweets that received no reply from twitter customer support can be categorized into 3 types: tough questions, insults or complains, or one-sided expression of opinions or feelings. Examples from each type can be seen in Figure 2.

Therefore, a hypothesis is made that the text can be classified into the 3 types mentioned above. Starting from the hypothesis, I used lexicon-based sentiment analysis method to classify those tweets into 3 categories (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). A sentiment lexicon library from Hu & Liu (Hu & Liu, 2004) together with an additional hand-picked word list specific for this task, such as '?', 'why', 'when' are used to classify the above mentioned three types of no-response messages. Since we only extract tweets that satisfy the defining feature words of each category, tweets that do not contain those features will not be categorized. 89% of the tweets can be classified in the above categories.

In conclusion, for a large portion of those no response tweets, the customer's intention is not to demand any service but to express feeling or opinion, complain or attack the brand, or to ask a tough question that is difficult for customer support employees to answer. Although it is no excuse to justify not replying customer’s messages, we can conclude that the actual service response rate is around 90%.

Figure 2 Tweets with no response from customer support

@AppleSupport When is the iPhone 7 going to come out?
@AppleSupport why do you never reply to questions?

@AppleSupport I DONT WANT YOUR UPDATE!! Quit sending me it!!!
@AppleSupport fuck you and your shitty products
@SamsungSupport you can't replace the drain pump without those rods. That's just wrong. Thanks for checking on this. #lostcustomer

@AppleSupport thank you!! (2%)
@HTCUSA I want! Sucks I am on AT&T & depend on the ATT Next to purchase...#Losinghope #loyalcustomer #Newphone #Searching

4.3 Response time

The response time is an important performance indicator of customer service quality. A frustrated customer expects to receive response as soon as possible after the initial contact with customer service. As can be seen from Figure 3 and 4, 75% of customers receive response within 40 minutes, with an average of 32 minutes waiting time. 98% of tweets are replied within 24 hours. The result indicates that mobile device market leaders provide fast and smooth customer services on Twitter.

Figure 3 Response time density distribution

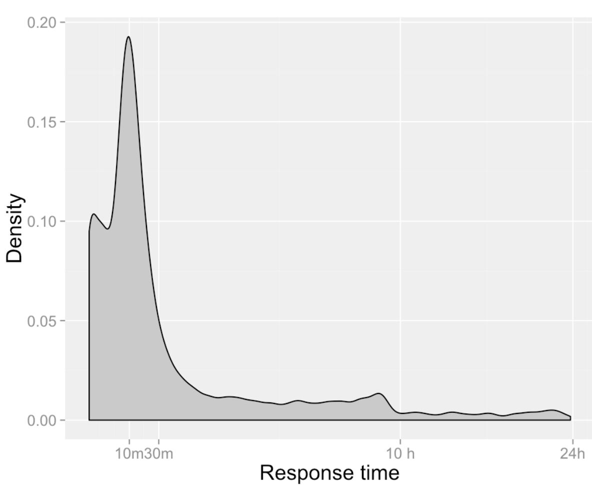
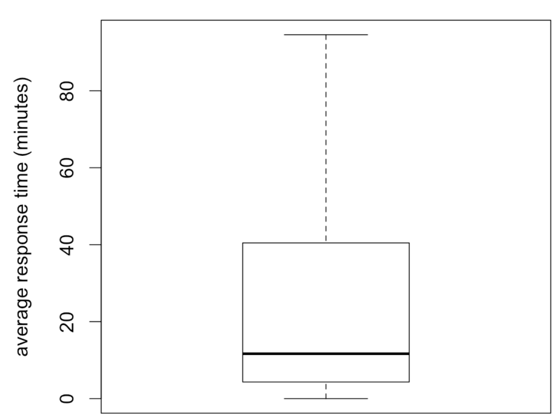


Figure 4 Average response time



4.4 Customer analysis

No doubt that influential twitter users with large number of followers play an important role in word-of-mouth diffusion on social media (Bakshy, Hofman, Mason, & Watts, 2011). Therefore, it is essential to evaluate customer's social presence and influence. Our data shows that customers who use Twitter customer service have an average of 1,222 followers, compared to 208 for average twitter users. Also more than 50% of those customers have follower base larger than average. We can then draw the conclusion that users of Twitter customer service are generally more active and influential on social media.

