Ziyad Mir (20333385), Jennifer Blight (20347163)
Stephanie Cheng (20346637), Chi Wai Kenneth Ng (20341536)

SYDE 212 Project

Introduction

Twitter is an online social media tool that enables users to send and receive 140-character messages, referred to in this report as "tweets". The focus of our statistical analysis is on the length of tweet against the time of day in which it was sent, as well as the emotion of the tweet against the time of day in which it was sent, as measured by the use of emoticons. The goal is to identify any possible relationships between these variables, and draw inferences upon our population of interest, the entire population of Twitter users.

Questions of Interest

1) Is there a difference in mean length of a tweet between the morning, afternoon, evening and night?
2) Is there a difference in the proportion of tweets that include 'smiles' and 'frowns' between the morning, afternoon, evening and night?

Methods

The Streaming API, made public by Twitter, generates a live stream of tweets referred to as the 'Firehose' which can be collected by opening a HTTP connection to the appropriate URL. One of the features of this API is the sampling feed known as Spritzer provides access to approximately 1% of all public tweets based on a sampling algortihm. According to Twitter, this sampling algorithm combined with a status ID assignment algorithm produces a result similar to a simple random selection. [1] Tweets are collected in the form of a .json string consisting of a comprehensive set of metadata including the tweet message, timestamp including time zone, user twitter ID, and numeric identifier. Our dataset of tweets was stored in an SQLite database, in order to facilitate quick data retrieval. Data was extracted from the twitter feed over the course of a weekend, providing 8GB of data for post-processing. After limiting our dataset to tweets originating in Canada and the United States, our sample consisted of a total of 687,012 pseudo-random tweets.

Analysis and Results

The analysis consisted of categorizing the collected tweets into one of four 6-hour time intervals, as outlined below:

Period	Time Interval
Morning	6:00 a.m. - 11:59 a.m.
Afternoon	12:00 p.m. - 5:59 p.m.
Evening	6:00 p.m. - 11:59 p.m.
Night	12:00 a.m. - 5:59 a.m.

Table 1 Time Intervals used to Categorize Tweets

The measurement variables of interest were the average character length of tweet, as well as the existence of happy and sad emoticons including ':)', '=)', ':(', or '=(', versus the time the tweet was created. These variables were directly accessible from post-processing parsing of the tweet text.

Our analysis will focus on two types of tests in order to measure the variables of interest:

1. Mean difference in length of tweet between each time interval based on confidence intervals
2. Chi-squared test for instances of ':)', '=)' and/or/versus ':(','=(' with tweet text

Mean Difference in Length

Using a 99% confidence interval, our group conducted hypothesis tests between the tweets from each time period in order to analyze the study variables and their correlation with the time they were created.

Statistics collected on our data set of 687,012 randomly generated sample tweets are for each time interval (period) including the respective sample size, sample mean, the average character length of each tweet (maximum 140), and sample standard deviation. A summary table of these statistics is given below:

	Total tweets/ interval	Mean	Standard Deviation
Morning	139002	68.7437	39.151
Afternoon	244774	64.593	37.434
Evening	238525	64.379	37.666
Night	64711	63.110	37.549
SAMPLE SPACE	**687,012**		

Table 2 Total, Mean, and Standard Deviation of Tweets per Time Interval

A hypothesis test was conducted to determine whether a significant difference existed between the mean tweet lengths for tweets posted in each respective time interval. The following table summarizes the results.

Sample Time Intervals	Difference in Mean	Standard Error	99% CI	Conclusion
Morning vs. Night	5.634	0.181	5.127 - 6.100	Null hypothesis rejected.
Morning vs. Evening	4.365	0.130	4.030 - 4.701	Null hypothesis rejected.
Morning vs. Afternoon	4.150	0.129	3.817 - 4.484	Null hypothesis rejected.
Evening vs. Afternoon	0.215	0.108	-0.493 - 0.064	Null hypothesis not rejected.
Night vs. Evening	1.269	0.167	1.056 - 1.911	Null hypothesis rejected.
Afternoon vs. Night	1.483	0.166	0.840 - 1.698	Null hypothesis rejected.

