Big Data at AT&T Regional Fall Case Competition

The University of Texas at Dallas

Group Member:

Hua Guo
Chao Li

Table of Contents

In this competition, the final goal is to rate the retail zones by zip codes based on customers’ satisfaction. We plan to extract customer reviews, ratings, geo-locations and other useful information from social media platforms (such as Twitter, Yelp, Google, and so on), and explore the correlation between social media and retail store performance. In 1st round, we’ll use Yelp, Google and Twitter as examples. We may collect data from other platforms in following round.

There are two main sections in our report: data collection and data analysis. In the data collection section, we explain: 1) how to fetch raw data from social media platforms, 2) how to select key factors from raw data for each social media platform. In the data analysis section, we focus on stating our findings from selected data, sentiment analysis of customers’ reviews and our ranking strategy.

1. Data Collection

We classified these social media platforms into 2 types. A type of platforms provides reviews and ratings (such as Yelp and Google), and the other type only provides reviews (such as Twitter). Please see details in Fig. 1. We’ll fetch raw data first and select useful data based on the raw data provided by distinct platforms.

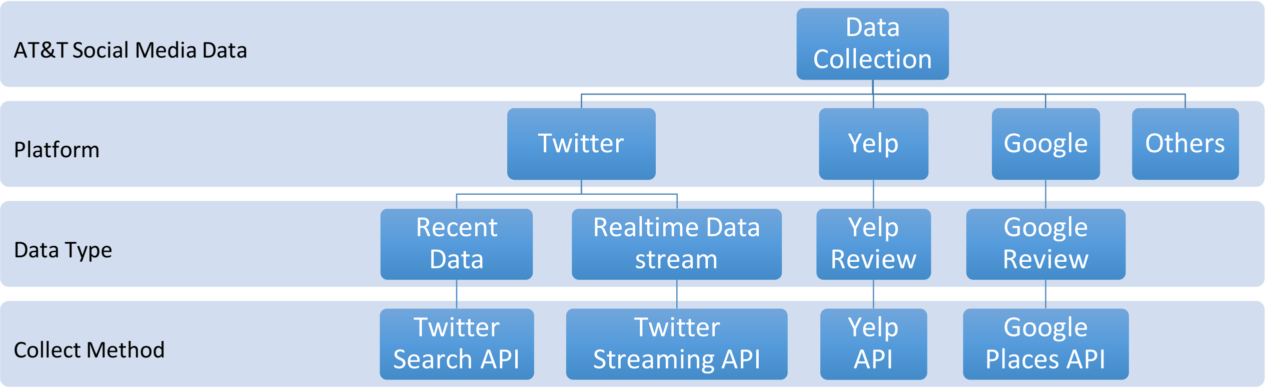


Fig. 1 data collection summary

1.1 Platform providing reviews and ratings

1.1.1 Yelp

We collect raw data of Yelp directly by using its APIs. Depending on different Yelp APIs location specifying methods, the search parameters for searching AT&T retail stores in Dallas area can be { ‘term’: ‘at and t’, ‘sort’: 0, ‘radius_filter’: 40000, ‘location’: ‘Dallas, TX’ } or {‘term’: ‘at and t’, ‘sort’: 0, ‘radius_filter’: 40000, ‘bounds’=32.675176, -97.023954|33.172869, -96.554031}.

Term is the search term (e.g. “food”, “restaurants”). The term keyword also accepts business names such as “Starbucks”. So we use “at and t” as the term keyword for AT&T retail stores. Sort = 0 means searching business best matched with term. Radius_filter is the search radius in meters. Location just means location where you want to search. We can use a city name like “Dallas, TX” or use a Geographical Bounding Box called “bounds” by representing the latitude &longitude at the southwest corner and the latitude &longitude at the northeast corner. 32.675176, -97.023954 is a latitude &longitude coordinate at the intersection of I-20 and President George Bush Turnpike while 33.172869, -96.554031 is a latitude &longitude coordinate at the intersection of 380 and Lavon Lake.

