**Big Data at AT&T Regional Fall Case Competition**

**Round 2**

**AT&T Customer Service Insights**

**via Social Media Analytics**

**The University of Texas at Dallas**

Group Member:

Hua Guo

Chao Li

Table of Contents

[1. Data Collection 3](#_Toc465973095)

[1.1 Platform providing reviews and ratings 4](#_Toc465973096)

[1.1.1 Yelp 4](#_Toc465973097)

[1.1.2 Google 5](#_Toc465973098)

[1.2 Platform without rating 6](#_Toc465973099)

[1.2.1 Twitter-recent public data 7](#_Toc465973100)

[1.2.1 Twitter-real time data stream 7](#_Toc465973101)

[1.3 Data from Facebook and other platforms 8](#_Toc465973102)

[Facebook 8](#_Toc465973103)

[Other platform 8](#_Toc465973104)

[1.4 Demographics Data 8](#_Toc465973105)

[1.5 Name and Gender data 8](#_Toc465973106)

[2. Data Analysis 8](#_Toc465973107)

[2.1 Social Analytics Metrics 8](#_Toc465973108)

[Sentiment 8](#_Toc465973109)

[Volume of mentions 9](#_Toc465973110)

[Engagement 9](#_Toc465973111)

[Impact 9](#_Toc465973112)

[2.2 Sentiment Analysis of reviews 9](#_Toc465973113)

[2.2.1 Data Preprocess 9](#_Toc465973114)

[2.2.2 Sentiment Analysis 9](#_Toc465973115)

[2.3 Ranking strategy 12](#_Toc465973116)

[2.4 Service aspects extraction based on LDA 13](#_Toc465973117)

[2.5 Customer Service Insights 14](#_Toc465973118)

[2.5.1 Hidden service aspects 14](#_Toc465973119)

[2.5.2 Switching reasoning 14](#_Toc465973120)

[2.5.3 Twitter Trend 14](#_Toc465973121)

[2.5.4 Demographics analysis 14](#_Toc465973122)

[2.5.5 Competitor telecommunication companies 14](#_Toc465973123)

[Reference: 14](#_Toc465973124)

This paper is an integration of our 1st round report and what we did in 2nd round, because many jobs are done based on 1st round work.

In the 2nd round, the final goal is: 1) learn what’s been said in the context of AT&T through social media. Some of aspects related to customer service analyzed are: technician dispatch, product installation, customer experience across various channels (store, live chat, service line), satisfaction around product usage, billing issues, responsiveness, quality, prices, reasons for churn, reasons for staying with the network and so on. 2) Perform location analysis to see if the services and product experience vary by zones or zip codes. 3) Investigate things like location based demographics to see if that plays a role in AT&T’s overall customer sentiment. Therefore, we extracted customer reviews, ratings, geo-locations, time and other useful information from various social media platforms (such as Twitter, Yelp, Google, Facebook and so on).

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 ranking strategy and AT&T customer service insights we found from social media data.

# 1. Data Collection

We classified these social media platforms into 2 types. A type of platforms provides reviews and ratings (such as Yelp, Google and Facebook), and the other type only provides reviews (such as Twitter). Please see details in Fig. 1. We fetched raw data first and select useful data based on the raw data provided by distinct platforms. The total number of data we get from social media platforms is 50808 (Twitter data dates from October 21 to October 30, details in Table 1.1). Furthermore, we also fetched reviews and related tweets from AT&T’s competitor companies like Sprint, TMobile and Verizon. We’ll compare our findings among all these 4 companies in Section 2.5.

Fig. 1.1 data collection summary

Table 1.1 Data summary in different social media platforms

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Yelp | Google | Facebook | Twitter |
| AT&T Data for Dallas area | 206 | 532 | 183 | 188 |
| AT&T Data US | 0 | 7786 | 0 | 5434 |
| Sprint Data | 119 | 8138 | 0 | 3045 |
| Verizon Data | 98 | 7716 | 40 | 4291 |
| TMobile | 86 | 8290 | 36 | 4560 |
| Total | 509 | 32462 | 259 | 17518 |

## 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.

Table1.2 The structure of data selected from Yelp

|  |
| --- |
| Yelp Review |
| Store\_id |
| Name |
| Display\_address |
| Postal\_code |
| (latitude, longitude) |
| Rating [1.0,5.0] |

Table 1.3 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 1.4 The structure of data selected from Google

|  |
| --- |
| Google Review |
| Store\_id |
| Name |
| Zipcode |
| Address |
| Overall Rating [1.0,5.0] |

Table 1.5 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 1.6 The structure of sentiment analysis corpus selected from Google

|  |
| --- |
| Sentiment Analysis Corpus |
| Reviews |
| Rating (0~3) |

Table 1.7 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](https://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 1.8 The structure of data selected from Twitter

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

Table 1.9 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 Facebook and other platforms

Facebook Because the limitation of Facebook APIs, we cannot get reviews of each retail store from their public pages. Fortunately, the number of reviews is not very large, so we directly extract all reviews manually. The total number of reviews for all AT&T retail store in Dallas area is 183.

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.

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.

## 1.4 Demographics Data

In order to analyze the relation between demographics data and our custom service insights, we downloaded Dallas area demographics information from US Census website. Main values include zipcode, population, median age, percentage of people age from 15~44, sex ratio, median household income and percentage of high school graduated people.

## 1.5 Name and Gender data

In order to analyze the gender information of each store’s review, we use the APIs provided from website “genderize”, which can return the estimated gender of a given name. The search sentence is like: https://api.genderize.io/?name=Jeremy. The returned the search result is {"name":"Jeremy","gender":"male","probability":0.99,"count":1840}. As the result says, people named Jeremy are categorized as male with probability of 0.99.

# 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. 2.1 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. 2.2 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:

Here is a sample of sentiment evaluation results:

Table 2.1 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 |

#### 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, we use our own sentiment database for tele industry which is fetched from Google reviews 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. We rescaled all Google review rating from 0~3 to {-1,0,1}. Rating 0 is converted to -1 as negative sentiment, rating 1 converted to 0 as neutral sentiment while rating 2 and 3 is converted to 1 as positive sentiment. Moreover, Google reviews we use as the corpus are extracted and sampled from US Top 100 cities by Google APIs. So this corpus can well represent customer’s sentiment all across US.

