Emotional Content in Tweets

Group Members: Wonjohn Choi, Brenna Gomer, Hye Soo Cho, Yekun Wang, Ethan Yung
12/9/2013

Introduction

Some lifestyle choices unique to the Millennial Generation are habitual attachment to mobile devices and constant streaming of social media. Such a lifestyle shows a drastic change in modes of emotional expression and regard for privacy. For this project, we are interested in the quality and quantity of emotions expressed via tweets. More specifically, how is the emotion content on Twitter changing with time and the effect of any big events. We collected Twitter posts from Nov. 25 2013 8p.m to Dec. 3. 2013 9p.m including Thanksgiving. Thus, we tried to answer the following questions through the project:

1) How is the general trend of sentiment score throughout day and week?
2) How people's emotion was during the holiday compared to the normal days?
3) Is the feeling during the holiday more positive or negative?
4) If there is unexpected changes of emotion, what's the reasons for that?
5) What is proportion of positively valences tweets and negatively valences tweets.
6) Where tweets are coming from and how the number of tweets per hour changes?

Data

To collect tweets from Twitter, we experimented with different R libraries that are designed to fetch tweets from Twitter. First, we attempted to use twitterR but tweets returned by twitterR did not contain information about their exact location and time. Then, we attempted to use streamR. streamR could only return tweets from Twitter in real time, and did not have access to archived tweets. streamR returned a data frame with 40 columns, where each row represented a tweet and each column represented an attribute of tweet. Since the data frame included exact time (in UTC timezone), timezone, latitude, and longitude of tweets, we chose to use streamR. filtered_tweet_stream.R script collected tweets using streamR with the following steps:

1. Using Pablo Barbera's publicly open-sourced R function "filterStream", tweets within the US were collected every minute.
2. Then, the tweets were parsed to a data frame using streamR's parseTweets function.
3. The tweets were appended to a filtered_tweets variable, and the variable was stored to a file every hour.

Between 11/26/2013 to 12/03/2013, the script collected about 1.2 million tweets, averaging about 6,250 tweets/hr.

During the data manipulation phase, cleanup_tweets.R script reduced the size of data and normalized created_at column (which showed the time a tweet was posted in UTC time zone) to local time. The script processed tweets in 4 steps:

1. It removed tweets filled with NA in any of created_at, text, lat, lon, utc_offset, and time_zone columns of data frame.
2. It removed tweets written in a language other than English.
3. It removed unnecessary columns in the data frame.
4. It normalized created_at column to local time by converting created_at to Posix date/time format and adding utc_offset column.

Once we acquired the data, we processed tweets for emotionality using a modified function developed for sentiment analysis by Jeffrey Breen called "score.sentiment" (citation needed). The approach for assessing emotionality in tweets was to match the words in the tweet to words in pre-established lexicons (i.e. emotional word lists). Because our initial exploration of these resources showed that some words within lexicons were a bit "odd" (for example, "audible" appeared in a positive word lexicon), we decided to instead pool together these resources.

The original scoring scheme takes the number of positive matching words in a given text and subtracts the number of negative matching words in the given text (this will yield an overall 'sentiment score'). However, some tweets are longer than others, and some lexicons are stranger than others, so we decided to modify and weight emotionality a little differently. The modified scoring scheme instead attempts to match words in a single tweet to one of several lexicons, and counts each one of those matches. A sentiment score of 2 then may indicate that there are two positive words in a tweet, or that there is a positive word in the tweet that matches 2 of the lexicons. We hope that this reduces bias towards longer tweets and indicates degree of emotion more accurately by assessing overlap of lexicons and giving those words more weight.