Figure 5 Customer's Twitter follower count

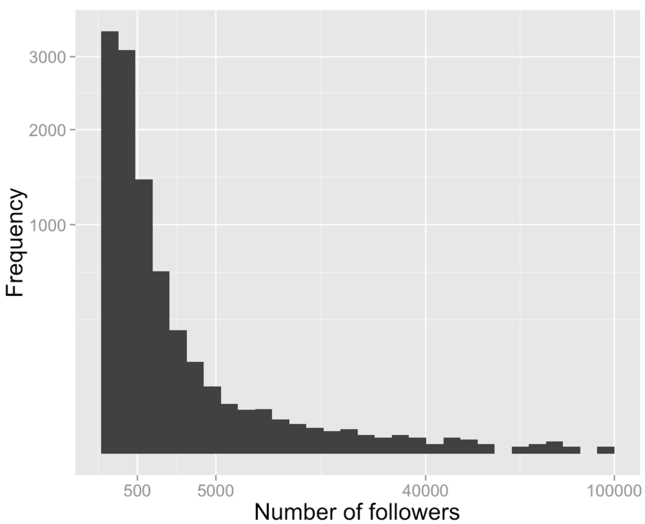


Table 3 Most influential Twitter customers

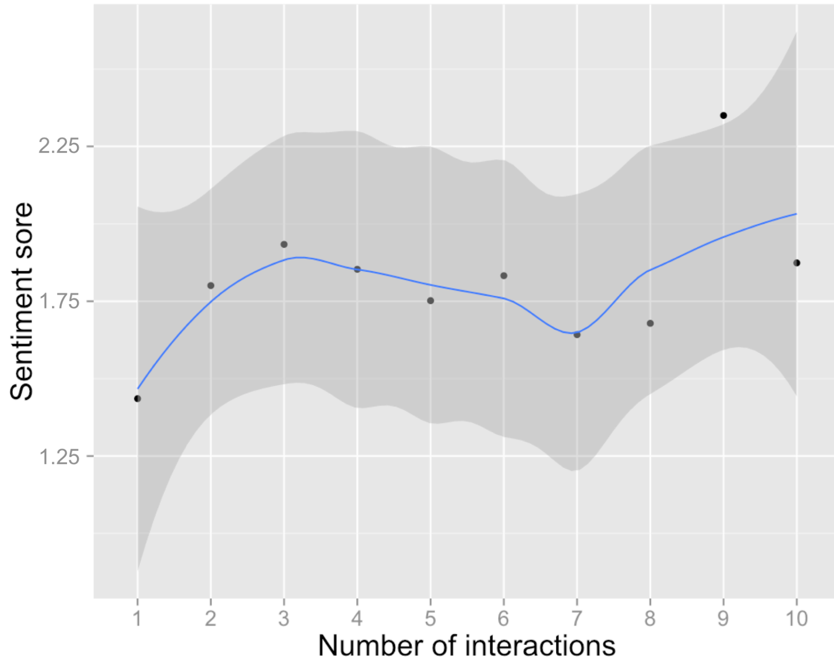
Username	*Number of followers*	*Number of tweets*
DrJimmyStar	1,257,615	168,069
MattPrior13	241,876	7,977
TheEricYoung	206,098	27,598
Jon4Lakers	158,030	42,430
Melofication	150,217	19,223

4.5 Sentiment analysis

Applying the trained Naïve Bayes Classifier on collected tweets, we are able to determine the sentiment score of all tweets on a scale of 0 to 4 (0=negative, 2= neutral, 4=positive). Figure 6 shows sentiments of tweets sent by customer, for each number of interactions. Due to the nature of customer service requests being a result of negative experiences, it is expected that the sentiment scores of customers' tweets tend to be negative. Similarly it is unsurprising that the average sentiment of brand's customer service tweets is a very positive score of 3.1 indicating polite and positive word of choice used by customer support staff.