There are tons of variables in the response raw data. Because our final goal is to rate the retail zones by zip codes based on customers’ satisfaction, we only select data as key factors related to store identity, geo-location and customer reviews (Table 1). Every AT&T store has an exact rating score. We can take the score as an important variable for final store rating calculation. However, Yelp API only returns 1 review comment. The returned review number is too small to help our analysis, so we don’t consider it in Yelp platform.

Table1 The structure of data selected from Yelp

Yelp Review
Store_id
Name
Display_address
Postal_code
(latitude, longitude)
Rating [1.0,5.0]

Table 2 Sample data from Yelp

Store_id	Name	Display_address	Postal_code	Latitude	Longitude	Rating
at-and-t-dallas-9	AT&T	["1152 North Buckner Blvd", "Ste 114-A", "Dallas, TX 75218"]	75218	32.8348	-96.7021	3.5
at-and-t-authorized-retailer-dallas-2	AT&T Authorized Retailer	["5618 E Mockingbird Ln", "Lower Greenville", "Dallas, TX 75206"]	75206	32.8363	-96.7715	5
at-and-t-mckinney	AT&T	["1681 N Central Expwy", "Suite 450", "Mckinney, TX 75070"]	75070	33.2142	-96.6380	5

1.1.2 Google

Google is an interesting platform, not only providing an overall rating but also giving at most 5 customer comments and their respective scores. We may collect all available comments of telecommunication retailer to serve as a corpus, and use it as training data to score the non-rating comments in our review sentiment analysis part.

We use Google Places API to fetch our raw data. There are 2 steps. First, we grab placeid of all AT&T retail stores from Google by text search request with parameters { query=‘ATT’,lat_lng={‘lat’: 32.9803, ‘lng’: -96.7674}, radius=50000}. ATT is the query keyword. (32.9803, -96.7674) is the center coordinate of our searching area for Dallas area. Radius defines the distance (in meters) within which to bias place results. The maximum allowed radius is 50 000 meters. In the second step, we get place details by place details request with placeid we grabbed in the previous step.

Google overall rating data

From Google raw data, we collect the following data as our key factors including store information, reviews and ratings of AT&T stores located in Dallas area (Table 3).

Table 3 The structure of data selected from Google

Google Review

Store_id
Name
Zipcode
Address
Overall Rating [1.0,5.0]

Table 4 Sample data from Google

Store_id	Name	Zipcode	Address	Rating
AT&T-75218-1	AT&T	75218	1152 North Buckner Blvd, Ste 114-A, Dallas, TX 75218, USA	2.5
AT&T-authorized-retailer-75206-1	AT&T Authorized Retailer	75206	5618 E Mockingbird Ln, Dallas, TX 75206, USA	3.8
AT&T-75070-1	AT&T	75070	1681 N Central Expy #450, McKinney, TX 75070, USA	3.3

Google comments with score data

We collect all available comments with their respective scores of AT&T retailors. Use the comments and scores as a corpus to further estimate AT&T customer reviews got from Twitter.

Table 5 The structure of sentiment analysis corpus selected from Google

Sentiment Analysis Corpus
Reviews
Rating (0~3)

Table 6 Sample data of sentiment analysis corpus from Google

Reviews	Rating
I was extremely happy with the service here. The account manager was friendly and helpful. It's out of the way and not busy like the ones further in town.	3
WORST SERVICE EVER. THIS STORE IS NOT EVEN A BUSY STORE. Walked on 12/22 and WAS NOT GREETED BY ANYONE. Two male employees was with a customer and did not greet me or tell me they would be with me momentarily.	0
Staff was extremely helpful and friendly and there is usually not a long wait like there is at so many of the other locations.	2

1.2 Platform without rating

Twitter provides thousands of tweets every second. So, it becomes a great social media platform for customers to interact with their wireless/internet service companies in real time. There three issues when we dealing with tweets data. First, those customer comments don’t have ratings. Second, a tweet may contain many elements, not follow common language rules, and use emoticons or hashtags convey messages. Third, we want to get real-time customer feedback. We will leave the discussion of the first and second issue in section 2. In this section, we mainly focus on the third one – fetching data in real time.

