Churn Prediction and Analysis

Neti Sheth
Zhenlong Wu
Soumya Roy
Raghav Dasari

Table of Contents

Executive Summary 1

Problem Significance 2

Prior Literature 2

Data Source and Preparation 3

Text Analytics Workflow 4

Exploratory data analysis 4

- Tweets counts for each carrier and each group 5
- Most frequent words in each carrier 5
- Counts for each carrier mentioned daily 6
- Sentiment polarity distribution for each carrier 7
- Overall sentiment polarity of each carrier 7
- Positive vs. Negative tweets for each carrier 8

Choice and rationale for text analytic methods, and results 9

- Churn Detection Analysis 9
  - Manually labeling 9
  - Train a machine learning model 9
- Identifying Reasons and Suggestions 10
  - Rule-based algorithm to detect user's churn direction 10
  - Rule-based algorithm to detect the reason of churning 11
  - Summarize the reason 11

key insights (must be useful and actionable) 12

references 13

Executive Summary

As every customer takes advantage of the network providers, same way the network providers. But this strategy worked well when there were just a couple of network providers and customers stayed with the same provider for a lifetime, once the technology changes started happening rapidly and as more new providers started in the market with more new plans to attract the customers, that is the time where customers started looking for a churn, to retain their customers every network provider started responding to churn in different ways like giving discounts, additional minutes, unlimited data plan

Taking into consideration that the market is being saturated and revenue from new subscriptions is increasingly deteriorating, mobile carriers tend to focus on customer service and high levels of customer satisfaction in order to retain customers and maintain a low churn rate. In this context, it is a matter of critical importance to be able to measure the overall customer satisfaction level, by explicitly or implicitly mining the public opinion towards this end.

online social media can be exploited as a proxy to infer customer satisfaction through the utilization of automated, machine-learning-based sentiment analysis techniques. Our analysis focuses on the three leading mobile broadband carriers AT&T, Verizon & T-Mobile, by analyzing tweets fetched during a 30-day period within October 2019, to assess relative customer satisfaction degrees.

Problem Significance

Taking into consideration that the market is being saturated and revenue from new subscriptions is increasingly deteriorating, mobile carriers tend to focus on customer service and high levels of customer satisfaction, in order to retain registered customers and maintain a low churn rate. In this paper, we try to examine if mobile wireless carriers can benefit from performing sentiment analysis through social media networks in order to enhance and improve customer service, which will lead to increased customer satisfaction, thus keeping a low churn rate.

Prior Literature

A lot of work has been done in the field of opinion mining for sentiment analysis for well over a decade now. Twitter is different from other forms of raw data which are used for sentiment analysis as sentiments are conveyed in one or two-sentence blurbs rather than paragraphs. There are various approaches to mining twitter data.

Twitter is much more informal and less consistent in terms of language. Users cover a wide array of topics which interest them and use many symbols such as emoticons to express their views on many aspects of their life. When using human-generated status updates, the

sentiment is not always obvious; many tweets are ambiguous and can use humor to maximize the opinion to other human readers but deflect the opinion to a machine learning algorithm [2]. Another consideration when using a dataset generated from Twitter is that a considerably large amount of tweets which convey no sentiment such as linking to a news article, which can lead to difficulties in data gathering, training and testing. Sentiment analysis provides a means of tracking opinions and attitudes on the web and determines if they are positively or negatively received by the public.

Performing sentiment analysis and opinion mining through Twitter is a research subject that has drawn the interest of many research teams throughout the world during the past few years. The challenge to accurately predict social mood based on text mined from Twitter, still remains a big challenge and is currently being explored in various market and academic segments.

