REGIONAL FALL CASE COMPETITION
FIRST ROUND

THE UNIVERSITY OF TEXAS AT DALLAS

Harsh Agrawal
Linh Phuong Nguyen
Achintya Sen
Varun Sriram

THE UNIVERSITY OF TEXAS AT DALLAS
Harsh Agrawal, Linh Phuong Nguyen, Achintya Sen, Varun Sriram

Objective

Our goal for this project is to collect data on AT&T's customer sentiment from different social media channels, sort them based on retail zones/stores, rank those zones/stores' ratings and eventually help AT&T find the correlation between social media and retail store performance.

Here are our approaches to achieve that goal:

I. Collecting Social Media Data

In order to have an updated and relevant idea of store ratings, we are going to only consider data for the past two years. Based on the distinct functions and uses of different social media platforms, we identified these four platform to be the most relevant and plausible sources of data when it comes to analyzing customer sentiment and satisfaction. Below are our approaches on how to collect data from each channels or feeds:

1. **Yelp**
   Yelp has a large volume of database of stores and geo-location. Extracting the data based on API services is possible by signing legal contracts with the company. The extract can be taken in JSON format, parsed through python *json* library and stored in a sqlite database or .csv format.
   *Limitation: The data scraping is limited through Yelp's Terms and Condition which needs to be resolved through contracts.

2. **Google Place Reviews:**
   The Google Places API Web Services provides store ratings through user reviews based on the corresponding place-id. The place-id is a textual unique identifier given to every location in google database. The place-id of AT&T stores can be fetched using Google Places API Web Service which returns all id's based on a search text in JSON format. We plan to parse JSON through python *json* library. Parsing individual comments will help in finding the individual review.
   *Limitation 150,000 free requests per day (after credit card validation)

3. **Twitter & Facebook**
   Facebook feed pertaining to pages, groups, and users can be fetched through facebook's ***graph API***. The majority of Facebook data can be associated with a geo-location, which can be beneficial to our geo filtering.

   Twitter data stored can be similarly extracted from feeds based on keyword searches in a SQLite database through parsing json data. The OAuth API Services can perform this function based on token access and user authentication.
   *Limitation: These social media sites also limit third party apps from scraping data through API's.

II. Data Cleaning and Integration:

- Resolve missing values, outliers, and noises through use of techniques like imputing and other methods if possible.
- Checking whether the data values have some sentiments attached to it. While extracting from various social media platforms we will get all sort of data values ranging from posts about new product or service launching, new store openings to news about AT&T and we need to filter out the required data for performing sentiment analysis.
- Data Scraping from various platforms will result in data in different formats and we need to properly integrate these together.

III. Geo Filtering

- Segment the data to generate smaller datasets pertaining to store locations or retail zones in the Dallas area using k-medoids clustering. This method will allow us to create clusters around the medoids chosen based on the latitude and longitude.

*The number of AT&T stores found on its official website is 82, but we can also divide the data into fewer retail zones as we see fit.

IV. Sentiment Analysis:

Sentiment analysis is a key tool used in marketing analytics. Most of the data is User Generated Content, or UGC. Extracting information from this type of content requires high level data mining techniques such as Information Extract and Natural Language Processing (syntactical analysis). Here are our steps:

- Identify 4-5 entities of interest: AT&T, AT&T customer services, AT&T cellular network, AT&T's deals, etc. based on their frequency in the data. Each entity will represent a product/service.
- Utilize parsers (paid or self-generated) to produce rules i.e. patterns present in the datasets. For example: "[entity 1] [word 1] [word 2] excellent" can be a possible pattern found.
- For each entity chosen, perform clustering analysis on those patterns to identify the groupings of the patterns as well as the classification of the social media posts
- For each entity chosen, assign labels to clusters, potentially:
    - Positive sentiment
    - Negative sentiment
    - Neutral sentiment

  By considering neutral sentiments, we will take into account long reviews with a small percentage of either positive or negative words/phrases.

V. Store Ratings:

- Assign a weight of 2 to the percentage social media posts of good sentiments, a weight of -2 to percentage of negative sentiments and a weight of 1 to percentage of neutral sentiments
- Calculate scores for each entity as product/service level rating scores and sum up those scores as the overall store rating score.

- Proper rescaling of these different scores on a scale of 5 is required to offset the effect of extreme values.

Alternatively, we can utilize other sentiment analysis tools such as Google Cloud Services to assign scores/ratings to the stores.