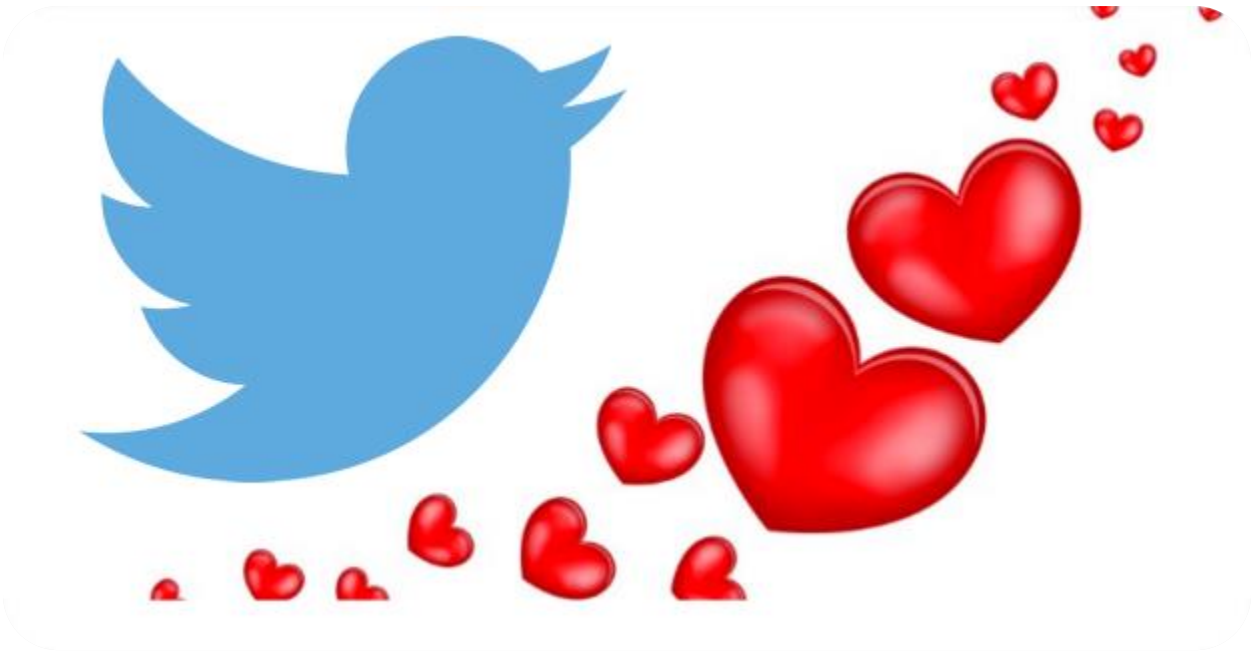

Sentiment Analysis of Valentine's Day Tweets

Veslava Ovendale, Arash Naderpour, Kyle Witt

March 2017

TABLE OF CONTENTS

INTRODUCTION

Valentine's Day is an annual holiday held on February 14th to celebrate romantic love, friendship and admiration (What is Valentine's Day?, 2017). Every year on this day people use social media to exchange messages of love and affection to partners, family and friends. In 2016, NetBase Solutions, California based developer of natural language processing technology, performed sentiment analysis of Valentine-related tweets (Chauhan, 2016). Their analysis showed that out of a total of 9 million tweets that mentioned Valentine's Day, the overall sentiment was overwhelmingly positive with approximately 4 million positive tweets and only 188,000 negative tweets. Also, one of the key findings was when weighed against the topics of Chocolate, Travel, and Engagement, Single Life was the most active topic (Chauhan, 2016).