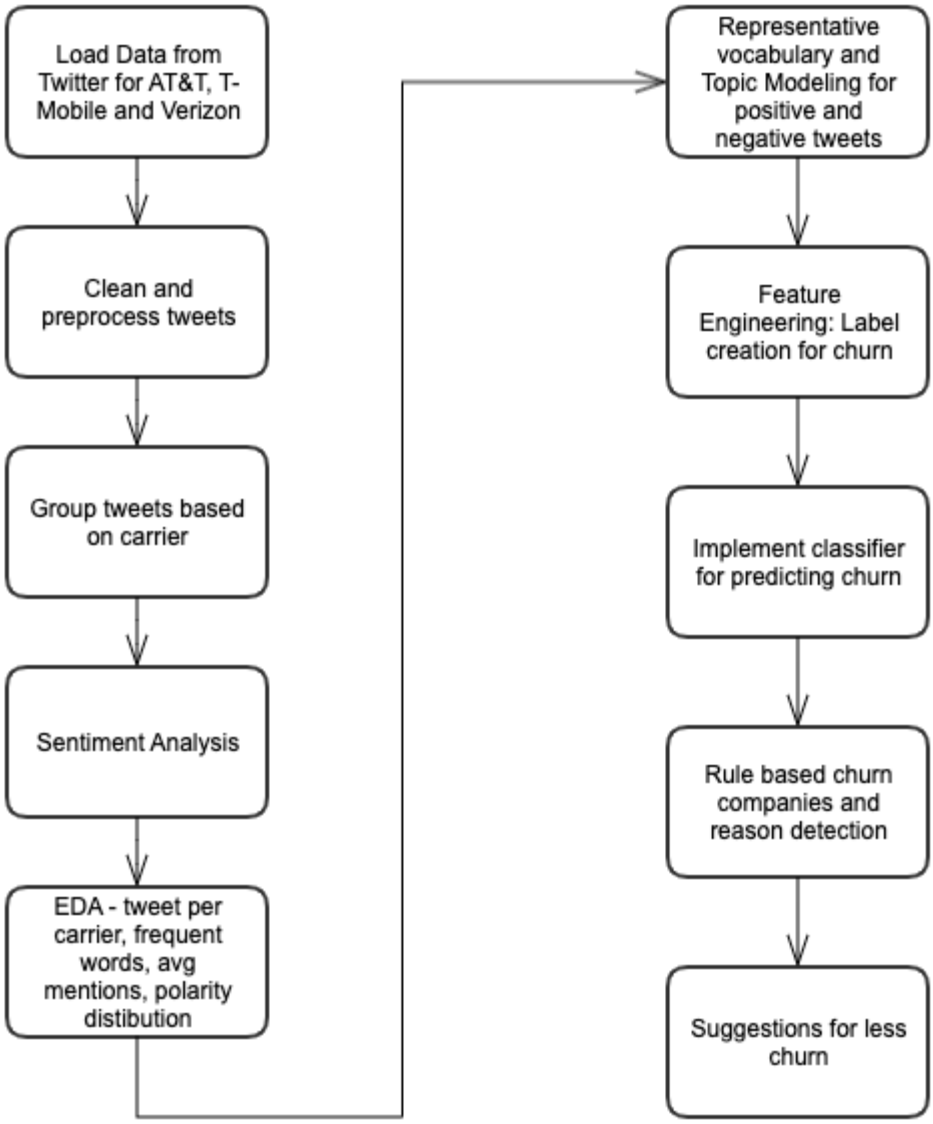
Data Source and Preparation

During the time between October 9, 2019, and October 23, 2019, we collected and analyzed a collection of over 213859 tweets using Twitter's Streaming API. The method of data collection centered on gathering tweets directly relating to the three major mobile carriers, including AT&T, TMobile, and Verizon. Twitter's official streaming API was used to search for keywords on "VerizonSupport," "VZWSupport," "ATT," "ATTHelp," "TMobile" and "TMobileHelp".

The corresponding dataset included a total number of 82395, 47746 and 31842 tweets respectively for AT&T, Verizon and T-Mobile, which were then categorized into a class of 2 and 3 mobile carriers such as [ATT, T-Mobile],[T-Mobile, Verizon],[ATT, Verizon] and[ATT, T-Mobile, Verizon].

The data was then cleaned and preprocessed for analysis. The process involved tokenizing text into words, removing stop words, links, hashtags, handles mentions and words with less than three characters.

