Twitter Bank Analysis

Hedel Torres

009046384

2016-03-25

# Introduction

Classification and sentiment analysis, using Twitter data, will be used to identify client sentiment relating to financial institutions across Canada, specifically TD and its main competitors. An attempt will be made to identify emerging client financial needs, such as new products and services, as well as opportunities for improvements. Analysis will attempt to identify potential markets and gaps.

Analysis will be performed by identifying keywords and metadata found within Tweets. Hadoop, Hive will be used to store, query and perform analytics on the data. Visualization will be made on Tableau, with the repetitive scripting being done on Excel.

# Literature Review

Finding clear instructions on how to extract Twitter data for a Python beginner was difficult. I engaged a past Ryerson colleague, [Rakesh Dote](mailto:rakesh.dhote@gmail.com?subject=Python%20Twitter%20Code), to assist me with the code.

**Tweets - Field Guide**

<https://dev.twitter.com/overview/api/tweets>

The Twitter Developers Documentation provided an extensive list of the data attributes, Field Names, Data Type and Description. I used this document to identify the key attributes that I would require for my analysis.

**Thornydev Blogspot**

<http://thornydev.blogspot.ca/2013/07/querying-json-records-via-hive.html>

The ThornyDev site is a great source that provides information on how to use JSON data in a Hive table.

The site provides some details on how to put all the data in a Tweet into a table with their associated headings.

**Stackoverflow – How to export a Hive table into a CSV file?**

<http://stackoverflow.com/questions/17086642/how-to-export-a-hive-table-into-a-csv-file>

The Stackoverflow site is a great source for additional information. The site provided details on extracting a Hive table into a .csv format. I found that the .csv format did not correctly separate the columns well. I used the same instructions but modified it to .txt instead of .csv.

**AblazeGroupBI - Tableau Training Create A Word Cloud in Tableau**

<https://www.youtube.com/watch?v=x4ByP-wzIO0>

The AblazeGroupBI site provides easy to follow instructions on how to import your data into Tableau and create a Word Cloud. The Word Cloud is a visual proportional representation of the words in a Tweet text. The size of the font is based on the number of times the word is found amongst all the Tweets.

# Dataset

I will be using Twitter feeds which includes the Twitter text and all its associated metadata. A Tweet consists of 82 attributes along with a number of sub-attributes. I will not be analyzing the sub-attributes as they have no bearing on the objective of the project, to identify client sentiment.

Of the 82 attributes, I have identified 26 attributes that I will be focusing on for my analysis. These include the actual text of the Tweet, the number of times the Tweet was retweeted, the number of “Likes” a Tweet receives and the geographic location. During my analysis of the 128 files of data, I identified a large number of non-bank client sentiment values, which include reference to contests, employment postings and foreign tweets. I will use a number of the meta data attributes to filter out the irrelevant Tweets.

A Python script is used to pull the Twitter feeds from the Twitter social networking website. Twitter makes available seven day’s worth of Tweets pulled by mentions/hashtags. I have collected Tweets weekly for the period starting January 26, 2016 and ending March 20, 2016.

I have attached below an Excel spreadsheet with a Data Catalogue (Data Dictionary) and an inventory of my source data files.

The below attached file contains the following,

* Definitions – a glossary
* Data Catalogue – a list of the data attributes, the data type, description and comments
* File Inventory – a list of the files and their respective naming convention, as well as which subject they belong to. I have broken down the data into their respective financial institutions and a single common “banks” group
* Sample betterbank-1-27-16 – a sample file set of rows of the file for reference
* User, Tweets, Entities, Places and Entities in Objects – Twitter data dictionary with additional information and examples.



|  |  |
| --- | --- |
| **Definitions** |  |
| Term | Definition |
| Hashtag | \*(On social media sites such as Twitter) a word or phrase preceded by a hash or pound sign (#) and used to identify messages on a specific topic. |
| Like | Used to identify agreement or appreciation of the Tweet. Users may also use Like to provide support of the comment. |
| Mentions | Is a method used to associate a specific word or phrase to a Tweet preceded by the “at” sign (@). |
| Retweet | The reposting or forwarding of a specific Tweet. |
| Tweet | \*A posting made on the social media website Twitter. A textual or media communication out to the Twitter user community. This analysis will only include textual communication as the media communication is out of scope due to the complexities involved in dealing with media (URLs/video/images/etc.) data. |
| Tweeter | A user who posts Tweets on the social media website Twitter. |
| User | A person that posts to the social media website Twitter. |
| User handle | The screenname (nickname) a user will use to associate their Twitter account with. |
| \*Source: Google Definitions | |

# Approach

# 

## Step 1: Pull Twitter feeds

Pull the Twitter feeds data using the Python script in Apple MAC OS Terminal. The requests included key words that identified each bank. The following link is the Python script. <https://github.com/HappyFeet75x/TwitterBankAnalysis/blob/master/Twitter_Pull>

## Step 2: Create Hive Tables

I will be using Excel and its Concatenate functions to repeat the code for each of the 128 associated files. I will combine it using the Inventory List, Github link below. The Excel Concatenate code can be found in the “Create Table” tab in the following attached Excel document.



Code for importing files into Hadoop HFS and importing data into Hive Table

<https://github.com/HappyFeet75x/TwitterBankAnalysis/blob/master/ImportFiles2>

Code for creating Labelled Hive Tables

<https://github.com/HappyFeet75x/TwitterBankAnalysis/blob/master/CreateHiveTablesv3>

## Step 3: Feasibility Validation of Word Count and Bank Comparison

Perform a dry run of the Word Count and Bank Comparison scenario as well as identify ways to reduce the number of non-bank client Tweets by running through the process with one file and perform a follow up validation with a consolidated Hive Table as described in Step 4.