Curious about how users interact with Twitter during Valentine's Day and thinking about the application of our research, we connected with two professionals in the industry. The first professional is a data scientist working for an online retailer. The second professional is a financial analyst helping various businesses while working at the Seattle-based consultancy firm. We asked each of these professionals what Twitter information regarding Valentine's Day would be most relevant to them. The data scientist noted, "We would totally care about how many people supported which retailer and also their location like country or city. Also, we would care about what they ordered." In his turn, the financial analyst stated, "If I was a card company, I would want to know at what time, how many times people are tweeting sarcastic versus romantic tweets so as to make cards in different parts of the season. Or if I am AAA, I would want to know whether people take vacation for Valentine's Day, how long do they stay and what demographic they are. Finally, if I was in the chocolate business, I would want to know how people respond to gifts of chocolates. Take for example, the tweet 'My love got me flowers instead of chocolates.' What is he trying to say? Am I fat? If I were a chocolate business, I would like to know about negative tweets as much as possible."

We found it interesting that sentiment analysis of holiday tweeter data can potentially be utilized by businesses to decide on product offerings. Given the fact that this was our first attempt at working with a large set of tweets, we decided to stay open-minded and kick off the research with the following two exploratory questions:

1. What are Tweeters' attitudes towards Valentine's Day?
2. What does Valentine's Day mean to Tweeters?

ABOUT THE DATA

Data Collection

To collect our Twitter data the Twitter Rest API (Twitter Rest API, 2017) was accessed using a modified version of the HCDE python module (HCDE Module) provided by Dr. David McDonald (McDonald, 2017). Using the parameters provided through the Twitter Rest API, tweets were collected only if considered by Twitter to be in English language and containing the term “valentine” in all derivatives regardless of case or position in the tweet text. In addition to information specific to each tweet, we also collected information specific to each user that authored the tweet and any provided place IDs.

Collection commenced at approximately February 11th (12:00 a.m. CMT) and ceased at approximately February 18th (12:00 a.m. CMT). The collection script ran continuously during that time and was able to achieve a maximum collection rate of approximately 60K tweets per hour due to rate limits set forth by the Twitter Rest API. In total, we collected approximately 4.6 million tweets which included approximately 1.6 million original tweets with the remainder being retweets. Figure 1 below provides examples of the tweets we collected.

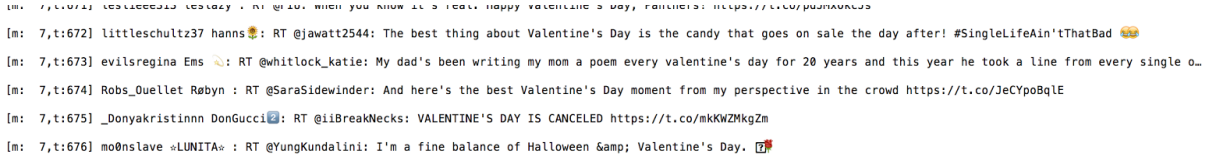

Figure 1- Examples of collected tweets

Data Storage

Collected tweets were stored in a MySQL database. MySQL Workbench was utilized to manage the database and contained data. The database consisted of three tables, one for tweets, a second for users, and the third for place ID information. Figure 2 below shows a view of data inserted into each of the aforementioned database tables.

Twitter Tweets Table

Name	Datatype
rid	**BIGINT**
tweet_id	BIGINT
tweet_id_str	VARCHAR
created_at	DATETIME
from_user_id	BIGINT
from_user_name	VARCHAR
from_user_screen_name	VARCHAR
lat	DECIMAL
lon	DECIMAL
tweet_text	VARCHAR
place_id	VARCHAR
entities	VARCHAR
favorite_count	BIGINT
in_reply_to_screen_name	VARCHAR
in_reply_to_status_id	BIGINT
in_reply_to_status_id_str	VARCHAR
in_reply_to_user_id	BIGINT
in_reply_to_user_id_str	VARCHAR
is_quote_status	TINYINT
scopes	VARCHAR
metadata	VARCHAR
retweet_count	BIGINT
retweeted	TINYINT
source	VARCHAR
truncated	TINYINT
collect_date	DATETIME