The lexicons used were MPQA Opinion Corpus' Subjectivity Lexicon, WordNet's Affect Emotion Lists, a lexicon developed by Hu and Liu (2004 citation needed), and a psychological list of 'feeling' words (the latter was used in order to strengthen and emphasize the scores for tweets with very obvious emotion content). Two further modifications were needed: removing the word 'like' from the lexicons (it was like, such a confound), and writing a small 'catch-all' list of chat abbreviations with very obvious sentiment. It was not our initial intention to match chat abbreviations, but after scanning the data and noticing a sizeable number of tweets containing 'lol' that would have been misclassified, it seemed like an appropriate measure. The lists follow below:

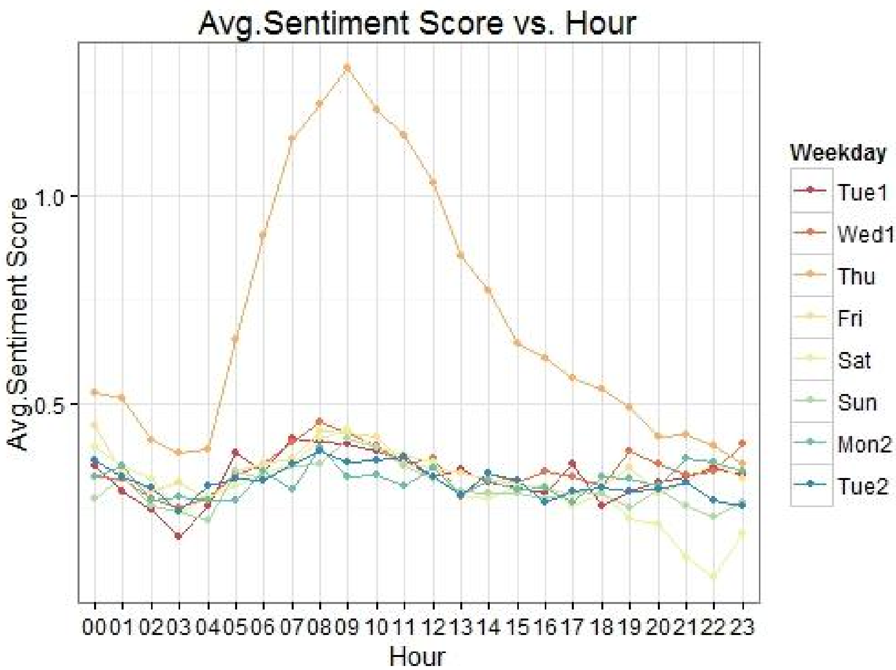
Positive:"lol","rofl","lmao","lmfao","jk","ilu","ily","thx","tu","haha","ha","hahaha","whoo","woohoo"
Negative: "wtf"

We didn’t see any negative chat abbreviations in tweets we glanced through, although we did see a lot of cursing. We noticed a lot of tweets with ‘lol’ or ‘haha’ and these patterns influenced the list we created.

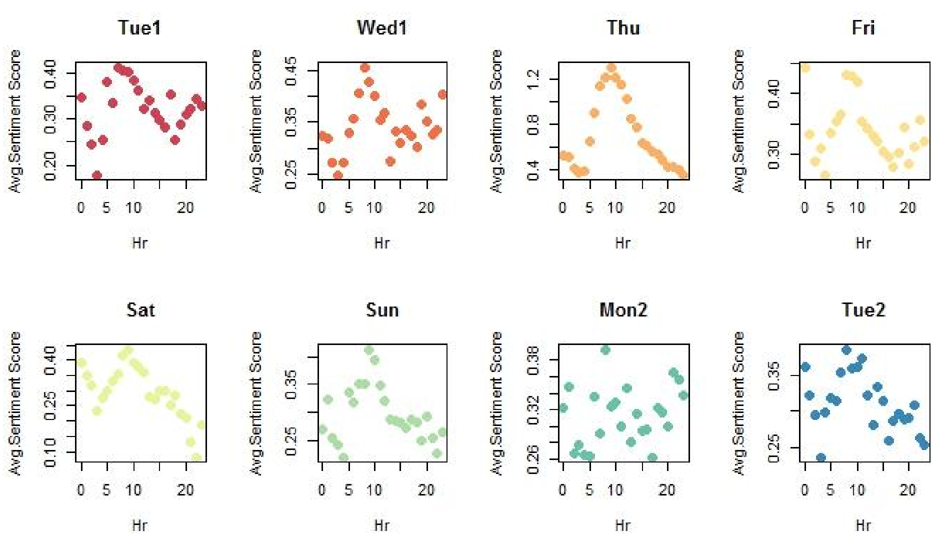
Because some tweets may have mixed emotional content or have words that match a negative and a positive lexicon, we took that information into account as well as the overall sentiment score calculated with simple arithmetic. Each tweet has sentiment information about its overall score, positive matches, negative matches, neutral matches, and ‘both’ matches. The latter two originate from MPQA’s Subjectivity Lexicon because it includes those valence categories. The “Surprise” Affect Emotion List was later incorporated into a lexicon to be used to asses for the ‘both’ category because the list of words indicated emotion but did not have a clear positive or negative valence (i.e. ‘shocked’,’baffled’,etc.).