Table 2.2 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, |

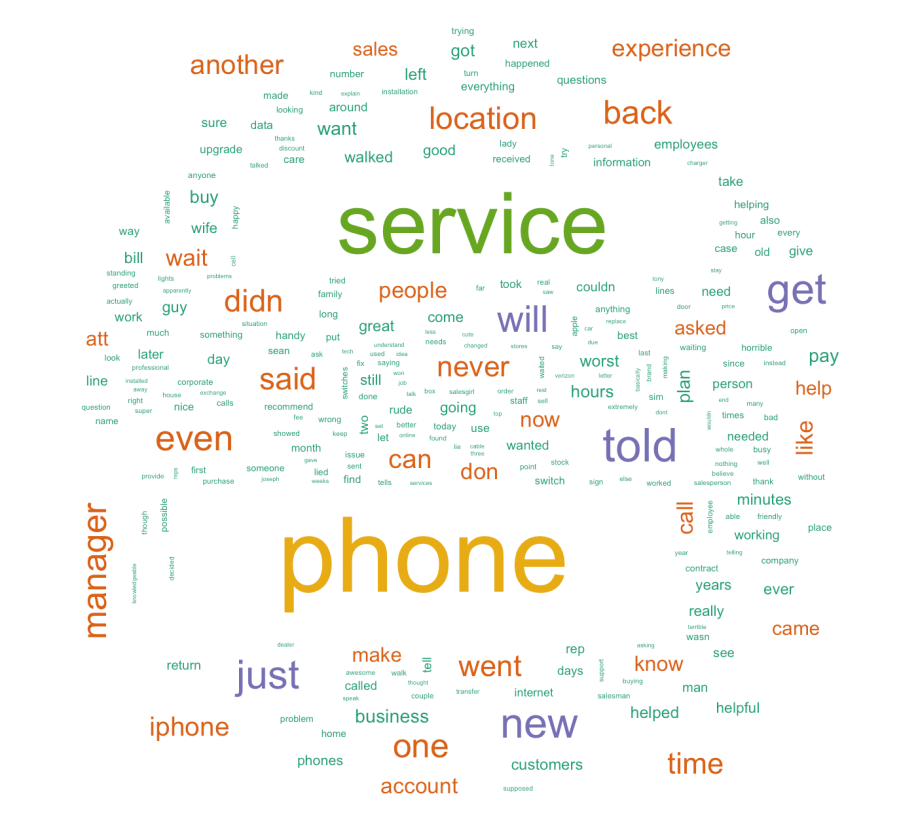


Fig. 2.3 Word cloud for Google reviews

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 use is bigram naïve Bayes. But still the predicted results are not very satisfied.

#### Support Vector Machine Based Sentiment Analysis

The final machine learning method we employed is Support Vector Machine (SVM) which gives us very high accurate in the validation step. We first do preprocessing for all raw review texts including punctuation removal, number removal and stop words removal. Then after converting all character to lower case, we get the final document term matrix. We split the data into 60% and 40% as 60% is treated as training data. Our SVM model has a very good performance with 93.2313% accuracy for training data while 91.0348% accuracy for test data. Table 2.3 is a sample of our predicted tweet sentiment. Positive sentiment is represented by 1, neural is 0 while negative is -1.

Table 2.3 SVM sentiment analysis results

|  |  |
| --- | --- |
| 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. | -1 |
| 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 | 1 |
| 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 | 0 |
| @ATTCares Worst customer service ever! You will be losing a long time customer very shortly. #att #attsucks | -1 |

## 2.3 Ranking strategy

Because currently we can fetch the overall ratings of each retail store from Yelp, Google and Facebook, our final ranking is calculated as the weighted sum of rating from Yelp, Google and Facebook (all ratings are scaled to 1~5 before final computation):