Twitter Users Table

Name	Datatype
rid	**BIGINT**
tweet_count	BIGINT
id	BIGINT
id_str	VARCHAR
name	VARCHAR
screen_name	VARCHAR
created_at	DATETIME
verified	TINYINT
geo_enabled	TINYINT
location	VARCHAR
lang	VARCHAR
time_zone	VARCHAR
url	VARCHAR
description	VARCHAR
favourites_count	BIGINT
followers_count	BIGINT
contributors_enabled	TINYINT
collect_date	DATETIME
entities	VARCHAR
statuses_count	BIGINT
listed_count	BIGINT

Twitter Place ID Table

Name	Datatype
rid	**BIGINT**
count	INT
id	VARCHAR
place_type	VARCHAR
full_name	VARCHAR
name	VARCHAR
country_code	VARCHAR
country	VARCHAR
url	VARCHAR
attributes	VARCHAR
bounding_box	VARCHAR
collect_date	DATETIME

Figure 2- Twitter Database Tables from left to right: Twitter Tweets, Twitter Users, Twitter Place IDs

DATA PREPARATION

Overview

When we started collecting tweets, we realized that we would have a challenge performing sentiment analysis as there were quite a few advertisements. Figure 3 provides a good overview of the data preparation process.

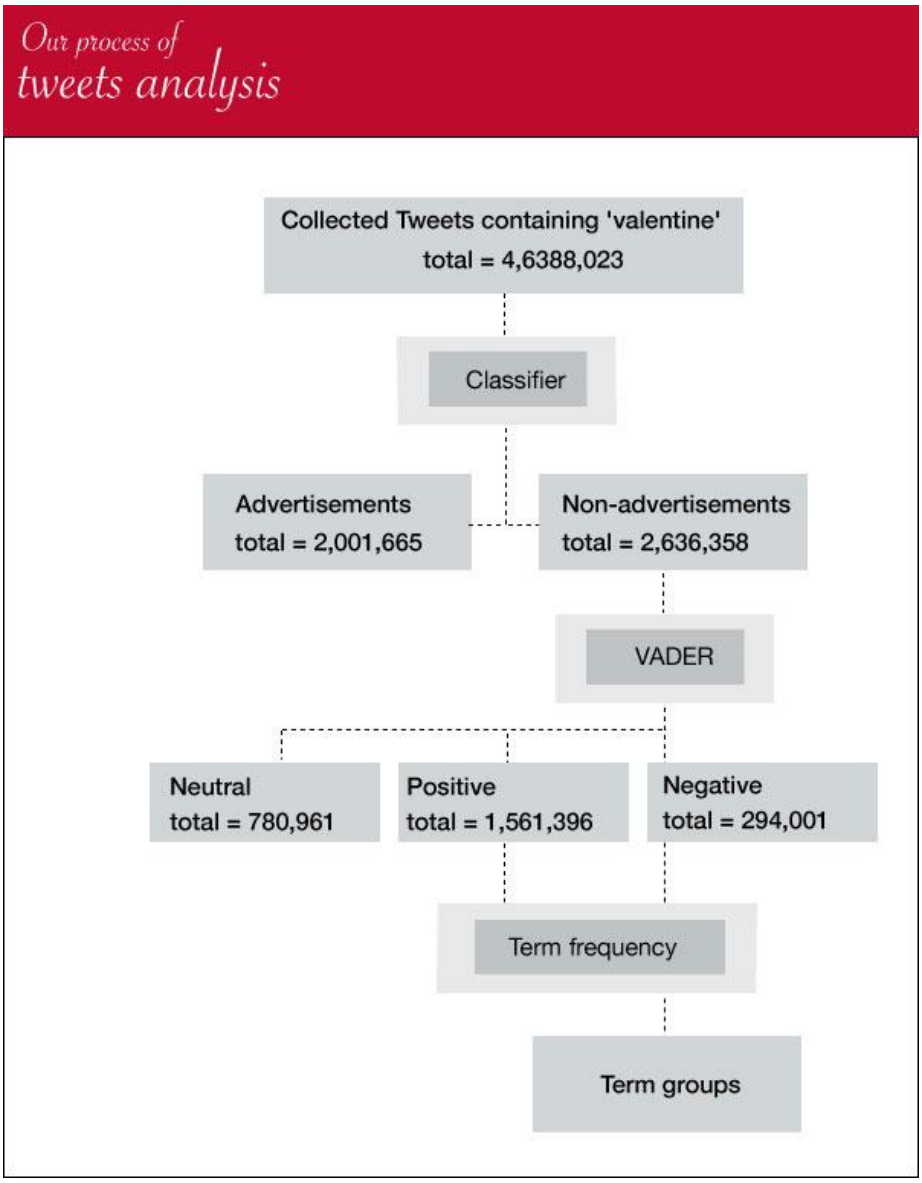

Figure 3- An overview of our data preparation process

Identification of Ads

To separate ad containing tweets from non-ad tweets we took advantage of the supervised machine learning script provided in the HCDE Module. First, we created a comma separated values (CSV) file, into which we dumped a random sample of 2750 tweets from the collected tweets database table. We then manually assigned labels to tweets, where a “pos” label corresponded to a tweet containing an advertisement, and a “neg” label corresponded to a tweet lacking an advertisement. Care was taken to avoid labeling of tweets that were not obviously ads or non-ads. An example of our labeling efforts is provided in Figure 4 below.

1955	neg	Your love is a blessing my sweet Valentine. LadyLuster Unveiled #LusterbyNadine #KCAPinoyStar #NadineLustre	831476662986420224
1956	neg	Your wcw is on social media saying she don't got a valentine but when you asked her she said she had plans	831532520457986053
1957	neg	· Happy Valentine♥ #valentine https://t.co/8V9qbya9l0	831404277306445824
1958	pos	. @TanyaNambiar is going to host "Date Face" speed dating by @thetrulymadly & @dlf_cyberhub {14 Feb} -… https://t.co/7zVlweTDyA	830828873793335297