The final modified score.sentiment function allows the user to specify which default lexicons to use for scoring, whether or not to include weights in scoring, and what threshold should be used to determine if a tweet is emotional or not emotional. The threshold is the minimum score required among the positive, negative, or ‘both’ emotional categories. It is sufficient to attain the minimum score in any of them, and is not necessarily required in the overall sentiment score itself.

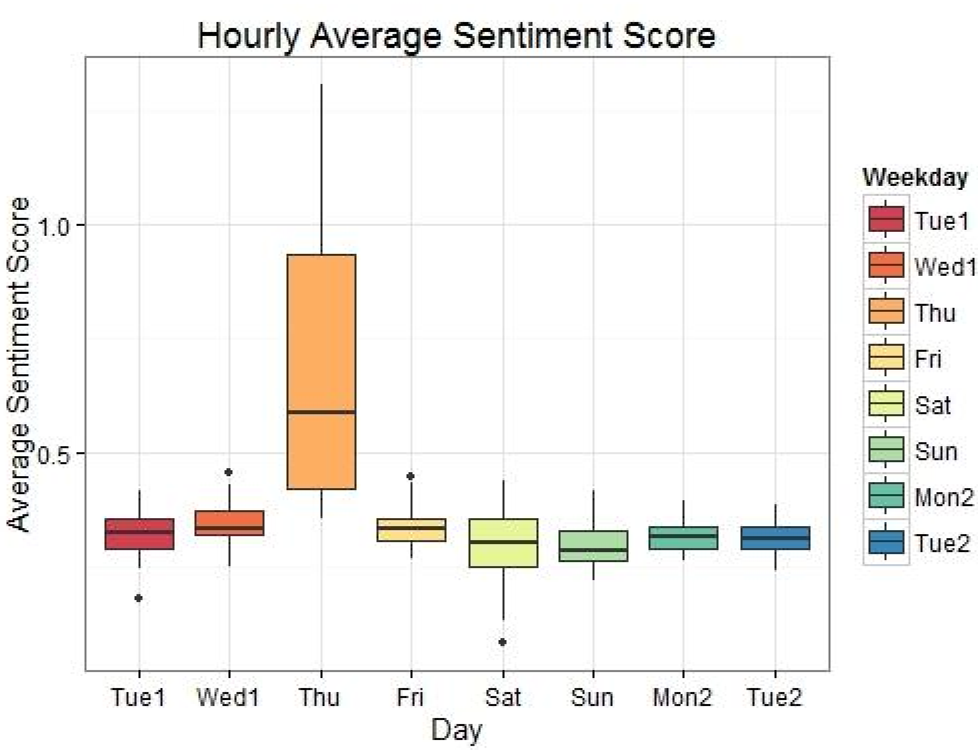
Results:



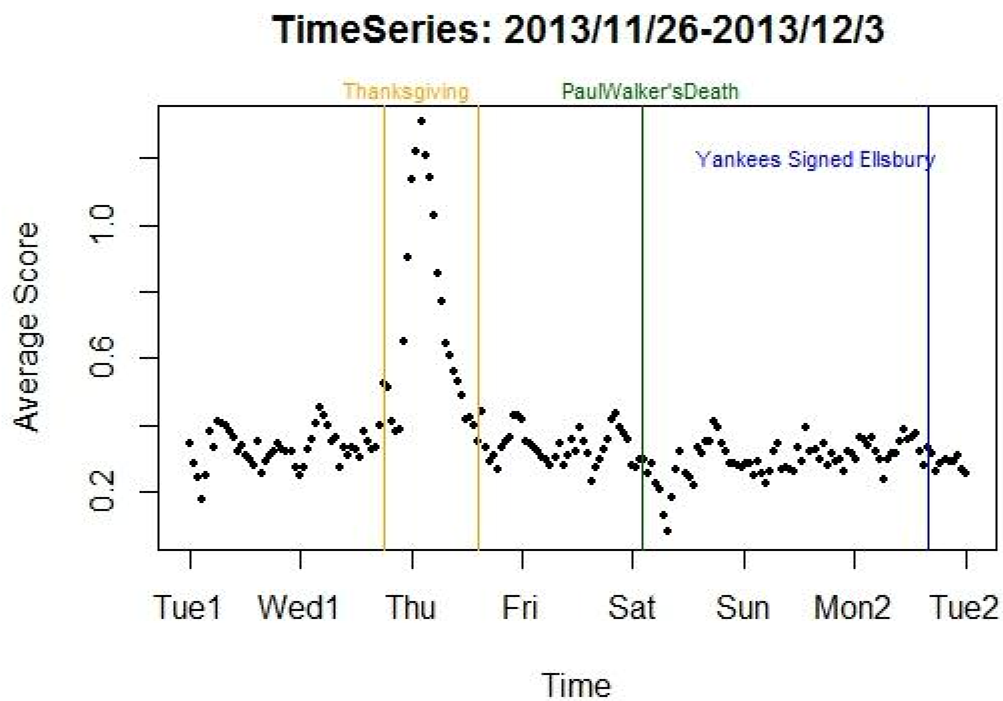
The average sentiment score vs. hour plot illustrates the hourly average sentiment score for each day during a 24 hour period. It is obvious from the plot that Thursday, the day of Thanksgiving, had significantly higher sentiment score than the rest of the week. This might be due to the cordial greetings that oftentime take place on holidays. Another key observation from the plot above is the general trend of sentiment score throughout day: the emotional content on Twitter seems to have two dips in sentiment scores around 3 a.m. and 6p.m. and a peak at 8 a.m. However, the trend in the afternoon is not so clear.



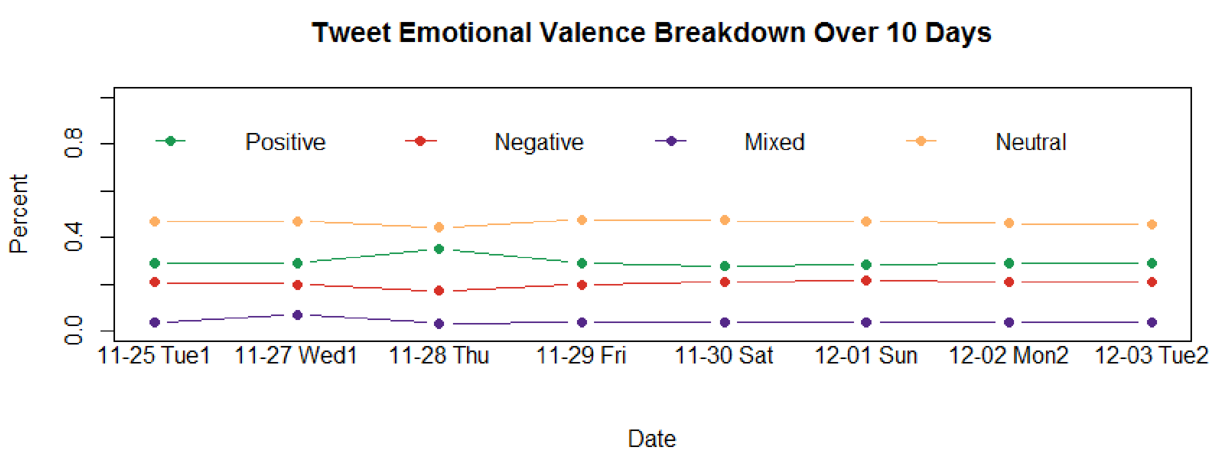
For a closer look at the trend within each day and compare trends for different days of the week, the hourly average sentiment score for each day is plotted separately. Consistent data shows that during the week, emotional content in tweets approaches zero around 3 a.m. in the early morning, peaks at about 8 a.m. and decreases in the afternoon around 6 p.m. However, the trend in the evening is not so clear, on the first Tuesday and Wednesday, the hourly average sentiment score dipped around 6p.m. in the afternoon and bounced back around midnight, whereas on Thursday and Saturday the trends continue to decrease and reach another low point for the day.