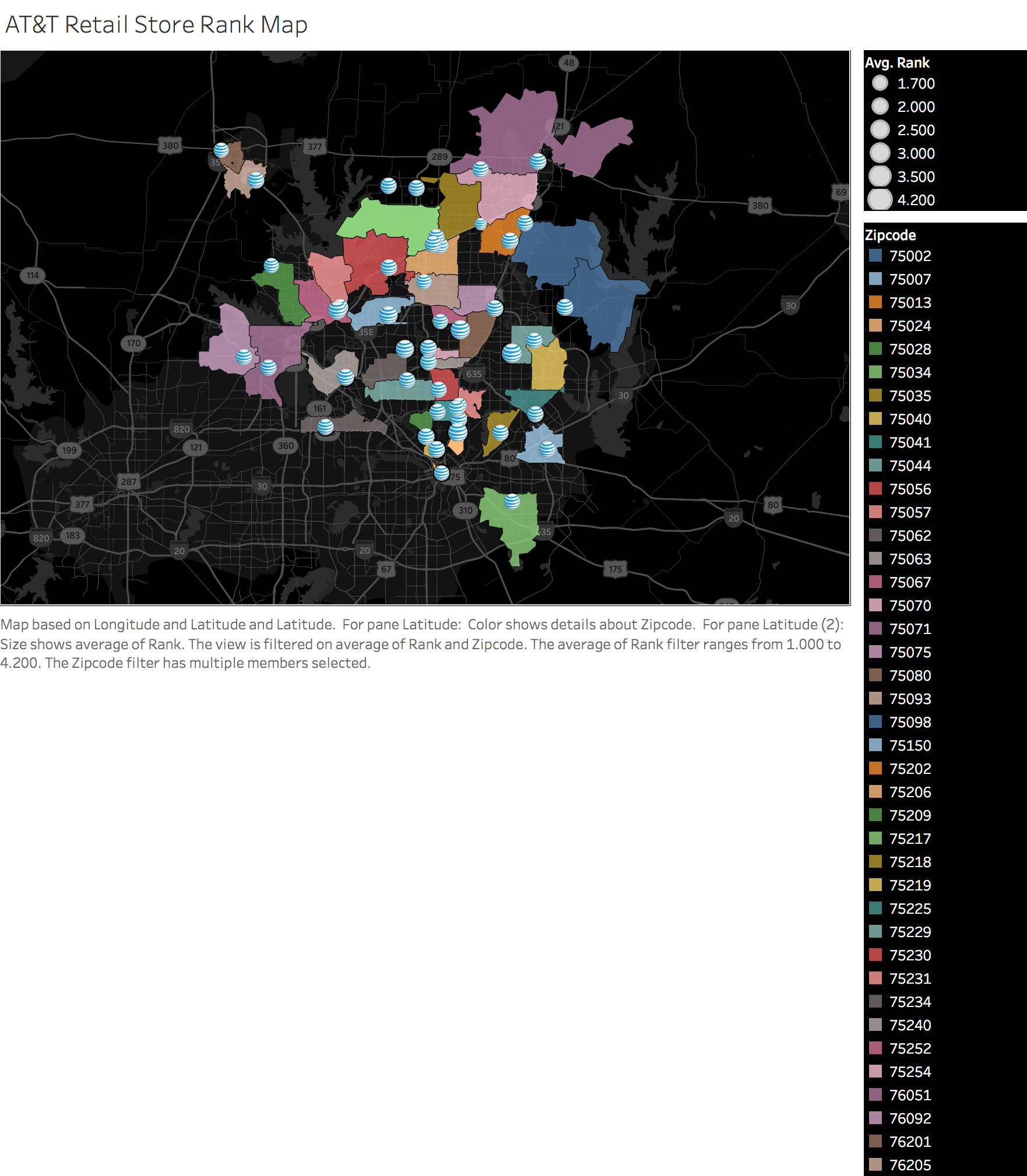


Fig. 2.4 Final rank map of each AT&T retail store we collect in Dallas area. Bigger AT&T brand icon means higher rank. This figure is created by Tableau (some zipcode area provided in Tableau is not correct like 75034).

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.

## 2.4 Service aspects extraction based on LDA

Latent Dirichlet Allocation (LDA) is well known as an excellent topic extraction method [2]. Therefore, we decided to employ such a method to extract service aspects from our collected data.

Suppose we have the following set of sentences:

I eat fish and vegetables.

Fish are pets.

My kitten eats fish.

LDA is a technique that automatically discovers topics that these documents contain.

Given the above sentences, LDA might classify the red words under the Topic F, which we might label as “food”. Similarly, blue words might be classified under a separate Topic P, which we might label as “pets”. LDA defines each topic as a bag of words, and you have to label the topics as you deem fit.

Therefore, as this small example shows, we can use LDA to extract the hidden service aspects from collected customer reviews. After indicating the number of aspects as 15, the extracted hidden service aspects are shown in the following figure (Figure 2.6). From this figure, we can find the most concerned service aspects are: too long waiting time, after-sale service and billing issues. Of course, many customers praised and recommended our AT&T service. The reps are knowledgeable and professional.

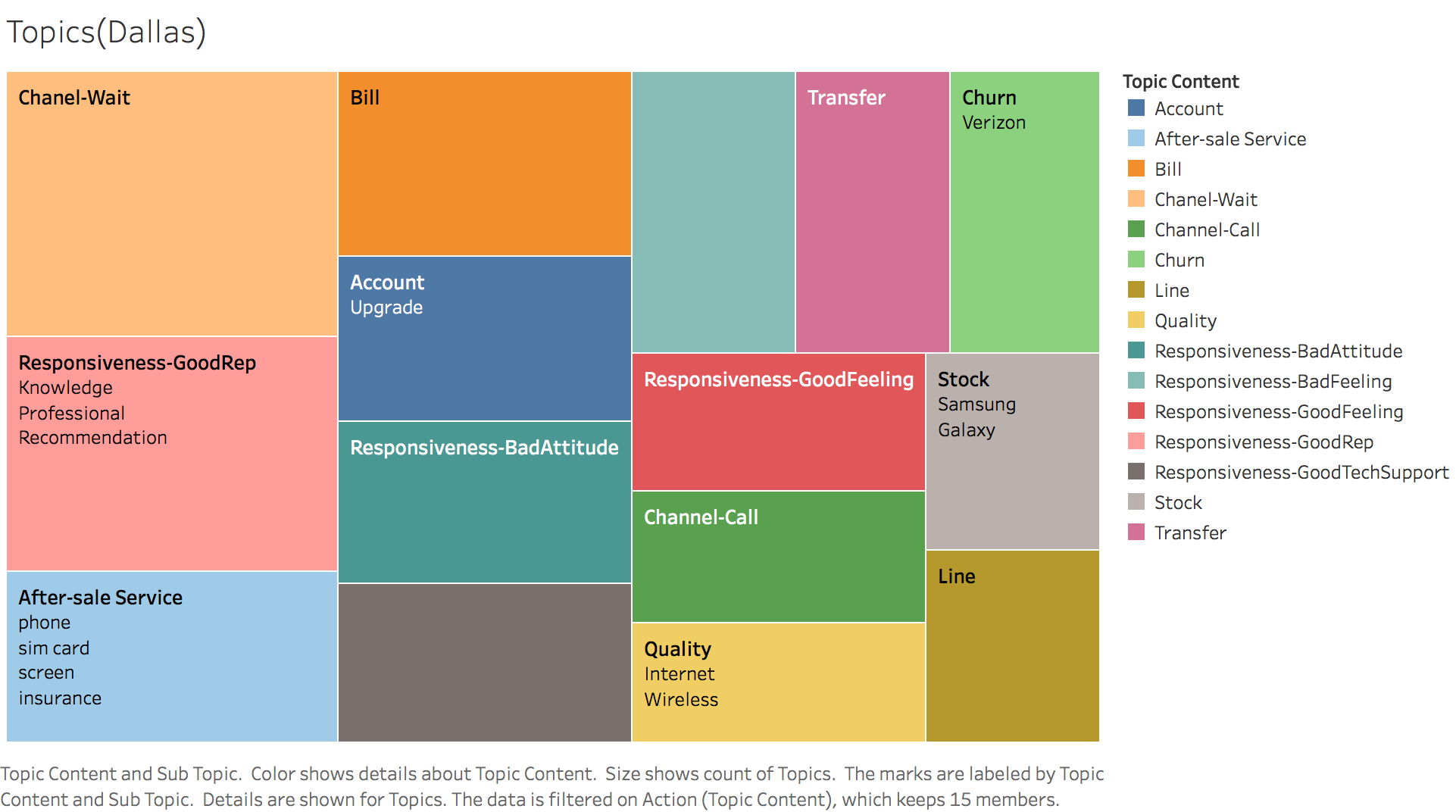


Fig. 2.6 Extracted hidden service aspects by LDA method

## 2.5 Customer Service Insights

In this section, we state customer service insights we found by exploration.

### 2.5.1 Rank analysis

The following figure is the AT&T retail store rank by zipcode. The top 3 high rating zipcode area is 75044, 75080 and 75206. The lowest rank is from area 75035 with rank 1.7. The average rank of all AT&T stores is 3.1 which ranks in the second place of all 4 telecommunication companies.