1959	pos	.@HackneyLife_ on #Periscope: HACKNEY LIFE- Valentine's Day Special https://t.co/dNTtXD4Ntq	831600923331985408

Figure 4- Example of tweets labeled as ads or non-ads

The CSV file we created was then used as a training set for the binary classifier script. The machine learning script classified approximately 2 million tweets as ads of the 4.6 million tweets collected. The accuracy of the classifier was determined to be approximately 75% when automatically labeling tweets as ads. Meaning that of all the tweets classified as ads, approximately 25% were actually non-ads. However, accuracy of classifying non-ads was above 99%. Meaning less than 1% of tweets classified as non-ads were actually ads. The dataset was then divided and the non-ads subset used for the remainder of the preparation and analysis.

Sentiment Determination

VADER (Valence Aware Dictionary and sEntiment Reasoner) was utilized to assess the sentiment of tweets. VADER is “a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.” (Hutto, 2017). For each tweet the VADER script provides sentiment polarity (negative or positive) and a relative intensity ranging from -1, most negative, to 1, most positive. Tweets that cannot be identified as either negative nor positive are assigned a polarity of 0. When running our 4.6 million tweets through Vader, we left the entities like @names, #hashtags, and urls.

Time Zone Filtering

Each tweet was assigned a time zone based on the user that authored the tweet. Not all users provided a time zone and therefore not all tweets were able to be connected to a

time zone. A subset of tweets with a time zone in the US was selected and used for term frequency determination and subsequent analyses.

Term Frequency Determination

Tweets were further divided by positive and negative sentiment. A list of stop words was compiled from a list provided in the HCDE Module, the NLTK stop words list, and additional custom words. Term frequency was calculated for each sentiment subset after stop words were removed, using a python script utilizing various functions provided by the python NLTK module. The top 1000 most frequently occurring words for each sentiment were selected.