Text Analytics Workflow



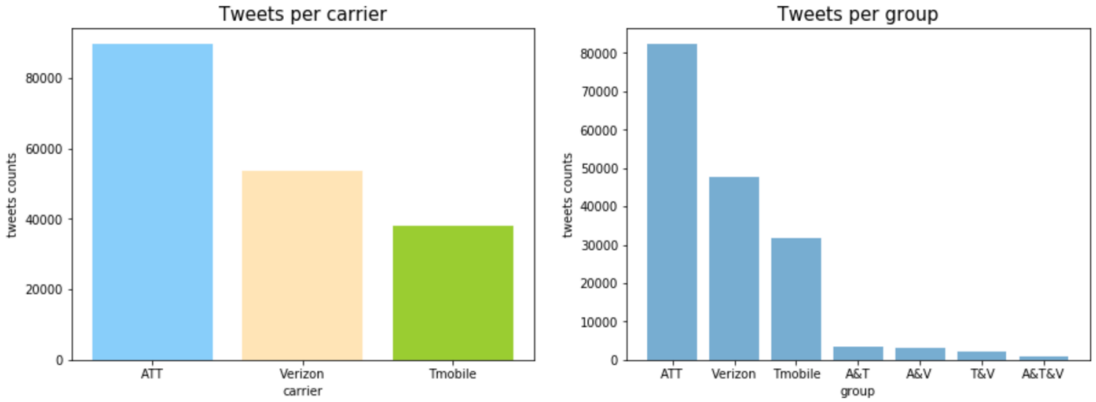
Exploratory data analysis

In the exploratory data analysis, we used all 213859 tweets from October 9, 2019, to October 23, 2019, to do overall data exploration. And these are several questions that can help us quickly have a preliminary understanding of the database:

- How many Tweets are there for the three major carriers?
- What are the most frequent words in each carrier?
- How many times have the three major carriers been mentioned over time?
- Sentiment Analysis:

    - What is sentiment polarity distribution of each carrier?
    - What is overall sentiment polarity of each carrier?
    - Positive tweets vs. Negative tweets of each carrier.
- Representative vocabulary in both Positive Tweets and Negative Tweets.

1. Tweets counts for each carrier and each group



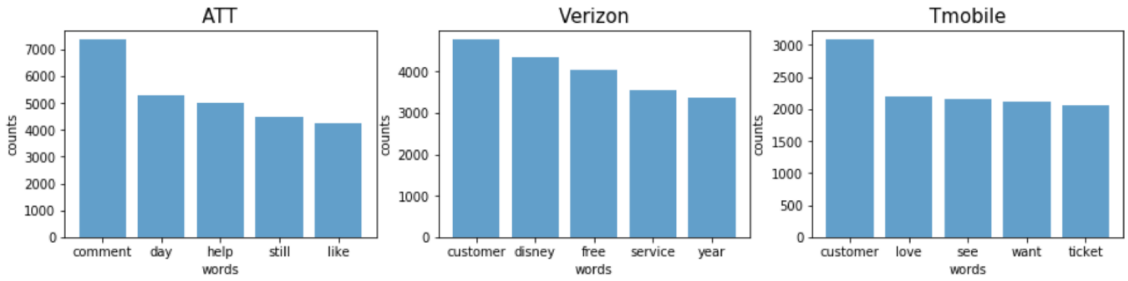
ATT is the carrier that has been mentioned the most times by tweets during this period. Followed by Verizon, and then TMobile. Since each tweet generally does not always mention only one carrier, we defined 7 groups based on different carriers mentioned.

- att_df: tweets that only mention ATT
- ver_df: tweets that only mention Verizon
- tmo_df: tweets that only mention TMobile
- tmo_att_df: tweets that mention both TMobile and ATT
- att_ver_df: tweets that mention both ATT and Verizon
- tmo_ver_df: tweets that mention both TMobile and Verizon
- tmo_att_ver_df: tweets that mention all TMobile, ATT, and Verizon

And only a small portion of the tweets that mention or compare multiple carriers.