## Step 4: Consolidate Hive Tables

Since institutions can be known by different names, the tables were combined into their associated subject, as follows,

* Banks
* TD Bank
* Scotiabank
* Bank of Montreal
* Tangerine
* CIBC
* PC Financial
* Royal Bank

Code for combining all tables belonging to the same bank topic.

<https://github.com/HappyFeet75x/TwitterBankAnalysis/blob/master/CombineTablesv4>

## Step 5: Clean Data

I identified the most frequent users and analyzed their Tweet and their user profile to validate whether their Tweet should be considered a Client Sentiment.

I identified commonalities that would be associated with the non-client users to identify users that provided true client-sentiment

The best method to filter out users that were not clients was by using keywords.

The following key words were identified as not being associated with clients. Even though a user could have such phrases in their screen name, the likelihood would be low and considering that we are looking for overall user sentiment, missing a user would have minimal impact.

* + bank
  + bmo
  + cibc
  + deals
  + employ
  + hotel
  + job
  + news
  + ontario
  + pcf
  + rbc
  + scotia
  + tangerine
  + td
  + tmj
  + invest
  + ticker
  + rating
  + market
  + financ
* watchlist
* mosaicguys
* money
* consumerfeed
* mktg
* intern
* expert
* analyst

Code for identifying most frequent users

<https://github.com/HappyFeet75x/TwitterBankAnalysis/blob/master/IdentifyMostActiveUsers>

Code for creating tables with the non-client users removed

<https://github.com/HappyFeet75x/TwitterBankAnalysis/blob/master/Table-NonClientRemoved>

## Step 6: Perform Word Count

Perform Word Count to break out the Tweets into separate words and provides a count of word frequency for each subject able.

Code for identifying the word count

<https://github.com/HappyFeet75x/TwitterBankAnalysis/blob/master/CreateWordCountTable>

## Step 7: Prepare Lexicon for use

Lexicon needed to be formatted for comparison. The Lexicon format was massaged to fit into a CSV format. A column that merged the Type (Strong/Weak) and Polarity (Positive/Negative) was added.

<https://github.com/HappyFeet75x/TwitterBankAnalysis/blob/master/ImportCreateLexiconTable>

## Step 8: Compare Word Count against the Lexicon

Perform a Join of the subject table with the Lexicon, using the word as the key.

<https://github.com/HappyFeet75x/TwitterBankAnalysis/blob/master/JoinLexiconBankv2>

## Step 9: Group counts by Type-Polarity subgroup

Group counts for each subject table by Type-Polarity to identify average sentiment by subject.

<https://github.com/HappyFeet75x/TwitterBankAnalysis/blob/master/AnalysisSum>

## Step 10: Extract to File

Once analysis is complete, send data to file to use in visualization software.

<https://github.com/HappyFeet75x/TwitterBankAnalysis/blob/master/ExtractCSV>

<https://github.com/HappyFeet75x/TwitterBankAnalysis/blob/master/AnalysisSum>

## Step 11: Identify Average Sentiment

In Excel, identify the average sentiment for each subject and combine all the subjects into one table for comparison.



## Step 12: Create Visualization

In Tableau, create Word Cloud for each subject and create a chart that compares the average sentiment of the different banks.

# Results

The word count sentiment analysis found that RBC had more positive Tweets of all the banks, with CIBC having a more negative presence in Twitter during the two-month period of January 19, 2016 and March 20, 2016.

|  |  |  |
| --- | --- | --- |
| **Bank** | **Sentiment Word Count** | **Rating** |
| RBC | 9748 | 3.88 |
| Tangerine | 6982 | 3.76 |
| Scotiabank | 29592 | 3.71 |
| PCF | 2448 | 3.70 |
| TD | 108040 | 3.59 |
| BMO | 50512 | 3.57 |
| CIBC | 17853 | 3.35 |



The Rating is a weighted average of the sentiment. To normalize the different sentiment, I assigned a value of 5 to “Strong Positive” and a 1 to “Strong Negative” as outlined below and I found the average sentiment by multiplying the sentiment value by its proportion.

|  |  |
| --- | --- |
| **Type-Polarity** | **Value** |
| Strong Positive | 5 |
| Weak Positive | 4 |
| Neutral | 3 |
| Weak Negative | 2 |
| Strong Negative | 1 |

I started with this hypothesis, and ran a number of scenarios using made-up (fake) and true data until I felt that the hypothesis was justified.

The attached file in Step 11 above provides a breakdown of my analysis.

Along with the average sentiment, I presented a Word Cloud to identify the most common sentiment words by subject, which also represented their frequency of use by size.



The following attachment is a slide presentation that summarizes the process, activities associated with each process, as well as all the results and final analysis.



# Conclusions

RBC received the highest percentage of positive tweets of all seven banks, followed by Tangerine. The bank that received the lowest percentage of positive tweets was CIBC.

Phrases such as “great”, “like”, “love”, “best” were common in tweets referring to RBC, with a few mentions of negative words such as “worse”, “weakness”, “break” and warning”.

CIBC on the other hand, had a high number of neutral words compared to positive words. The few positive words included, “great”, “support”, “proud”, “thank”, “love”. Even though CIBC did not have a high count on any specific negative words, there was a large variety of negative words associated with CIBC, such as “volatile”, “crime”, “beware”, “fraud”, “insincere”, “weakness”, “rude”, “worst”, “bad”, “blame”.

With some filtering and massaging, the analysis was very useful in identifying what Twitter users think of an organization and its products, and allows the organization to address them. Using the Twitter feeds directly with no analysis, will provide incorrect results as a result of the massive amounts of corporate media Tweets, which drown out the client sentiments.