ANALYSIS

Overview

We used Tableau to perform all our data analysis. Depending on the question we were trying to answer, we built a series of worksheets. For example, to explore the relationship between ads vs. non-ads, we imported the data, which had already gone through machine learning script into Tableau. Only data fields relevant to our analysis were imported into Tableau. Some of the database table fields included time zone, Ads vs. Non-Ads, sentiment, and time of tweet. Filters were created using these fields to aid in data exploration.

Advertisements

We found that advertisers appear to take advantage of the Twitter social networking service during Valentine's Day. We concluded so by analyzing and visualizing the number of ads versus non-ad tweets in Tableau and plotted them per hour. As mentioned previously, at this stage of our analysis, we are focusing on time zones in the US. The time series chart shown in Figure 5 below makes the trend clear: the pattern of advertisements mimics that of non-advertisements. For example, when the volume of non-advertisements increases on February 13th at 3pm., so does the volume of advertisements (to be exact, the number of non-advertisements picks up 1 hour later at 4pm. on February 14th). At that time, the number of advertisements spikes from 7,500 tweets per hour and continues to increases for the next 8 hours. Between February 13th 11am and February 15th 11pm, the number of advertisements stays about the same at a rate of approximately 24K advertisements per hour. In the next few hours, the number of advertisements continuously decreases from 22,789 on February 15th at 11pm to 2,784 advertisements per hour on February 18th at 12pm. The glance at the pattern of two lines from 10am on February 16th through 12pm on February 18th is mesmerizing

as the lines almost coincide. This observation makes us think that advertisers are monitoring user activity on Twitter so as to tailor their offerings.

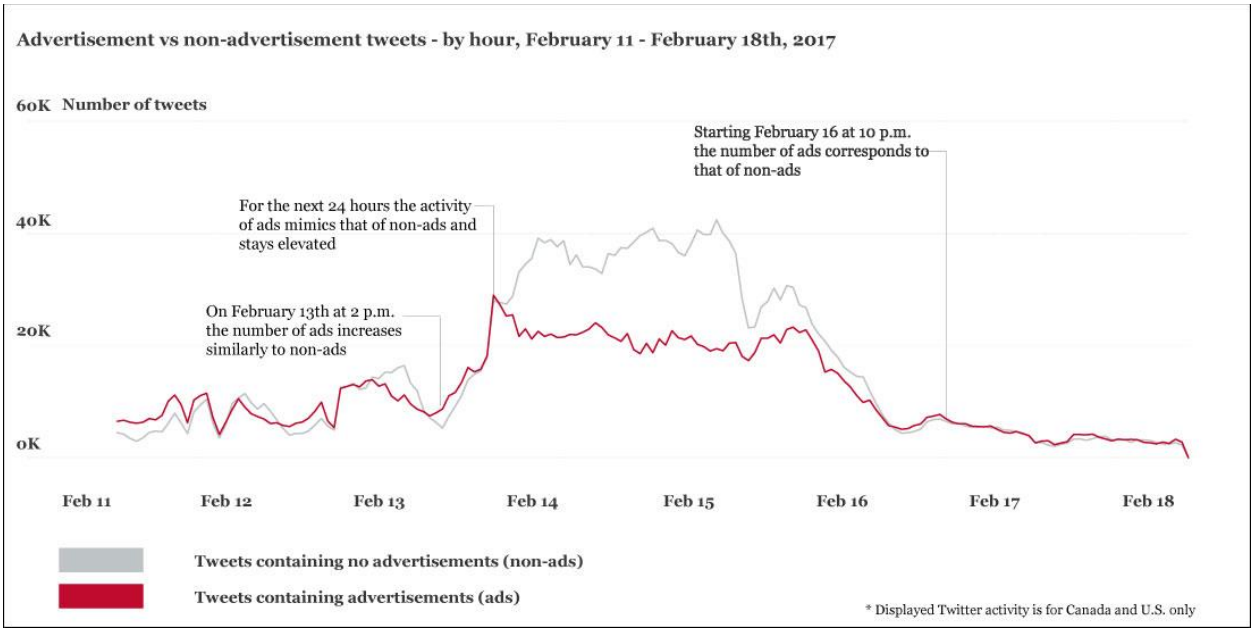

Figure 5- Advertisements vs Non-advertisements over the course of our data collection (hourly)

