YORK UNIVERSITY
CSDA1050F18S1

15 July 2019

OVERVIEW

1. Project Background and Description

This is for the fulfillment of the York University's Advanced Analytics Course Capstone Project. The aim of this project is to uncover insights from the social media space through programmatic means

2. Project Scope

Here are the boundaries of the project:

- *Social media channel: Twitter (to include Facebook if time permits)*
- *Social media scope: Major Canadian Financial Institutions (FI) like BMO, CIBC, RBC, Scotiabank, Simplii, Tangerine, TD*
- *Comparison of the following insights across the above FIs: Sentiment Analysis (polarity and categorical); Word Cloud (conversation drivers); Key-word dendrogram (blend of sentiment and conversation drivers); Network Analysis (demographics and product segmentation). Paraphrases of these insights are given in the "Research Questions" section below*

3. Research Questions

Here are the research questions for this project:

- *Which bank has the most favourable / unfavourable trending opinion?*
- *What are the current financial products being discussed?*
- *Which bank has the most favourable / unfavourable trending opinion?*
- *What are the current emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) towards each bank?*
- *What are the current sentiments towards trending financial product segments / categories (and the general network of terms being tweeted)?*

4. Literature Review

Here is a summary of the literatures that are relevant to this project

- https://ai.googleblog.com/2006/08/all-our-n-gram-are-belong-to-you.html extracted on 14 July 2019
    - Example of N-grams analysis, which determines the likelihood of next word or character given N previous terms, by Google
- "Detecting Sarcasm in Text: An Obvious Solution to a Trivial Problem", http://cs229.stanford.edu/proj2015/044_report.pdf extracted on 14 July 2019
    - Unlike humans, it is very difficult for machines to detect, let alone understand, sarcasm. This paper attempts "to design a machine learning algorithm for sarcasm detection in text"
- https://github.com/mjockers/syuzhet extracted on 14 July 2019
    - Documentation on the R package, syuzhet, for sentiment analysis. It is based on concepts of "fabula" and "syuzhet" where the former is "the raw material of a story (chronology)" and the former is "the way a story is organized (technique of the narrative)". As such, the syuzhet R package helps reveals "emotional shifts that serve as proxies for the narrative movement between conflict and conflict resolution"
- https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html extracted on 14 July 2019
    - This article describes mining of online opinions, and then determine whether the opinions are positive or negative (opinion polarity)
- http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.html extracted on 14 July 2019
    - This article helps take sentiment analysis to the next step. That is, besides sentiment polarity (positive or negative), we now have the ability to categorise sentiments into the eight basic human emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust)
- https://igraph.org/r/doc/aaa-igraph-package.html extracted on 7 July 2019 o Documentation on how to implement "follower graph" network analysis using the R igraph package

5. Data Sources

The data source will be via Twitter's Application Programming Interfaces (APIs), which can be obtained via https://dev.twitter.com/.

6. Methods

The following outlines the methodology used for this project

- Data ingestion will be via Twitter's Standard (free developer's access) API
- Data storage will be on local PC for testing and eventually on MongoDB
- Data analytics will be via the R programming language
- The primary R package used will be "twitteR"
- Other methodologies used will be Natural Language Processing and network analysis plus those described in the literature review

7. Limitations and constraints

The following outlines the limitations and constraints of this project

- Twitter's Standard (free developer's access) API provides up to seven days of tweets
- The analysis will be limited to the following Financial Institutions in Canada: BMO, CIBC, RBC, Scotiabank, TD
- Insights will be limited to the following: Sentiment Analysis (polarity and categorical)
- Word Cloud (conversation drivers)
- Key-word dendrogram (blend of sentiment and conversation drivers)
- Network Analysis (demographics and product segmentation)

8. Timelines

The following timelines will be used for tracking the progress of the project

- Week 1 (02-08 July 2019): Prepare project proposal
- Week 2 (09-15 July 2019)
    - Data ingestion, assembling methodology for Exploratory Data Analysis
    - ETL into data warehouse (local PC for testing; experiment with MongoDB)
    - Exploring data and documenting issues/limitations/needs understanding needs/limitations regarding research question and scope (might need to adjust question, scope, data, etc.)
    - Submit project proposal
- Week 3 (16-22 July 2019): Sprint #1 submission
    - Post exploratory data analysis to Github
    - Collect/Augment/Refine project according to Sprint #1 findings
    - Develop analytical methodologies to help derive and visualise insights
- Week 4 (23-29 July 2019)
    - Complete data collection
    - Finalise analytic methodology
- Week 5 (30 July to 05 August 2019): Sprint #2 submission
    - Post work progress to Github
    - Includes R codes, summary report, and plan for analysis
- Week 6 (06-12 August 2019)
    - Final sprint (finalise codes, methodology, insights, etc.)
    - Start writing final project report
- Week 7 (13-19 August 2019)
    - Complete R codes with comprehensive comments
    - Finalise project report
    - Upload documentation and R codes to Github
- Week 8 (20-26 August 2019)
    - Final project report submission
    - Project presentation

PREPARED BY

Christopher Tan

Student ID: 303428

School: York University, Toronto, Canada