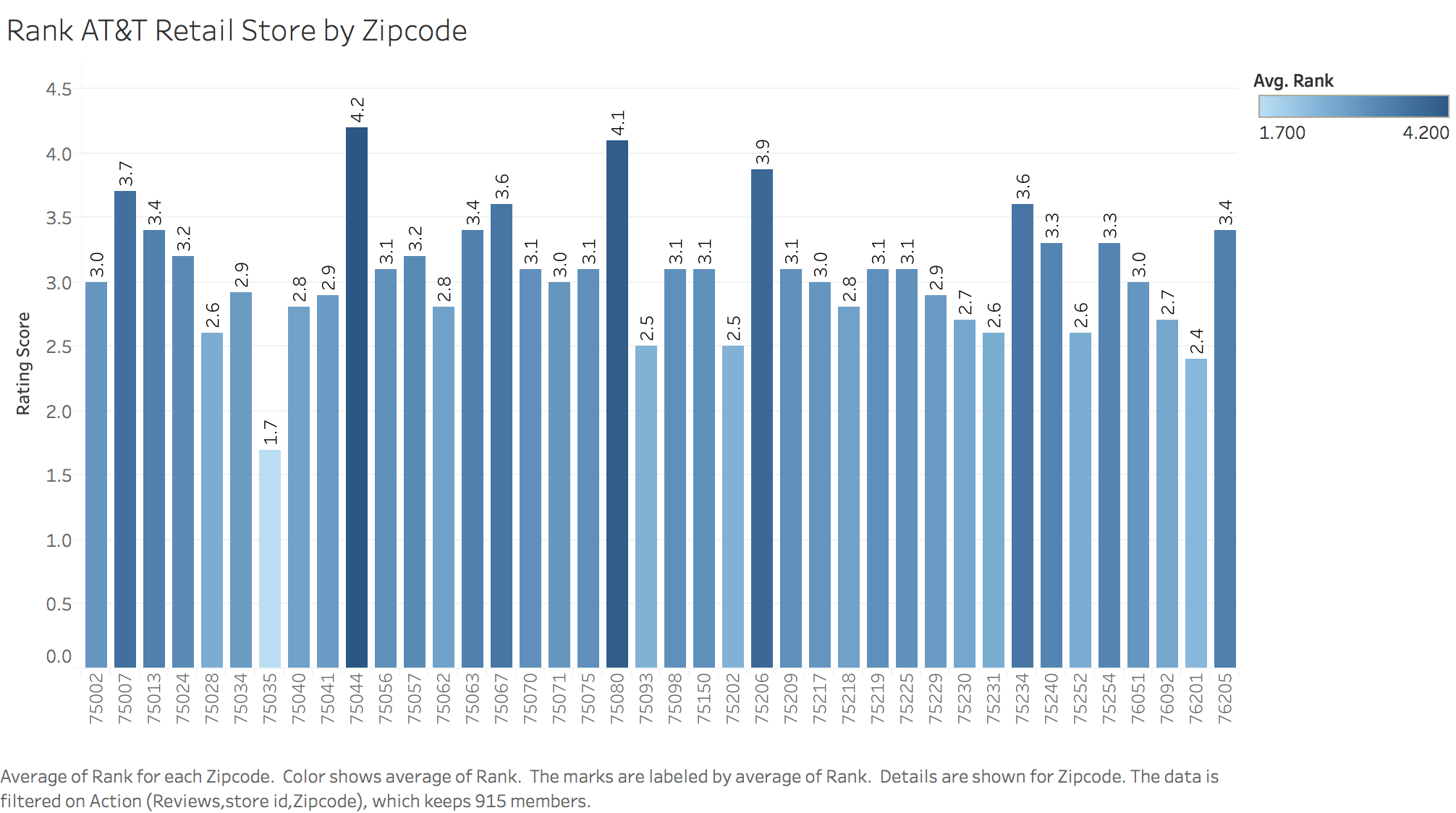


Fig. 2.5 AT&T retail store rank by zipcode

### 2.5.2 Hidden service aspects

In this section and the following sections, topic and service aspect is interchangeable.

After we getting the hidden service aspects from customer’s reviews, we can draw the distribution of all topics like Fig. 2.6. The rank 1 issue customer complain is waiting too long time in the store. The rank 2 topic is thanks and praise from customers. The least concerned issue is customer service hot line. From this type of analysis on hidden service topic, we can recommend that AT&T can train their employees mainly on these aspects like responsiveness, after-sale service and bill dealing strategy.