1.2.1 Twitter-recent public data

Before talking about real-time twitter data stream, we first look at another data collection method provided by Twitter Search API – twitter recent public data. Twitter Search API allows queries against the indices of recent or popular Tweets and behaves similarly to, but not exactly like the Search feature available in Twitter mobile or web clients, such as Twitter.com search. This twitter recent public data is a sampling of recent Tweets published in the past 7 days.

We collect two types of recent public data: tweet information and user information. Tweet information includes tweet text, retweet count and so on, while user information provides user location, followers count and so on. Particularly, tweet text represents customer comments. Tweet.location and user.location together provides us geo-location information of retail stores if we assume customers will go to the nearest store away from their home. Retweet count and favorite count is used to represent the impact of comments. Followers count is used to represent the impact of a user.

Table 7 The structure of data selected from Twitter

Tweet	User
id	id
tweetID	screenName
tweet text	location
favoriteCount	followersCount
retweetCount	
screenName	
location	

Table 8 Sample data from Twitter

id	tweetID	text	favoriteCount	retweetCount	screenName	location
21	7883720000000XXXXX	Credit to @ATT for responding to resolve the issue quickly. System still way too complicated and confusing, though. https://t.co/RdJbnqykwO	1	0	tXXXX	NA
69	7883720000000XXXXX	I love coming home from work everyday to internet service that never works (: @ATTCares	1	0	AXXXX	NA
1225	7883720000000XXXXX	RT @ATT: @ajanata Who doesn't love a happy surprise like that? Enjoy your new place and the great speeds, Andy! https://t.co/mmGN7TEbVS	0	1	LXXX	NA

1.2.1 Twitter-real time data stream

We can fetch twitter data stream in real time by the streaming APIs. As described in Twitter's developer website, the Streaming APIs give developers low latency access to Twitter's global stream of Tweet data. The mainly used parameters for streaming API request is {'track'='ATT', 'language'='en', time=300, 'locations'=-97.023954, 32.675176, -96.554031, 33.172869}. Track is the keyword for tracking tweet data stream. Language equaling to 'en' means we focus on tweets in English. Time is measured by second, indicating the tracking time. Locations are longitude and latitude coordinates of a Geographical Bounding Box.

Because we cannot track twitter data stream all the time. It is not allowed by Twitter. Therefore, our real-time data collection strategy is to capture data for 5 minutes and wait for 2 minutes to process raw data. For real-time data stream, exact tweet match is not supported by Twitter. We use the waiting time to extract data related to AT&T cares or helps such as tweet with "@ATT" or "@ATTCares". We also extract geo-location information from data stream by visiting label place_lat and place_lon. Sometimes, place_lat and place_lon are missing. We'll track user's location by getting user profile location with user_id or screen_name. One real-time tweet example is as follows: "I'm keeping my #Note7 I upgraded just for this damn phone. Can I get my upgrade back on my line if I return it? @ATT @SamsungMobileUS".

1.3 Data from other platforms

There should be many other social media platforms which can play an important role in analyzing retail store performance. For example, AT&T has its own forum to help solving customer's problems. We may scrape this forum to fetch useful information like customer reviews and geo-locations.

Facebook has a public page for business companies like AT&T and it provides detailed page insights for page administrators.

Other platform like AT&T's e-chat also has potential to provide millions of useful data, but we do not have authority to access them. If AT&T is glad to share a sample of such kind of data, maybe we can dig out something valuable.

2. Data Analysis

2.1 Social Analytics Metrics

We found some interesting things during the data analysis process. We'll describe our social analytics findings by breaking down metrics as sentiment, mentions, engagement, and impact.

Sentiment is the most important social media metric to measure retail store performance. Sentiment refers to the emotion and attitude of a customer behind the social media mention – is the customer happy, annoyed or angry? Sentiment has a very direct correlation to retail store performance: positive meaning satisfaction and good service while negative meaning anger and bad performance (more details in section 2.2).

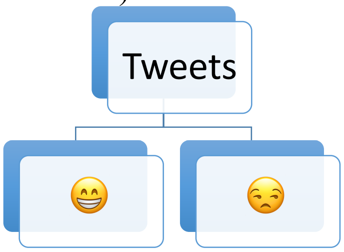


Fig. 2 A simple sentiment classification: Positive/Negative Tweets

Volume of mentions is the size of conversation. From the mention volume we can know the average mention numbers of everyday, the peak of time, and monitor the abnormal. From the result we may infer local store performance, although it provides limited help.