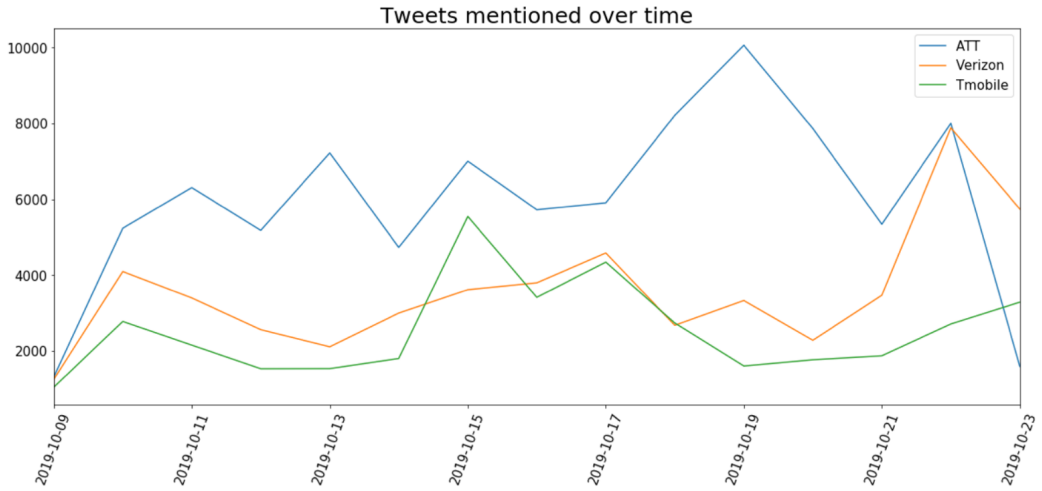
2. Most frequent words in each carrier

```
ATT : comment, day, help, still, like
Verizon : customer, disney, free, service, year
Tmobile : customer, love, see, want, ticket
```



According to the BOW model, we got the top 5 common words of each carrier. For Verizon, users like to mention Disney, free service, and yearly plans in their tweets. Similarly, for TMobile and ATT, users are talking about comments, tickets, and customers.

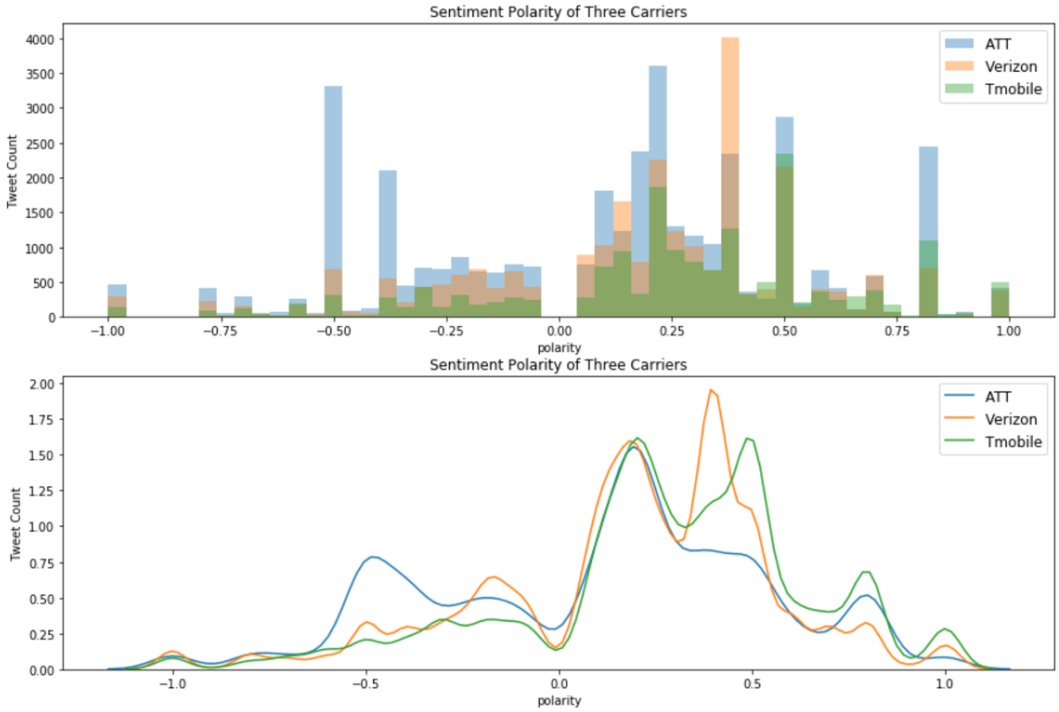
3. Counts for each carrier mentioned daily



During the time between 2019-10-09 and 2019-10-23, we can observe how the number of tweets mentioning each carrier varies. Some interesting observations are:

- On 2019-10-15 Tuesday, the number of tweets for TMobile is maximum, the possible reason can be t-mobile Tuesday promotional offers.
- On 2019-10-19, a lot of ATT users were facing network issues. That can be the reason behind the sudden rise in the number of tweets.
- On 2019-10-22, Verizon tweets were the maximum. They announced a free year of Disney+. The users might be comparing it with ATT offers.