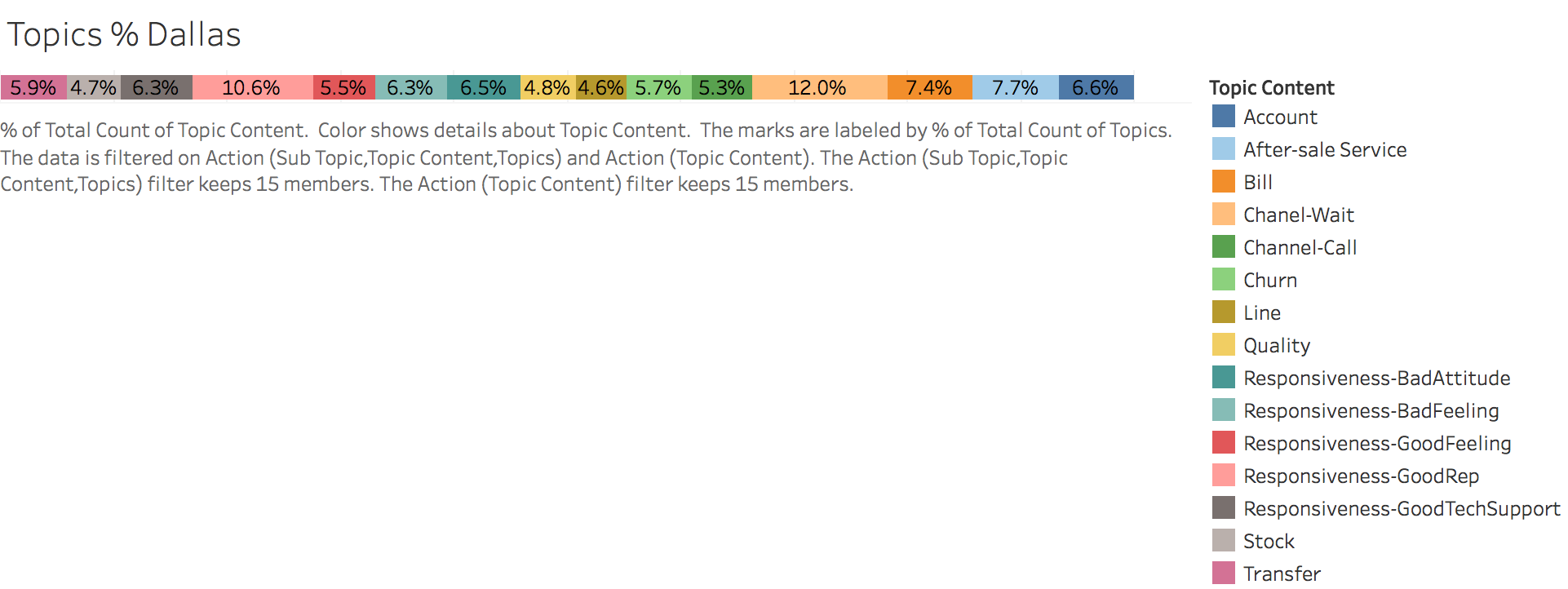


Fig. 2.6 Hidden Service aspect percentage of total in Dallas area

If we drill down to a specific store, like the following Fig. 2.7 showing. We can rate all hidden service topics by averaging the rating of topic related reviews. Of course the highest rating is related to good responsiveness. The biggest problems this store faces are bad representative attitude and quality problem. Fig. 2.8 is a similar figure to show the customer’s satisfaction of an aspect of service. This figure shows more negative sentiment than positive one for aspect “Channel-Wait”. That means customers are not satisfied with too long waiting time.