Engagement is another key factor of social media metric. When people take time to favorite or retweet on comments in Twitter, they're actively engaging with the comment tweets. Engagement will highlight the good service provided by a retail store through Twitter but also will amplify a bad performance influence as fast as broadcasting a rumor.

Impact is the last social analytics metric we describe here. The retweeting count of a positive or negative tweet may amplify customers' sentiment effect to the AT&T image. We'll rate high impact tweets by a relatively larger weight when score the retail stores.

2.2 Sentiment Analysis of reviews

2.2.1 Data Preprocess

The data got from Yelp and Google is almost clean. So we mainly do data preprocess for data from Twitter. As we mentioned before, a tweet may not follow common language rules and use lots of emojis or hashtags or the "at" symbol @. So, we use text parsing methods first to cut out hashtags and @keyword, and only keep tweets with #ATT or @ATT or @ATTCares. Second, we remove all unimportant words from fetched tweets like punctuation, numbers, retweet entitles, html links and unnecessary spaces. In the next step, we convert all capital letters to lower case letters. Finally, we tokenize each text and extract words like adjective (good, bad, etc.), verb (like, dislike, hate, etc.), noun and some adverb (well, hard, etc.). After this data processing, for each left tweet, we have a bag of words ready for further sentiment analysis.

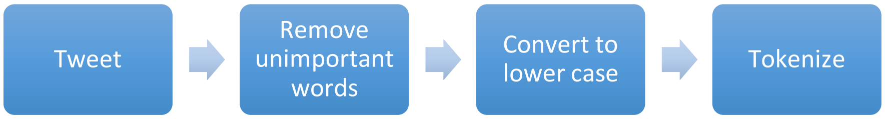


Fig. 3 Tweet preprocess

2.2.2 Sentiment Analysis

Word Frequency Counting Method

After data preprocess, sentiment of a tweet can be automatically determined based on customer's description words by employing machine learning approaches. Usually such a task can be helped by a lexicon full of related words.

Currently, the sentiment analysis method we employ is to count the positive words and negative words in each tweet's bag of words. We download a sentiment lexicon from Liu's sentiment analysis webpage [1]. For each word of a processed tweet, we explore the matched sentiment word in this lexicon. If we find one positive word, the count of positive words is added by one, and so is the negative count. The final sentiment of this processed tweet calculated by the following formula:

$$\text{Estimate Sentiment} = \text{positive count} - \text{negative count}$$

Here is a sample of sentiment evaluation results:

Table 10 sentiment evaluation results sample

Original tweet text	sentiment
@ATT your TV is aids, and your internet is cancer. Let me know when you get your pathetic shit together. Retarded fucks.	-4
RT @ATT: @ajanata Who doesn't love a happy surprise like that? Enjoy your new place and the great speeds, Andy! https://t.co/mmGN7TEbVS	5
Sweet. Thanks @ATT. Almost feel like it's a smarter risk/reward to just not buy a case now for my new iPhone https://t.co/sSYPM8vvR6	3
@ATTCares Worst customer service ever! You will be losing a long time customer very shortly. #att #attsucks	-2

Bigram Naïve Bayes Method

Although the evaluated sentiment results we get look very good, we still think we can improve the predicted results if we use the Google reviews as our training set. It is important to use a corpus that is related to the domain we want to classify text for. A generic lexicon sometimes may not help this AT&T case. For example, one popular lexicon, Hu &Liu's opinion lexicon, is for digital products review. Another famous corpus, Stanford Twitter sentiment corpus, is based on all kinds of tweets rather than comments for telecommunication stores.

Therefore, in the future, we plan to use our own sentiment database for tele industry which is fetched from Google in section 1.1.2, and use it to score tweets. According to our analysis on Google reviews, there are some keywords or phrases indicating retail store performance in tele industry. Details are in Table 11. From the word cloud of Google reviews (Fig.4), we can also find the aspect that customers care most is the SERVICE and the number 1 service that customers require is phone related stuff.