Table 3 Hypothesis Test for Difference in Mean of Each Time Interval

Based on the conclusion of our hypothesis tests, we are 99% confident that there is a significant difference in mean tweet lengths between all time periods *except* for tweets between the evening and the afternoon.

Use of Emoticons

Chi-squared tests were conducted to determine whether or not the proportion of 'smiles' and 'frowns' remains constant over the 4 time intervals. A tweet containing one of the following characters (':)',':D',':P','=)','=D','=P') was defined to have a 'smile', and a tweet containing (':(','=(') was defined to have a 'frown'. The terms 'smile' and 'frown' act as simplified indicators of the mood of a tweet, however, this does not absolutely gauge the specific mood associated with a given tweet. The null hypothesis associated with the chi-squared test is that the proportion of 'smiles' or 'frowns' is independent of the time interval in which the tweet occurred. The data collected, as well as the calculated chi-squared values, are shown below.

	Total	Smiles	Proportion of Tweets with Smiles	Frowns	Proportion of Tweets with Frowns
Morning	139002	8835	0.0637	1636	0.01177
Afternoon	244774	12394	0.0506	2332	0.009527
Evening	238525	13539	0.0568	2647	0.01110
Night	64711	3450	0.0533	720	0.01113

Table 4 Propertion of Tweets Containing Smiles and Frowns

	Total	Observed		Expected	
		Smiles	No Smiles	Smiles	No Smiles
Morning	139002	8835	130167	7732.585	131269.415
Afternoon	244774	12394	232380	13616.607	231157.393
Evening	238525	13539	224986	13268.980	225256.020
Night	64711	3450	61261	3599.828	61111.172
Total	687012	38218	648794		

χ^2	295.091
Cutoff Value ($\alpha = 0.001$)	16.27

Table 5 Relative Chi-Squared Analysis Calculations for Smiles

	Total	Observed		Expected	
		Frowns	No Frowns	Frowns	No Frowns
Morning	139002	1636	137366	1484.078	137517.922
Afternoon	244774	2332	242442	2613.371	242160.629
Evening	238525	2647	235878	2546.653	235978.347
Night	64711	720	63991	690.898	64020.102
Total	687012	7335	679677		

χ^2	51.576
Cutoff Value ($\alpha = 0.001$)	16.27

Table 6 Relative Chi-Squared Analysis Calculations for Frowns

Using a significance level of $\alpha = 0.001$, we can reject the null hypothesis for both the 'smile' and 'frown' tests. Based on this analysis, the use of either of these emoticons is not independent of time.

Conclusion

From the results of our analysis, it is possible to answer both of our original questions of interest.

Firstly, based on hypothesis testing performed using 99% confidence intervals, there appears to be a significant difference in the mean length of tweets created at different times of day, with the exception of the afternoon to evening case. The data suggests that tweets are longest in the morning from 6am to noon and shortest late at night from midnight to 6am.

Secondly, the results of the chi-squared tests indicate that the proportions of tweets with smiles and frowns are different depending on the time of day. While the chi-squared test only rejects the hypothesis that the proportions are equal, the observed data suggests that there are proportionately more smiles during the morning than at any other time, while the period with the least frowns appears to be the afternoon.

It should be noted that while our sampling methods ensure that the tweets collected are representative of the entire stream at the time of collection, the content of Twitter is very much influenced by what is happening in the world at the time. Thus our data is only representative of the weekend that it was collected and data gathered on different days, such as weekdays, might yield drastically different results. Further investigation would require data collected over a longer period of time to more accurately represent the entire tweet population.

Works Cited

[1] Twitter. *Streaming API Concepts | Twitter Developers*. November 24, 2011 https://dev.twitter.com/docs/streaming-api/concepts#sampling