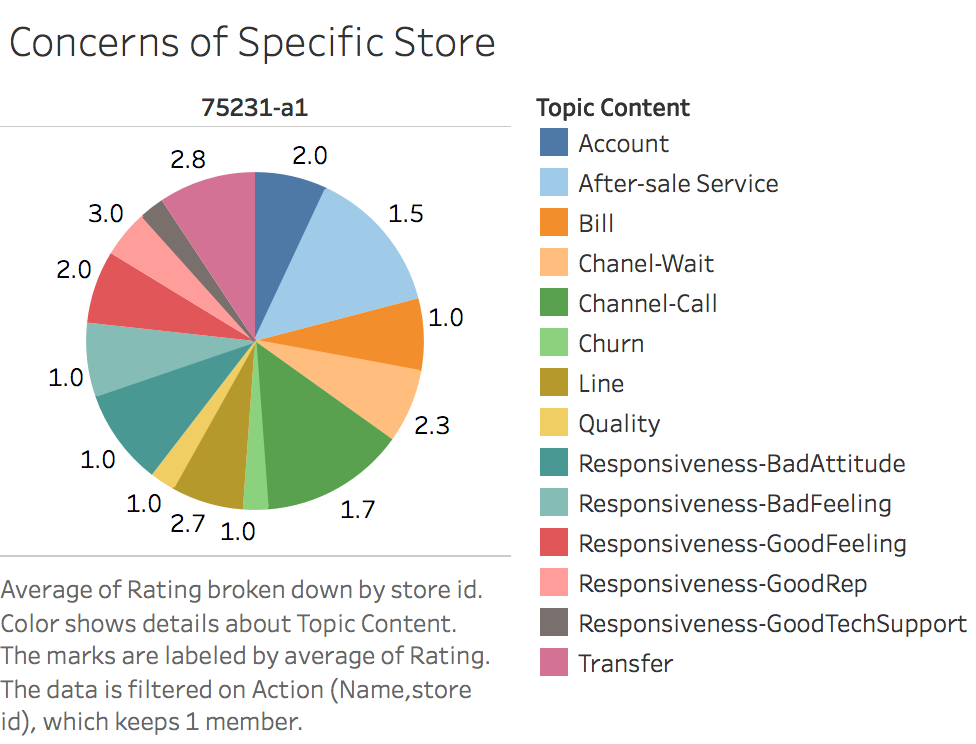


Fig. 2.7 Hidden service aspects rating of a specific store: 75231-a1

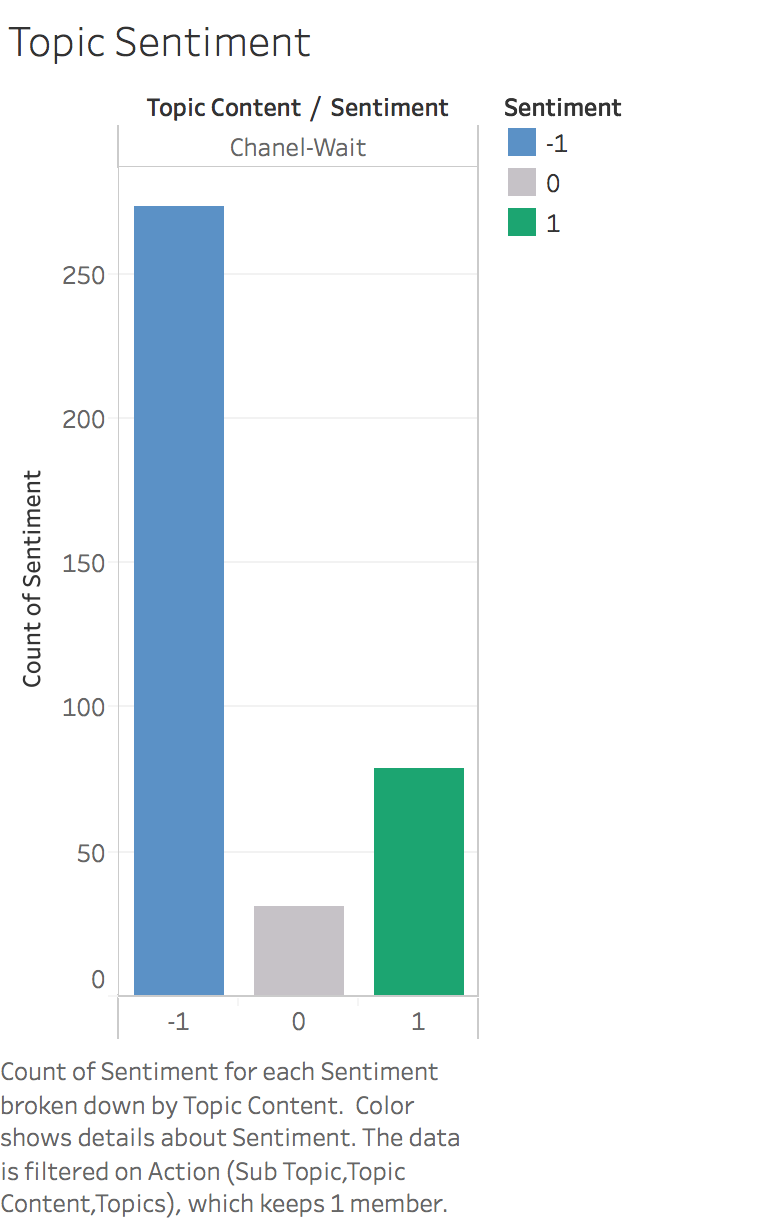


Fig. 2.8 Sentiment distribution of a specific topic—Channel: Wait

### 2.5.3 Twitter Trend

We also extracted and analyzed hot topics from the twitter. Those topics are created by LDA method as well. In Fig. 2.9, blue bar represents the number of a specific topic all across US and purple bar represents the number in Dallas area. Of course, those tweets we got by Twitter APIs is a small number of sample of all tweets, but the ratio of topic distribution should stay the same. Fig. 2.10 shows the trend of a selected topic “Verizon”. As AT&T’s most powerful competitor, I’m afraid many customers will switch to Verizon if they cannot get satisfied customer service especially in weekend.