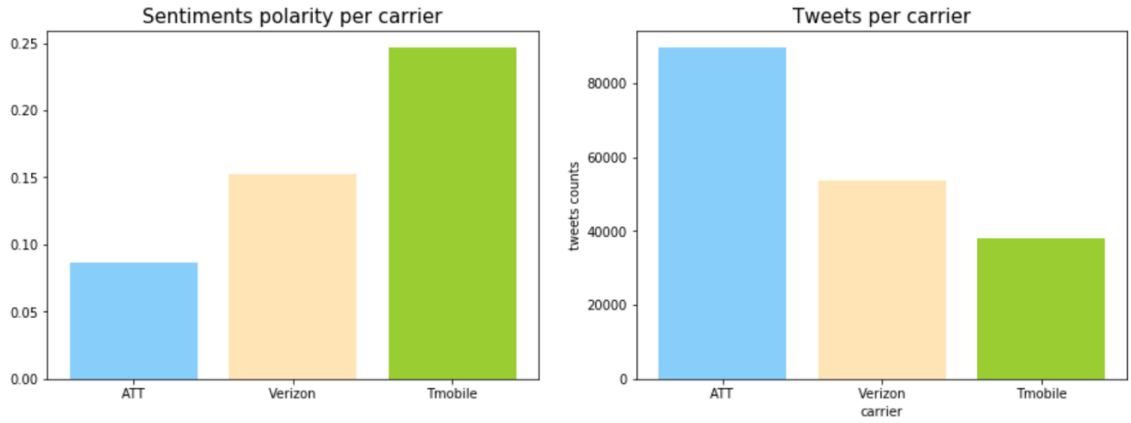
4. Sentiment polarity distribution for each carrier



We can see how the sentiment varies for each tweet mentioning different carriers. It helps us to understand how many tweets were positive and negative. For example, ATT had maximum number of tweets with -0.5 polarity (negative). Most of the tweets for each carrier were positive.

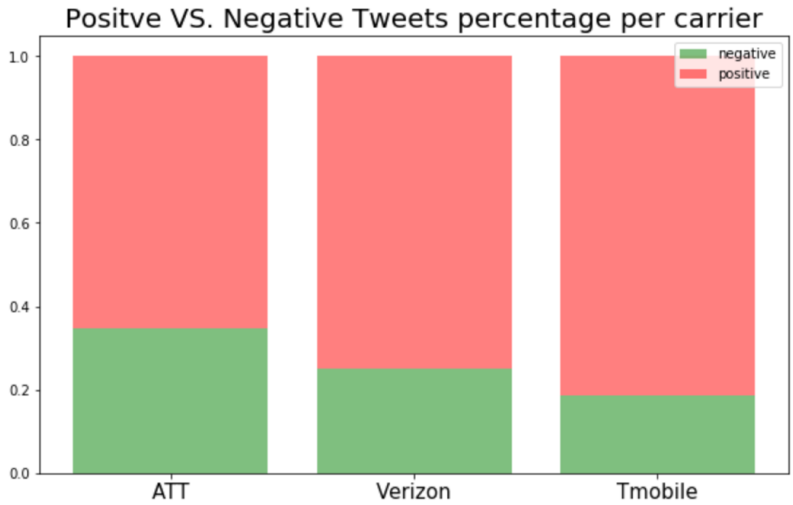
5. Overall sentiment polarity of each carrier

	carrier	sentiment
0	ATT	0.086422
1	Verizon	0.152118
2	Tmobile	0.246736



As we can see from the overall sentiments graph, ATT gets the lowest sentiment score, and TMobile gets the highest. An interesting observation is that carrier with more tweets mentioned has less sentiment polarity score.

6. Positive vs. Negative tweets for each carrier



From the graph, we can notice that in all three carriers, as the amount of tweets increases, the proportion of negative numbers also increases. Maybe users like to complain more than praise on tweets.