By using Pearson's product-moment correlation method, the link between number of interactions and sentiment score has a correlation of 0.61, indicating there is likely a relationship between these two variables. However by taking a closer look and combining text analysis results, we conclude that the sentiments should be examined by each interaction length.

Figure 6 Customer tweets sentiment score by number of interactions



From Figure 6 we can tell messages that received no response are the most negative ones. From semantic analysis, we observed that apart from common vocabularies shared across all lengths of interactions, strongly negative and even insulting languages are frequently identified in those messages. For example, ('suck' 'sucks', 289), ('shit', 174), ('never' 'ever', 146), ('fuck' 'fucking', 98), ('worst', 52). The offensive nature of those texts can be one of the reasons they received no reply. From length 2 interactions onward, however, no offensive words appeared more than 30 times in every category, as oppose to the frequent appearance of those in non-response tweets. Tweets with more than 2 interactions also tend to focus mainly on the problem, packed with descriptive nouns of device, service, product and the existing problems. Polite words, such as “thank you”, “please” increase in frequency as the conversation continues.

There is a positive turn of sentiment at length 3 where the customer replies to the customer service a second time after receiving replies form the company. Considering that 55% of interactions have a length of 2-4 tweets, this could potentially indicate the problem being solved. However, as the interaction gets longer, the sentiment score slowly slide downwards, with negative indicators like “still”, “:(” rising in frequency, until rising up again at length 7 mark.

From sentiment analysis and text mining, we confirm the previous hypothesis that some tweets receive no response not because of customer support failure, but because of its offensive nature. The sentiment graph also shows that if the customer service can solve the problem in one interaction, it tend to improve customer satisfaction significantly. However, if it takes more interactions to solve the problem, customer satisfaction falls gradually, to a point where the customer can be even more dissatisfied than before receiving the service. Nonetheless, if the case is complicated enough to take more than 7 interactions, customer’s sentiment became more positive again as the customer service devotes more effort into solving the problem.

4.6 Common topics and most used keywords

From text mining the tweets, a list of most frequent product related noun is extracted. These are the products and services that with most customer service request, indicating problematic, faulty, confusing or unsatisfying experience with those products. This feature can be used to monitor product and service quality, customer feedback, to name a few.

Table 4 Most problematic products and services

Frequent problems	*Frequency*
Update	600
iTunes	482
Screen	426
Battery	401
iCloud	312
Emoji	271
Wifi	180
Charger	179
Security	110

5. Conclusion

Social media have revolutionized the interaction between consumers and companies. To study the rising phenomenon of social customer service, we benchmarked the performance of 5 leading consumer electronics brands' customer service on Twitter, by collecting Twitter data and applying text mining and machine learning algorithms for analysis.

With 32, 312 tweets analyzed, the average response time is as short as 30 minutes, and most customers receive service within 40 minutes. The result indicates that market leaders are providing highly professional and timely customer service on social media, aiming to offer seamless experience across platforms, and serve customers on their preferred platform. This confirms to the empirical evidence that social media is becoming another essential customer service platform. We also discovered the users of Twitter customer service tend to be more active and influential Twitterers than average, which further address the importance of social customer service for positive word-of-mouth effect on the internet. It is important to find out that customers expect fast problem resolution, usually from company's the first reply, otherwise customer's satisfaction level declines as the problem takes longer time to solve. However, when handling complicated cases, long service sessions do pay off with increasing customer satisfaction after a certain point. From mining the tweet texts, a list of most problematic or confusing products and services can be identified through frequency word count, which also includes most wanted features.

