CS 410: Final Project Progress Report

Brand Sentiment on Twitter using Sentiment Analysis

Team Stanley (Fall 2021)

1. Which tasks have been completed?

I have completed the following tasks so far:

a) Write a program to generate a dataset by collecting the tweets from Twitter API for a given brand over a period of time.
b) Write a program to submit the dataset to Amazon Comprehend for sentiment analysis.

2. Which tasks are pending?

a) Create a sentiment trend graph based on the result of the sentiment analysis for data visualization. (4+ hours)
b) Collect recent brand's news or events. (1+ hours)
c) Analyze the sentiment trend graph, and link any sentiment shift with the recent news or events. (3+ hours)

3. Are you facing any challenges?

There were a few challenges I encountered:

a) Twitter API v2 is not ready for prime time

   Twitter currently supports two API versions: v1.1 and the new v2 in early access. I started off with the new v2 API under the impression that v2 should be superior. Unfortunately, it turned out the v2 API is still in early access and it lacked some of the functionalities I needed. After spending a few hours with the v2 API, I switched

back to use the v1.1 API.

b) Query abilities and rate limiting with Standard Developer Account

My Twitter developer account is standard (instead of premium or academic research), and I encountered several restrictions:

i) For standard developer accounts, Twitter focuses only on relevance and not completeness in search results. This means that some tweets could be missing from search results.
ii) The standard developer account has a limit which only allows querying the tweets in the past 7 days.
iii) Each query request could return up to 100 tweets, but there is a rate limit allowing only 180 requests (with user authentication) in a 15-min window.
iv) There is also a monthly tweet cap usage which allows only 500,000 tweets to be pulled.

All these issues make it difficult to retrieve large amounts of tweets over a long period of time for sentiment analysis for this project. To mitigate these, I decided to focus on performing sentiment analysis on the popular tweets in the past 7 days for this project.

c) Truncated text in tweets

Inspecting the retrieved tweets apparently revealed that many tweets were truncated. It turned out that Twitter originally supported only 140 characters in tweets, and later extended the support to 280 characters for certain languages. The Twitter API returned the tweet in compatibility mode which truncates the text by default. Changing the query to extended mode fixed the issue.

d) Tweets in different languages

Twitter is a global service available in over 200 countries, and the tweets are written by the users all over the world in various languages. To analyze the tweets, this presents a challenge as the tweet might not be written in a language which sentiment analysis would understand. For simplicity, I decided to only analyze the tweets in English for this project.

e) Retweet

After analyzing the tweets, it became apparent that many Twitter users didn’t write their own tweets. Instead, they often favorited or retweeted what others had written.

This presents another challenge as many tweets returned from Twitter had identical text. This is not ideal as each of these tweets with identical text is counted towards our developer account's rate limit and monthly usage cap. Hence, I decided to retrieve only the most popular tweets, instead of all the tweets. Retrieving the most popular tweets would return each tweet once with a retweeted count, and this helps address the concerns around rate limit and monthly usage cap. That said, retrieving only the most popular tweets means we would ignore the long tail of unpopular tweets created by many individuals, and this could affect the overall accuracy of the brand sentiment we would like to determine in this project.

f) Cost of public cloud's NLP service for sentiment analysis

It turns out that it is cost prohibitive to perform sentiment analysis on large volumes of text with public cloud's NLP service. Take Amazon Comprend for example, it charges $0.0001 per unit (up to 10M units) which each unit is 100 characters. Simply analyzing 10,000 tweets each with 200 characters alone would cost $2. There are 500 millions tweets put up on average every day on Twitter. Even if 0.01% of the tweets are relevant to our brand for sentiment analysis, that's 50,000 tweets per day. Hence, instead of performing sentiment analysis on all the relevant tweets over a period of time, I decided to scale back and only analyze the popular relevant tweets, as the total number of popular relevant tweets is several orders of magnitude less than the total number of individual relevant tweets, and the cost would be much more affordable for this project.

4. Plus anything specifically mentioned in the reviews to cover.

N/A