7. Representative words

From the figure1 in the appendix, we used topic modeling to get 5 topics from each carrier's positive tweets and negative tweets. And we found that ATT positive topics are about wifi and service, ATT negative topics are about internet, ticket and game. Verizon positive topics are about iphone and data, Verizon negative topics are about samsung, bill and outage. TMobile positive topics are about family, speed and offer, TMobile negative topics are about coverage and service.

These topic words help us find the reason for user churning in subsequent feature engineering.

Choice and rationale for text analytic methods, and results

1. Churn Detection Analysis

a. Manually labeling

To build any classification model, may it be the simplest one or may it be the most complex ensemble ones, the first thing that we need is the ground truth. Unless we have labelled truth data in hand, it's almost impossible to train machine. Hence, we started labelling our corpus of tweets manually. This task was divided among all the team members and each one of us had to read the tweets and label it as 'churn' or 'not churn' manually by understanding the context and the words used in the tweets. Once this was done, we shuffled the distribution and every member of the team had to validate the labeling done by others in the team. This way we were able to manually label around 4000 tweets and validate them with each other's views on the same tweets within a week. However, this wasn't enough for the large corpus that we had. So, we decided to merge this approach to an approach where we are training a model in an iterative process. This has been discussed in detail in the following section.

b. Train a machine learning model

Here,we used the above manually labelled tweets as an input to a machine learning model to predict the labels for a set of unseen tweets. The steps that we followed is as follow:

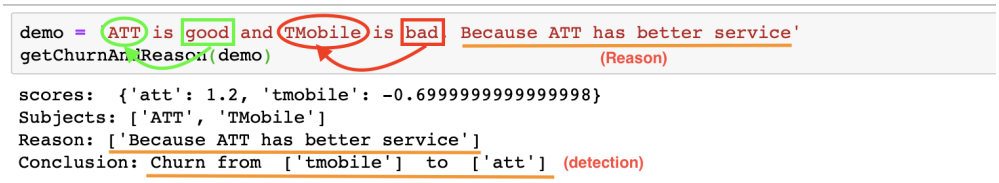
- ❏ Split the manually labeled input dataset of 4000 to train and test sets of 80% and 20% respectively.
- ❏ Use the train data set to train the model and test it on the test dataset. Tune the model hyperparameters and repeat. Test it on unseen data outside of the test set previously provided.
- ❏ Validate the results manually, refine the input and predict the next 500 labels with the help of the model. Now add this new labeled set to the training dataset and predict another set of 500 tweets with the updated training dataset.
- ❏ In this way we were able to label another 2000 tweets and validate them manually.
- ❏ We used two models,namely Naive Bayes and XGBoost, however, XGBoost proved to outperform the others with an accuracy of 73%.

- ❑ Now, we had 6000 labelled tweets. However, it still wasn't enough. Hence, we decided to develop a rule based algorithm on top of this approach and identify the labels, their reasons to churn and provide suggestions in a heuristic way (Described in the following section).

2. Identifying Reasons and Suggestions

According to the previous classification model, we can successfully classify each new tweet text into churn or not-churn groups. However,if the tweet was classified to churn group, the classification model cannot continue to detect the reason and the new carriers that the user wishes to switch. Therefore, we started to design a new tool to implement understanding of each tweet from the user. We try to identify the reason of each tweet in churn group, and also detect that the user wants to churn to which new carrier from which old carrier.

a. Rule-based algorithm to detect user's churn direction



There is a very easy-understanding example 'ATT is good and TMobile is bad. Because ATT has better service'. The reader can easily figure out that this user might want to churn from TMobile to ATT. The reason is that ATT has better service. Then, how does the machine do the semantic analysis and come to this conclusion?

Our idea is to build a Rule-based detection algorithm based on sentiment analysis. Firstly, we only identify three subjects, 'ATT', 'TMobile' and 'Verizon' in each tweet. Then, we try to disassemble the original text and find out a match between words or sentences and the subject. In the above example, we need to make the machine understand 'good' is describing 'ATT' and 'bad' is describing 'TMobile'. In reality, people's language has many expressions, and this simple and neat language format is rarely encountered on Twitter. Therefore, we have summarized some semantic rules by reading a lot of tweets. Finally, we used these rules to segment the text and add some scores to each subject. We assume that users always want to churn from a low-scoring subject to a high-scoring subject.