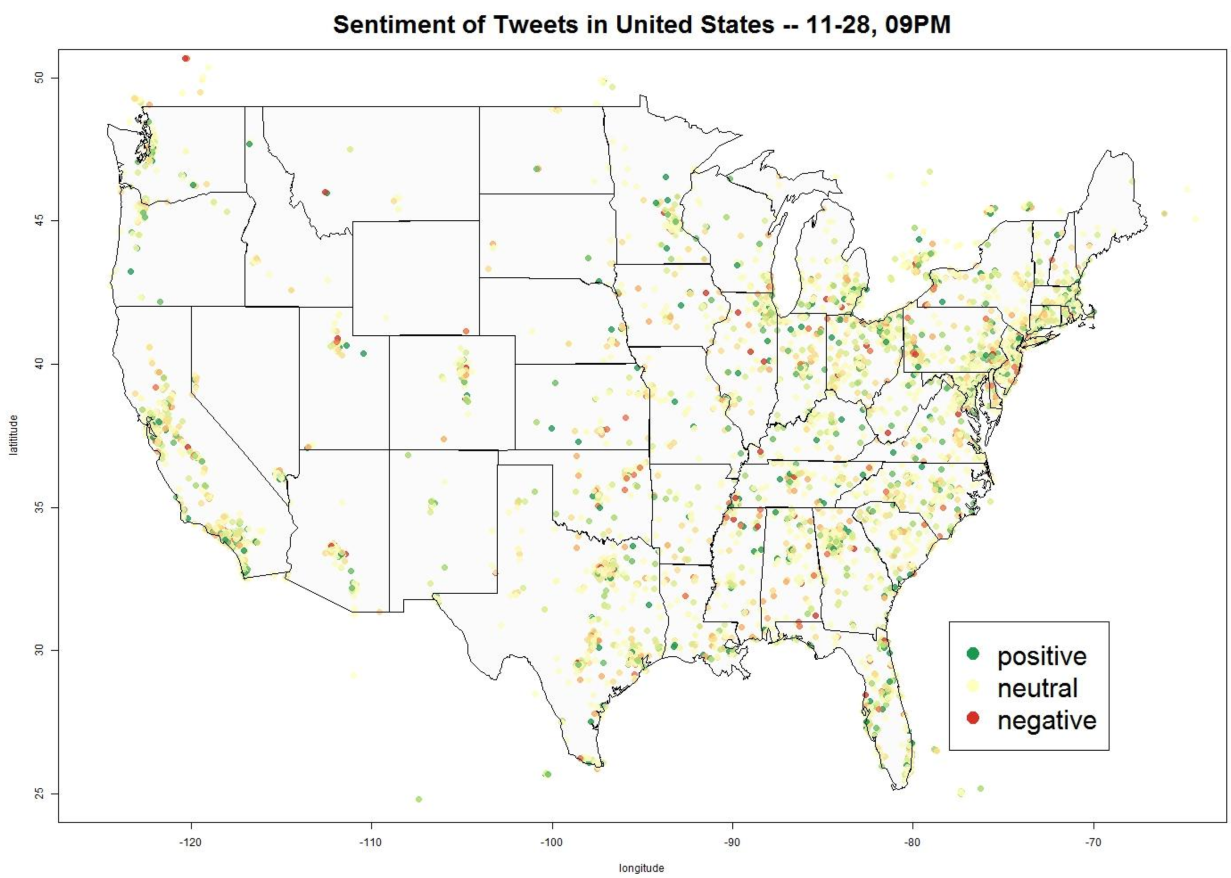
Box plot confirms that the hourly average sentiment score is consistently higher and has a wider spread on Thursday, the day of Thanksgiving, than any of the days in the week. Sentiment scores seem to have a wider spread on both Saturday and Sunday. Not including Thursday, sentiment scores seem to have similar mean and spread during weekdays.