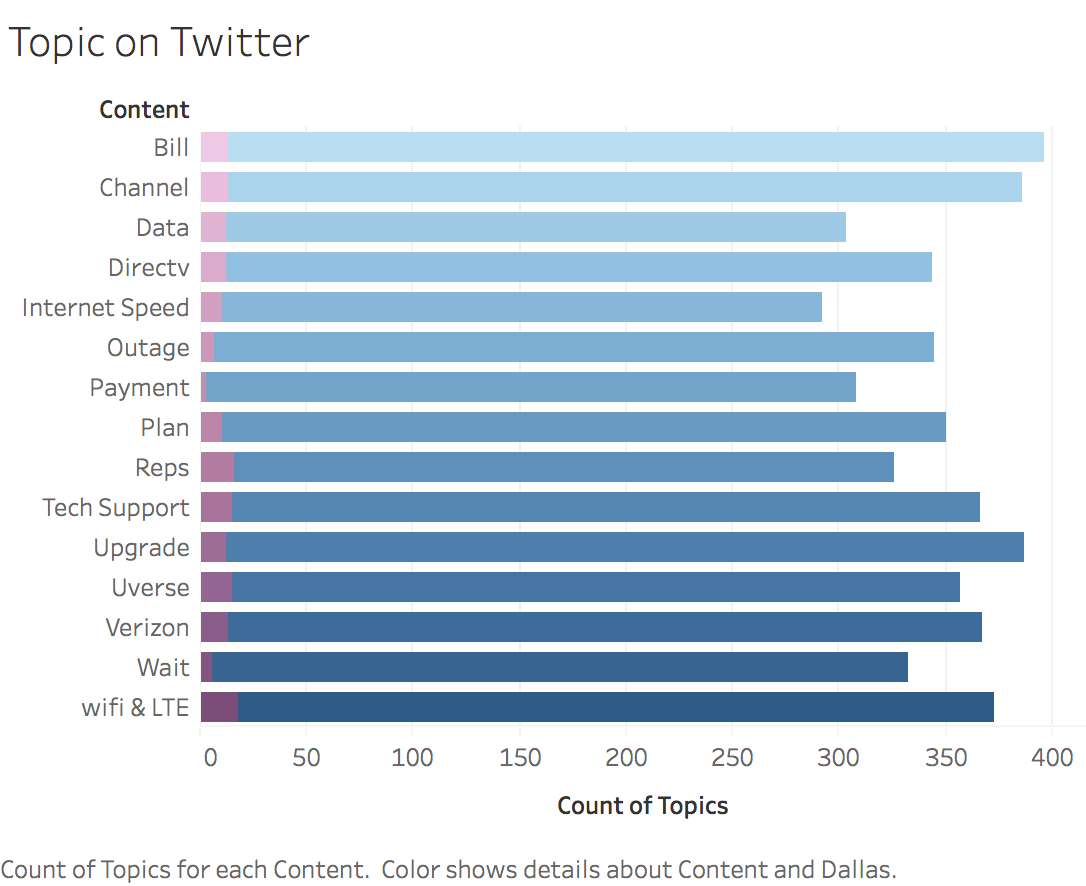


Fig. 2.9 Topic distribution on Twitter

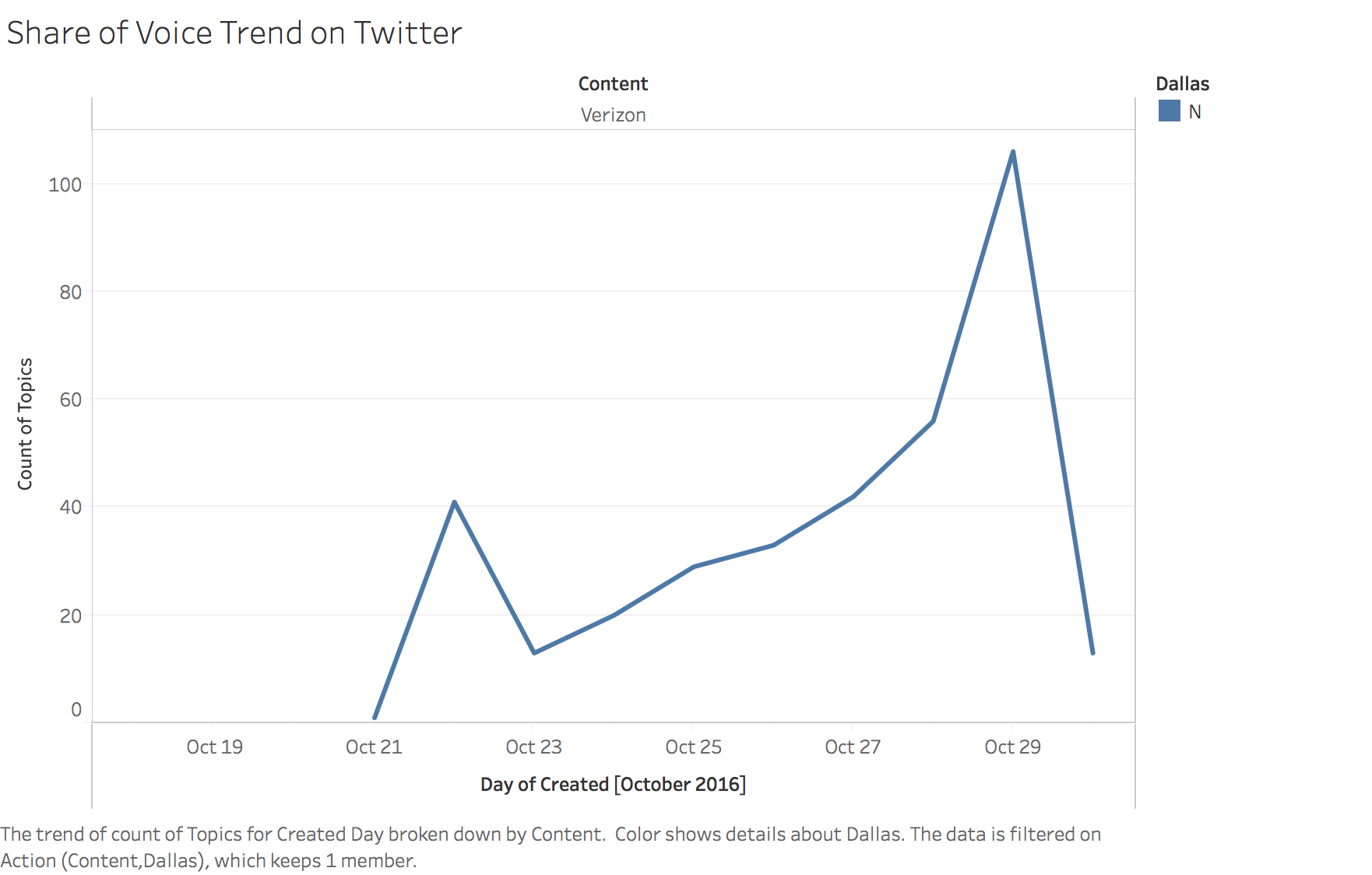


Fig. 2.10 The trend of topic “Verizon” from Oct10 to Oct 30 across US

### 2.5.4 Demographics analysis

We didn’t find any strong correlation between our collected demographics data and AT&T rank/ number of reviews.

Figure 2. shows the computed correlation results and 2 specific correlation line figures.

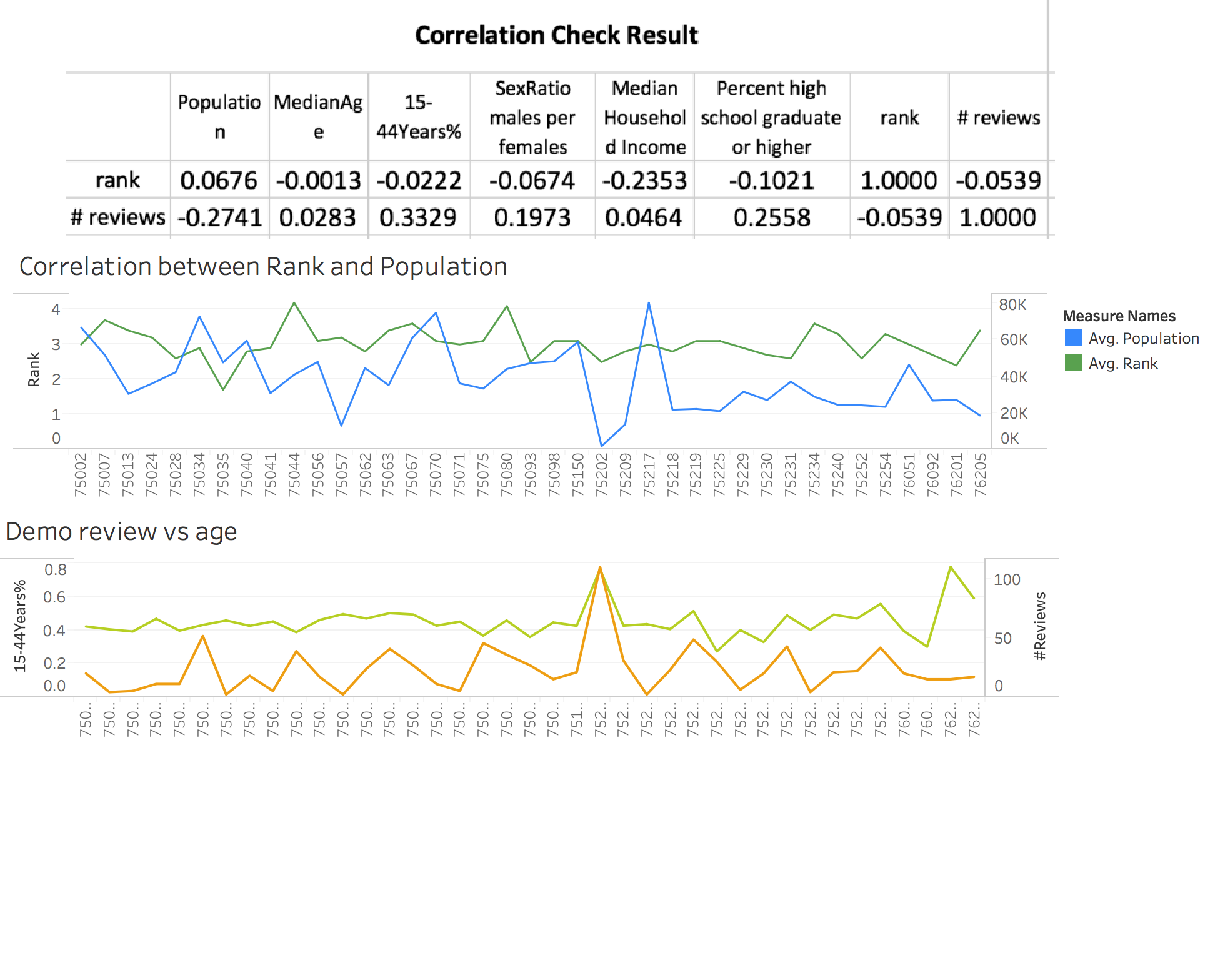
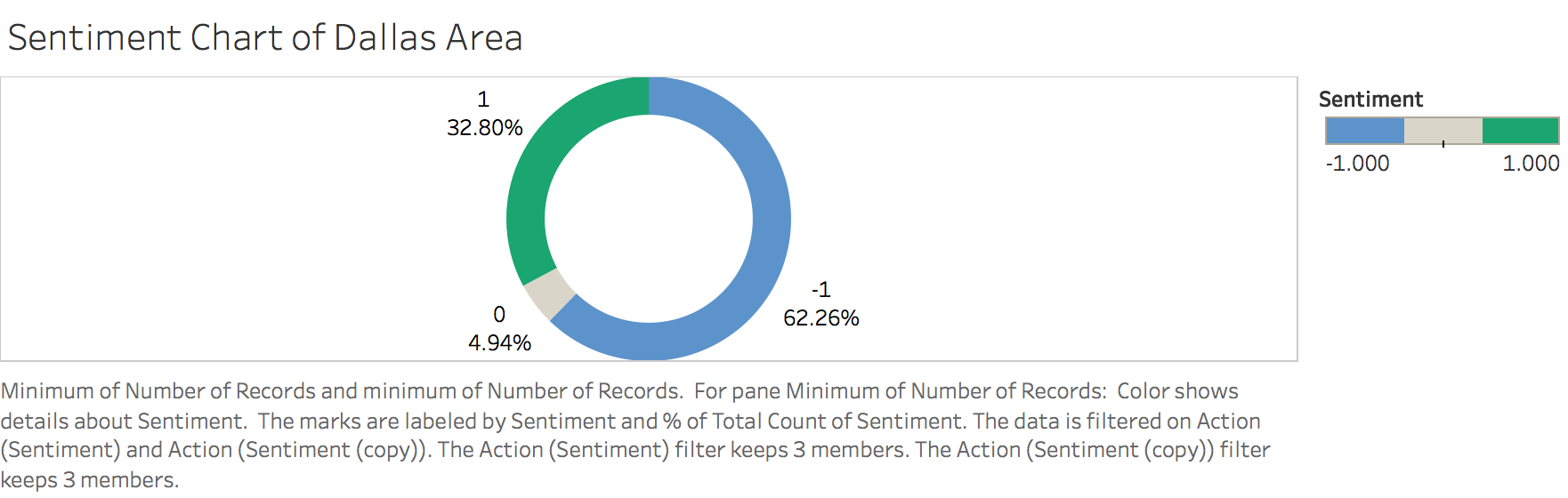
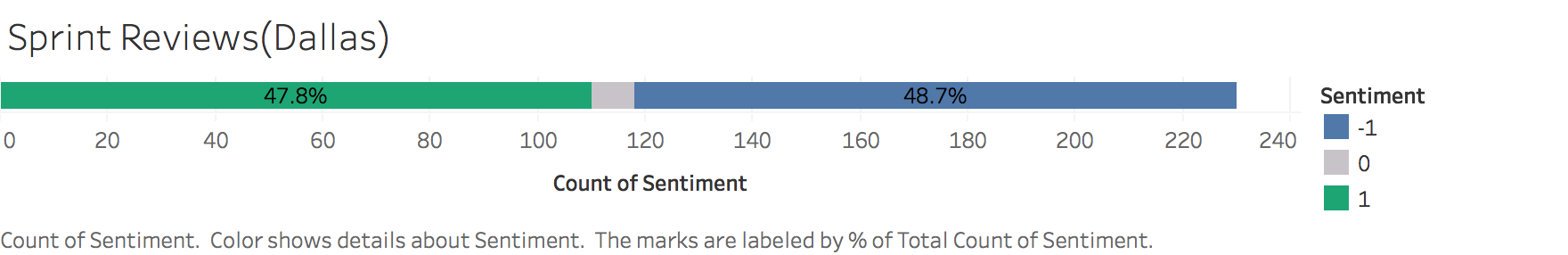