Here are some of our rules: "switch from… to…", which is a typical sentence that clearly states where the user wants to churn from which carrier. So the carrier in "from" group will decrease score, and the score of the carrier in "to" group will increase. Synonyms of switch, such as transfer, try to, free from and so

on, all follow the same rule. The carrier appearing after “like” or “hate” will also increase or decrease the score, since it implies that the user has a potential intention to churn. Similarly, people might use a lot of positive vocabulary to describe the new carrier they want to go, or use many negative vocabulary to describe the current carrier to imply the intention of churn. Therefore, we will eventually add the sentiment polarity value to the score for each subject. The conjunction words, such as “and” and “but ”, are very crucial signs that help us segment the text and find the words that describes the certain subject.

The challenge of this algorithm is to identify more semantic rules and define the weights of score of each rule. Our current approach is to artificially collect and add rules, then adjust weights manually.

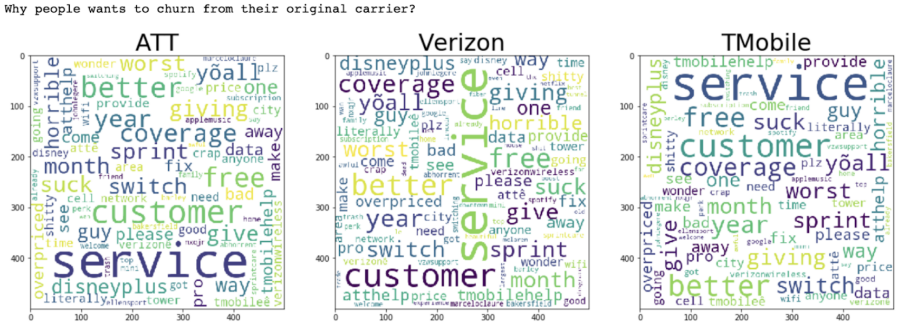
b. Rule-based algorithm to detect the reason of churning

For the churning reason, there are some indicators that users often mention. Then we try to figure out and collect these indicators, such as “service”, “coverage”, “data”, “wifi”, “price”, “family” and so on. If one of these words appear in a sentence, we will roughly attribute it to the reason. And comparative adjectives, such as ‘smaller’, ‘cheaper’ and so on, also imply the comparative relationship between the two carriers. These comparative adjectives are also the indicators of the churning reason detection.

The challenge of this algorithm is to find out a method that makes machine to find indicators automatically and improve the scalability.

c. Summarize the reason

When we identify the carrier that the user is unsatisfied with and extract the reasons for dissatisfaction by rule-based algorithm for each tweet, we build out three corpuses of unsatisfied reasons for each carrier. Through the analysis of these reasons corpus, we have obtained some conclusions. We can further summarize the problems of each carrier.



Key Insights

When we identify the carrier that the user is unsatisfied with and extract the reasons for dissatisfaction by rule-based algorithm for each tweet, we build out three corpuses of unsatisfied reasons for each carrier. Through the analysis of these reasons' corpus, we have obtained some conclusions. We can further summarize the problems of each carrier

Based on the above research we have done, there are several reasons why the churn happening in each network provider, these reports helps individually to the concerns network providers to have a look on each issue that we reported based on the tweets to take more actions to keep their customers stay with them to reduce the churn rate.

References

https://pdfs.semanticscholar.org/f26e/3d935ec705428b292f630d4aae63c24443a5.pdf

https://pdfs.semanticscholar.org/f700/9270213ece551e14e52760598a0c75429c29.pdf

https://monkeylearn.com/blog/analyzing-customer-support-interactions-on-twitter-with-machine-learning/

To see the output please execute the scripts in the below order:

1. Corpus Building (step1_Corpus Build.ipynb)
2. Project Exploration (step2_Project Exploration.ipynb)
3. Churn Labeling (step3_ChurnLabeling.ipynb)
4. Feature Engineering (step4_Feature Engineering and identifying reasons.ipynb)
5. Dashboard Creation (step5_DashboardCreation.ipynb)