Sentiments

We again used a time series chart to visualize the sentiment associated with Valentine-related tweets collected between February 11 and February 18th, 2017. We find it interesting that the spikes in both positive and negative sentiment occurred almost at the same time throughout each of the seven days. For example, February 12 at 3am, February 12 at 10pm, February 13 at 4pm, February 14 at 4pm, and February 15 at 4pm are times when the number of both positive and negative tweets started climbing. We find it interesting that while both positive and negative tweets started building up numbers on February 14th at 4pm, the positive tweets took a sharp spike within a couple hours and maxed at about 11,464 tweets per hour at 10pm on the same day. Figure 6 below displays an overall polarity of Twitter data over a period of time, and acts as a starting point for the analysis. We found that social activity with people expressing their positive and negative feelings towards Valentine's Day picks up right around the holiday and despite decreases in the next couple of days.

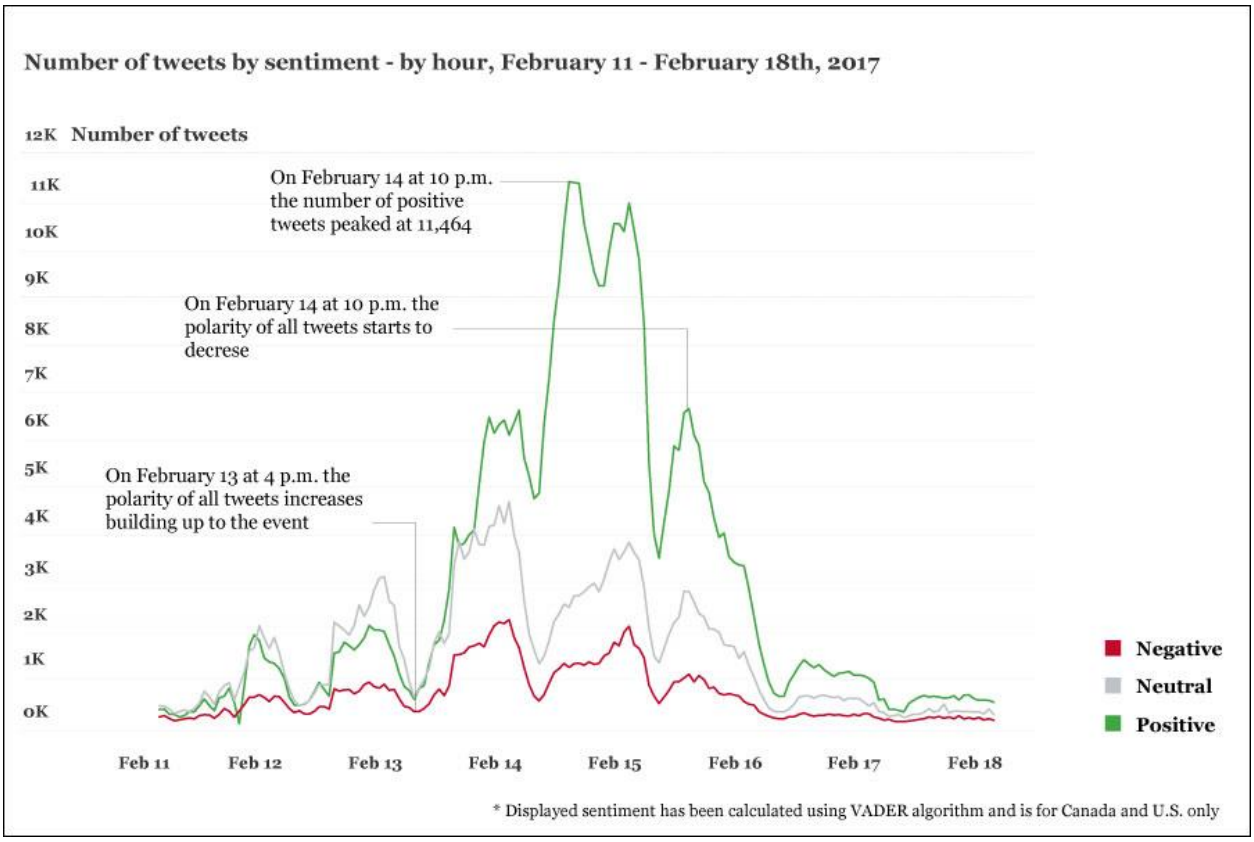

Figure 6- Number of tweets by sentiment (hourly)

Frequent Terms

While it was great to know the overall sentiment, and whether it was positive or negative, we believe analysis is more valuable and interesting when we identified what about the holiday was relevant to people. For that, we identified sentiment expressed in relation to the terms occurring within tweets themselves. Here, we first identified the top 1000 most frequently occurring tweets in our tweet collection. Next, we grouped semantically related common terms like “love”, “loves”, and “loving” and summed the term frequencies. Depending on whether terms came from positive or negative tweets, we assigned red (negative) or green (positive) colors to group terms. The size of each term group reflects the frequency of occurrence within each sentiment. For instance, as illustrated in Figure 7, the positive term group “candy” appears larger than the term “movies”, suggesting that Twitter users used terms listed within this group more frequently. We found Tableau very useful as we quickly filtered out words like “ass”, “fuck”, “cat” etc. As mentioned before, the term groups have been developed arbitrarily by us and the visualization can be edited by displaying the terms that are currently under the filter “other.” There are about 743, 477 terms within this filter.