Table 11 sample positive and negative keywords to evaluate store performance

positive	negative
recall, replacement, replace, good service, awesome service, helpful, friendly, in stock, great job, upgrade, knowledgeable, low pressure	worst service, bad service, wait, waiting, out of stock, incorrect bill, never greeted, never helped, careless, pushy,

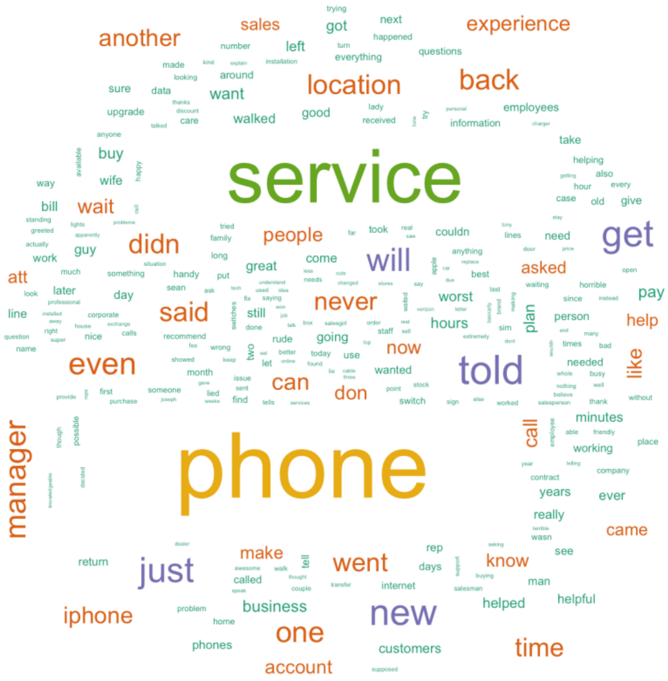


Fig. 4 Word cloud for Google reviews

Obviously, if we use method like unigram based naïve Bayes to predict the sentiment of reviews from other platforms without ratings like Twitter, we'll fail. Two-word phrases have opposite sentiment meaning to one word. So the next machine learning method we plan to use is bigram naïve Bayes. With Google reviews and their ratings as the training corpus and bigram naïve Bayes method, the prediction performance will be improved a lot.

Further, we can divide each customer review into smaller groups of words like different services – TV/Internet/Wireless/Cellphone or different rating aspects – professional knowledge and service attitude etc. Based on this detailed clustering, we can rate the performance of stores from different views and improve their service more specifically. We can also add time information to track the performance retail stores in a certain time period or predict the trend of their service quality.