The study result also demonstrates the information richness of social media records. The dialogues between customers and companies can be used to monitor customer service performance, conduct market research, acquire customer opinion and feedback, monitor competitors – the list goes on.

This research is a good starting point to recognize social media as a rising customer service platform. It provides an initial framework for analyzing social customer service performance. Moreover, this study is also an attempt to apply natural language processing and machine learning into business studies, and to use web data mining as the research data source in place of traditional questionnaires, for its large scale data volume, objectivity and timeliness advantages.

6. Limitations

The study used Twitter data to examine the current status of social media customer service. Twitter and Facebook are currently the biggest social media platforms. Since this study does not include data from Facebook, it might be biased by unique characteristics that only exist in Twitter platform.

Due to time limitation, the sentiment analysis classifier is trained and tested with other tweet datasets, rather than the tweets collected in this study. This is not a optimal solution but it is simply impractical to manually classify a large enough dataset to test the data within the timeframe of this research.

During the data collection period, due to system shutdown unexpectedly, a small portion of conversations is recorded incompletely. However, considering the large size of overall data volume, the effect on the actual result should be negligible.

7. Future research

Future research can be done in mining the profile of customers contacting companies on Twitter, which reveals the customer's age, gender, location, occupation, interests, device, etc. Moreover, it would be interesting to track the social activities of customers who have used Twitter customer support, in order to monitor the customer's sentiment, opinion and word-of-mouth effects on the target company. Such research would also verify the empirical observations stating that negative experiences with brands tend to spread faster and wider than positive ones. Such research can provide data on the correlation between social customer service experience and electronic word-of-mouth effect on the brands and products. The text mining of social customer service data can also be used to reveal faulty products designs, inferior quality and unwelcomed features, acting as an convenient and inexpensive source of feedback for new products and service launching.

Reference

An exhaustive study of Twitter users across the world. (2012). Retrieved from http://www.beevolve.com/twitter-statistics/#b1

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011, February). Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 65-74). ACM.

Bifet, A., & Frank, E. (2010, October). Sentiment knowledge discovery in twitter streaming data. In *Discovery Science* (pp. 1-15).

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1-8.

ComScore. U.S. Smartphone subscriber market share reports July 2015. (2015, September 3). Retrieved from https://www.comscore.com/Insights/Rankings/comScore-Reports-July-2015-US-Smartphone-Subscriber-Market-Share

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*, 12.

Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.

J.D. Power and Associates. Poor social media practices can negatively impact a businesses' bottom line and brand image. (2013, February 14). Retrieved from http://www.jdpower.com-/press-releases/2013-social-media-benchmark-study

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In *Machine learning: ECML-98* (pp. 4-15).

Liu, B. (2010). Sentiment Analysis and Subjectivity. *Handbook of natural language processing*, *2*, 627-666.

McCallum, A., & Nigam, K. (1998, July). A comparison of event models for Naive Bayes text classification. *In AAAI-98 workshop on learning for text categorization* (Vol. 752, pp. 41-48).

Nielsen. State of the media – the social media report 2012. (2012, March 12). Retrieved from http://www.nielsen.com/us/en/insights/reports/2012/state-of-the-media-the-social-media-report-2012.html

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, *2*(1-2), 1-135.

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). Servqual. Journal of retailing, 64(1), 12-40.

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. *In IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46). IBM New York.

Saif, H., Fernandez, M., He, Y., & Alani, H. (2013). Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold. In: 1st International Workshop on Emotion and Sentiment *in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013), 3 December 2013, Turin, Italy.*

Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of twitter. *In The Semantic Web–ISWC 2012* (pp. 508-524).

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267-307.

Taylor, S. A., & Baker, T. L. (1994). An assessment of the relationship between service quality and customer satisfaction in the formation of consumers' purchase intentions. *Journal of retailing*, *70*(2), 163-178.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. ICWSM, 10, 178-185.

Twitter usage & company facts. (2016, March 13). Retrieved from https://about.twitter.com/-company