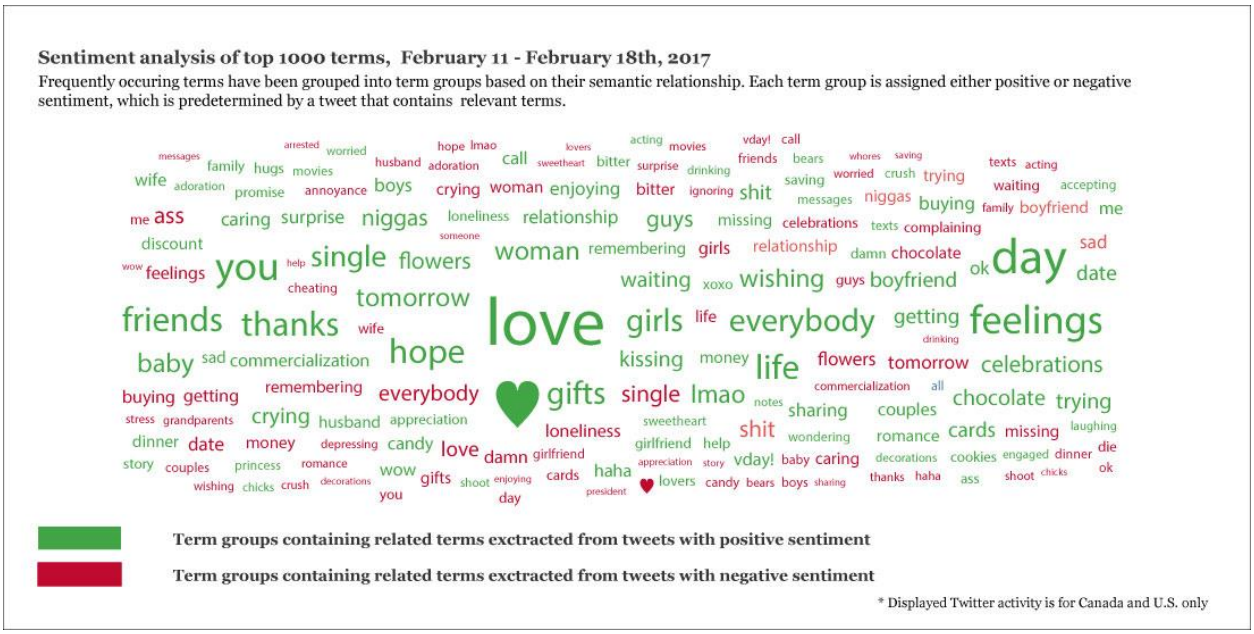

Figure 7- Top 1000 most frequently occuring terms in positive and negative sentiments

PROJECT REFLECTIONS

When working on this project, we came up with a few takeaways. Doing some level of user research is very beneficial. Having gone through the effort of collecting tweets, we felt positive that we had a large dataset consisting of 4.5 million of tweets. On the other hand, we were uncertain what to do with it. Our interviews with professionals in the field guided us in creating our visualizations. For example, knowing that business would be interested in looking at Twitter data overtime gave us an idea to visualize our tweets over a period of several days. Learning that businesses were interested in the feelings that people associate with their tweets encouraged our initial thought of doing a sentiment analysis.

Working with data involves risk. At the beginning of our research we were curious how people's sentiments changed depending on location. Having collected our tweets, only about 30,000 tweets contained geodata. The issue is that location services feature of Twitter is off by default and a user needs to opt in so as to use it. Since only a small number of our tweets contained geolocation data, we decided to omit that analysis.

We found removing stop words to be challenging. While stop words refer to the most common words in a language, there is no single universal list of stop words. Due to time

limitations of this project, we used the list of stop words provided in class. However, we plan to develop a more comprehensive list for the next stage of the project.

It is challenging to acquire the entire collection of streaming data. We tried to acquire as many tweets as possible by having each of us collect tweets. However, there is no test that would tell us what proportion of all tweets we collected.