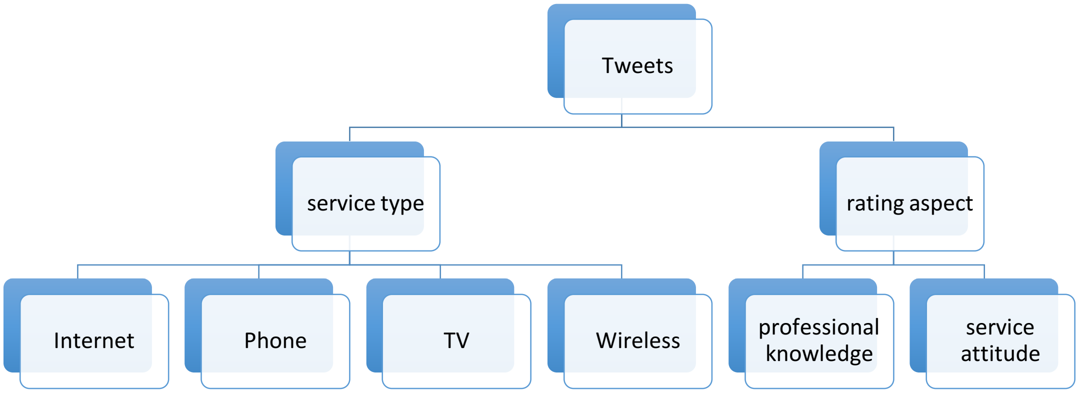


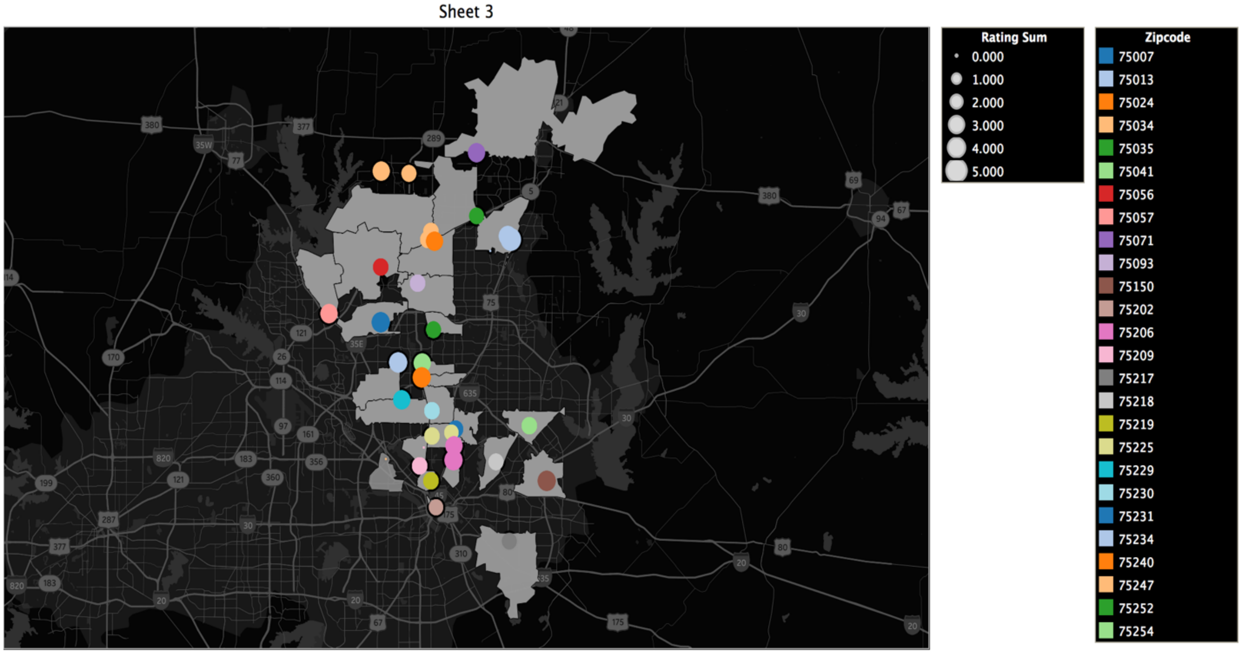
Fig. 5 Subdivided positive/negative sentiment

However, maybe we need to add comparison with competitors, because the majority of comments in Twitter is negative. People always complain through Twitter. We plan to analyze the comments to Verizon and Tmobile, and use the negative comments percentage as a comparison metric to rate retail store performance in a zipcode area. In the same zipcode area, if the AT&T stores have less negative comments percentage than their competitors, that would mean better service.

2.3 Ranking strategy

Because Yelp and Google can provide accurate geo-location information to arbitrary retail store, we give their ratings higher weights in the final ranking computation (all ratings are scaled to 1~5 before final computation):

$$\text{Rank} = 0.35 * \text{Yelp rating} + 0.35 * \text{Google rating} + 0.3 * \text{Twitter sentiment rating}$$



Map based on Longitude and Longitude and Latitude. For pane Longitude (2): Color shows details about Zipcode. Size shows Rating Sum. For pane Longitude: Details are shown for Zipcode.

Fig. 6 Final rank of each AT&T retail store we collect in Dallas area. Circles with the same color means stores in the same zipcode area. This figure is created by Tableau (some zipcode area provided in Tableau is not correct like 75034).

As we mentioned is section 2.2.2, if we subdivide each customer's review into smaller categories, we can provide suggestions for each retail store on how to improve their service. For instance, if the professional knowledge factor of one retail store is rated very low while its service attitude is rated higher, the training of professional knowledge can be strengthened in this store.

Reference:

[1] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html