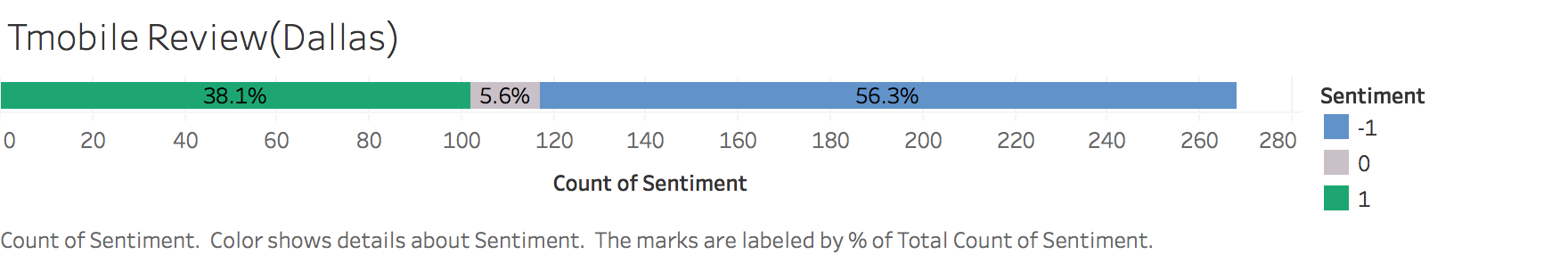
Fig 2.11 Correlation between demographics information and final store rank.

### 2.5.5 Competitor telecommunication companies

The final insight is the comparison among 4 famous telecommunication companies. From Fig. 2.12, we can find AT&T has the most negative review from its customers while sprint has the least.







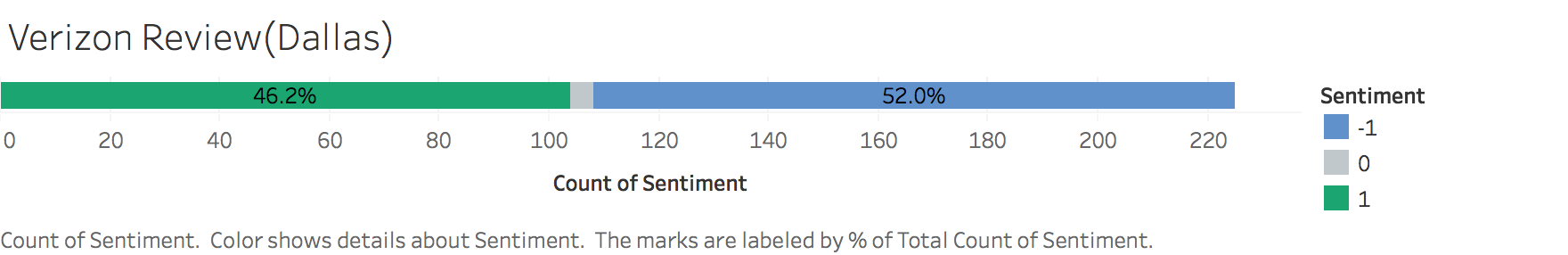


Fig. 2.12 Sentiment distribution comparison among 4 telecom companies.

# Reference:

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

[2] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.