We understand that we have to cater to users' needs when presenting data. When using Tableau, we created a whole series of graphs, including a stacked area chart displaying the number of tweets with advertisements vs. non-ads overtime. When presented with a choice of a stacked area and time series chart, the user noted that the latter illustrated the point better.

Finally, we found it helpful to go into the project open-minded and ready to explore. For instance, at the beginning we emphasized our project on sentiment analysis and analysis of geodata. We were planning to do a word cloud of most frequent terms just for fun. As we spoke to users, we found that they are actually very interested in the word cloud as it gave them ideas of how to tailor their business offerings to individuals active on Twitter.

REFERENCES

Bleier, S. (2017, February). *NLTK's list of english stopwords*. Retrieved from GitHub: https://gist.github.com/sebleier/554280

Chauhan, S. (2016, March 4). *SENTIMENT ANALYSIS SHOWS THAT VALENTINE'S DAY IS FOR SINGLES – REALLY*. Retrieved from NetBase: http://www.netbase.com/blog/sentiment-analysis-valentines-day-singles/

Hutto, C. (2017, February). *VADER-Sentiment-Analysis*. Retrieved from GitHub: https://github.com/cjhutto/vaderSentiment

McDonald, D. (2017, January). *Meeting Schedule*. Retrieved from Computational Techniques for HCDE: http://www.pensivepuffin.com/dwmcphd/syllabi/hcde530_wi17/python/hcde-user-module-January.05.2017.zip

Twitter Rest API. (2017, February). Retrieved from https://dev.twitter.com/rest/public

What is Valentine's Day? (2017, February). Retrieved from Roses Only: https://www.rosesonly.com.au/what-is-valentines-day