The time series is plotted and some important events that influenced the emotional content on twitter are denoted. It is clear that many of the trends are driven by important events. Thanksgiving happened on Thursday, Nov. 28th, and the data shows a clear peak. Actor Paul Walker was injured and soon passed away on Nov. 30th, and the plot shows a dip in average sentiment scores. Moreover, on Tuesday, Dec. 3rd, Yankees signed Jacoby Ellsbury and the sentiment score shows a decreasing trend.



The plot above depicts emotional valence of tweets over the 10-day period. The overall Tweet makeup appears to be consistently dominated by neutral, non-emotional tweets over emotionally valenced tweets when considering each valence category separately.There is a slightly larger proportion of positively valenced tweets than negatively valenced ones, but the gap begins to close after Thanksgiving. It is possible that there is an unusual surge of positively valenced tweets over this 10-day period because it makes up the beginning of the 'holiday season'. Surprisingly, there is a noticeable proportion of tweets with 'mixed' emotional content (i.e. there is both positively and negatively valenced content within a single tweet). This may be a valid pattern but also there is a possibility that these results are due to the particular lexicon and sentiment analysis approach used. It is interesting to note that there are only minor fluctuations in the emotional valence category proportions represented over the 10-day period. This may suggest that there is no relationship between the proportion of emotional tweets and the day of the week (as far as valence categories are concerned).



video of maps over time: https://www.youtube.com/watch?v=dyg3MPbFC50

The maps depict individual tweets and their emotional valence score on a color spectrum from red (negative) to yellow (neutral) to green (positive). Extreme outlier tweets, in terms of score, were excluded.

Plotting the tweets and their sentiment on a map over time provides insight on where tweets are coming from and how the number of tweets per hour changes. It's clear that the tweets center around major cities (e.g., San Francisco, Los Angeles, Chicago, New York City, Dallas). Also, the number of tweets per hour fluctuates in a daily cycle; it dies down in early morning after midnight, grows throughout the morning, and grows again through evening. More detailed analysis could examine how specific regions differ in emotional valence (e.g., which city is more positive or negative in emotion?).

Confounding Factors and limitations

Initially, we set out to investigate how emotions change throughout the course of a day or week. However, we were limited to about a week's worth of data (which still amounted to many megabytes of information), and the general trends were partially confounded by significant events. For example, Thanksgiving had an overwhelmingly positive effect on emotion, or at least emotion conveyed over Twitter. The other two causes of significant emotional change (Paul Walker's death, and the signing-on of Jacoby Ellsbury to the Yankees) were surmised by the timing of public news releases, and by reading some of the raw tweet text.

Additionally, time zone designation was tricky to deal with. The tweets are grouped by their local times, which makes sense for analysis by time of day. However, significant news events will break almost instantaneously across the entire U.S., and thus their emotional impact will be captured across different local times. Additionally, there were some tweets found with impossible local times for tweets in the U.S. (likely because outside time zones were somehow set to those tweets), and thus some tweets in the data are incorrectly timed.

Conclusion

The emotional content of Twitter posts was found to follow a regular cyclical pattern throughout the course of day: most negative in the early morning, reaching a peak at around 8 am, decreasing through the afternoon, and increasing again in the late evening. However, emotional content was also highly sensitive to significant events, which confounded our findings. Our process is validated by the fact that the sentiment scores followed these significant events appropriately (e.g., Thanksgiving was overwhelmingly positive). Also, by geographic mapping, we found that tweets cluster around major cities. Further analysis can be done by teasing out the effect of significant events versus cyclical patterns, collecting data for longer periods of time, accounting for events that are instantaneous across time zones, and investigating the differences in emotional content between particular regions. Additionally, the link from emotional content of Twitter posts to peoples' actual emotions over time should be further examined.

References

http://mpqa.cs.pitt.edu/lexicons/
http://www.psychpage.com/learning/library/assess/feelings.html
http://wndomains.fbk.eu/wnaffect.html
http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
streamR package: http://cran.r-project.org/web/packages/streamR/streamR.pdf
stringr package: http://cran.r-project.org/web/packages/stringr/stringr.pdf