Here an approach to understanding sentiment using spatially augmented Twitter data is presented. This work uses the dataset collected by Soliman and Wang (2015) to understand how sentiment relates to the unexpected social patterns the previous authors discovered. We use two different analytic approaches to determine sentiment differences between users of educational services at educational service locations and non-educational service locations. For the purposes of this work, educational services are defined as spatial areas of Chicago that represent and education related landuse type (e.g. k-12, college and university). Our first analysis method assessed the density of sentiment term over spatio-temporal windows of data. Our second analysis method used a co-occurrence based data-mining approach to interrogate the polarity of sentiment within the data. We demonstrate that sentiment related differences do exist, which is a novel discovery that has many implications for future works in urban mobility.

Introduction

Sentiment analysis of social media is an approach to discovering knowledge about our society. Social media generally provides an unsolicited set of information from individuals about current topics and trends happening in their daily lives. In their paper “Where Chicagoans tweet the most: Semantic analysis of preferential return locations of Twitter users” (Soliman and Wang 2015), the authors present a framework of understanding how twitter users are using various urban resources of Chicago using only geo-tagged information. In their work, urban resources are represented as various land classification types to denote various city services (e.g. residential, education) to understand daily urban mobility. However, the result of the hypothesis testing leaves more questions unanswered than answered. Further, the tweets associated with the use of educational services of the city have unexpected trajectories.

While our approach can be generalizable to all services, we focus on educational services in this effort because: 1) educational services users express unique spatiotemporal trajectories as observed by Soliman and Wang (2015), 2) represent a distinct demographic sub-group of society (i.e. younger), 3) have a large number of tweets (active users).

Here, we investigate the twitter message data using spatio-temporal binning and sentiment analysis to better understand the relationship between Chicagoans and the education services used. Specifically, we seek to determine if users of the education space and non-users of the educational space post Twitter messages with sentiment laden terms. If sentiment laden terms are used, we will determine the differences in the sentiment term use, which may indicate further research possibilities to discriminate between spatio-temporal user communities.

These broader questions were distilled into testable hypotheses. It should be noted that there is no general reason for the stance of the hypotheses listed (e.g. that sentiment differential may exist between the two groups); however, these hypotheses will guide the types of investigations pursued in this work.

We develop and present a sentiment analysis method which we use to attempt to answer the following questions:

1. Is there a difference in the happiness of individuals who spend more time using educational services versus individuals who spend less time?
2. Of the users who use the educational services, do more frequent users of educational services exhibit different or similar sentiment than those users who are not frequent users of educational services?

3. Further, is the sentiment of users of educational services temporally consistent?

It is expected that the answers to these questions will illuminate possible avenues of future related research for other urban services in Chicago. These lines of inquiry also augment the spatio-temporal understanding of urban usage and twitter analysis.

Related Work

Twitter Data

In “A scalable framework for spatiotemporal analysis of location-based social media data”, Cao et al. (2015) describe Twitter’s exceptional growth as a social media platform stating that it “...has grown at an exponential rate since its founding” (Cao et al. 2015). The monthly number of Twitter users exceeds 3.9% of the global population and nearly 18% of the United States. The overall data volume of the platform is nearly 300 Billion tweets (Cao et al. 2015). In their work, “Enhanced Twitter Sentiment Classification Using Contextual Information”, Zhou et al. (2015) describe Twitter data as being extremely noisy, providing individual tweet messages that have limited the unstructured content.

These tweet messages are then ascribed a treasure-trove of metadata for each tweet provided by the social media platform. The structured metadata allows each tweet to be situated in a spatio-temporal context and linked to other tweets through relationships between Twitter users, topic related hashtags, or message syntax and language. Exploiting specific features in particular ways in order to better understand sentiment is a mandatory endeavor when trying to employ the resource for further application. Much research has gone into leveraging Twitter data to understand human mobility and the sentiment society feels towards trending topics in popular culture.

Twitter poses unique challenges not only because it limits user input to 140 characters or less, but it occupies a niche in the overall Social Media product space. Facebook for example, provides a very rich heterogeneously typed content platform providing greater flexibility of use to the user to share, contribute, and organize many types of data. From the user’s perspective, Twitter provides a method to contribute and share tweets and enough functionality to link with other users to form friendship and follower networks. It is also important to understand how Twitter is used because this information can inform the semantics, ontologies, and latent structures of the dataset. Twitter users create messages that are 140 characters or less and can ascribe topic-based hashtags to the tweet to associate the message to a larger social context. In their paper, “What is Twitter, a Social Network or a News Media? Categories and Subject Descriptors”, Kwak et al. (2010) demonstrate that message propagation through the twitter follower graph is near instantaneous to the 4th hop once it makes an initial hop. Stated differently, a retweeted message propagates very quickly once it is retweeted initially. The authors conclude that the structure and behavior of Twitter is not as similar to other social media platforms as originally thought (Kwak et al. 2010) and acts more similar to a news service.

Sentiment Analysis

Sentiment analysis is the process by which information is extracted from a provided source to understand the sentiment, or mood, expressed by the source’s authors. Sentiment analysis, on the whole, focuses on extracting the mood from a source through analysis of a set of provided features rather than what the author’s opinion of a topic may be. For instance, a user may have a negative opinion of the discussed subject but be happy to discuss the topic in a broader sense, leading to a positive sentiment— or mood—while holding a negative opinion. This is why some sentiment classifiers are broken down into two classifiers, subjectivity and polarity classification Jiang et al.’s (2011).

Sasahara (2014) generate emoticon co-occurrence networks to demonstrate the interactions between verbal and non-verbal communication. Different languages also use different sets of emoticons with different frequency. While English speakers generally use a single emoticon to express a happy smile (:-)), Japanese are known to use up to five different emoticons to express nuances of the same feeling (Sasahara 2014). Sasahara (2014) notes that the use of emoticons within tweets varies. Some emoticons are heavily used, while most are rarely observed. Further, many tweets do not contain emoticons so analysis of Twit-
ter using emoticons as a semantic tag for sentiment requires a large volume of tweets. Sasahara’s (2014) work demonstrates the formation of networks demonstrating how the use and frequency of co-occurring emoticons changes through time and oscillates between different sets of emoticons. The work applies this theory to assess the affect the nuclear incident had on the mood of Japanese society (in an immediate sense). The paper demonstrates that substantial information is gained by observing the changes of specific emoticon usage by a community of Twitter users through time. For example, during the nuclear disaster, the use of happiness related emoticons plummeted; such that the structure of the emoticon co-occurrence network changed its structure dramatically.

**Spatial Trajectories**

In the simplest of terms, spatio-temporal data mining of Twitter data is the situating of individual tweets on a map. While this may be the earliest form of spatio-temporal data mining more recent research has developed greater sophistication for extracting spatio-temporal patterns in the form of spatial trajectories to gain new insights on human behavior. The classification of users into trajectories provides a ranking facility to group users into communities based on their spatio-temporal utilization of the landscape.

In Cao et al.’s (2015) work, “A scalable framework for spatiotemporal analysis of location-based social media data”, the authors introduce a method of understanding spatio-temporal movement paths of Twitter users. These paths are known as trajectories. Trajectories represent the motion of a Twitter user in both space and time, instead of understanding singular tweet locations or locations of clustered tweets from different users.

In their paper, “Where Chicagoans tweet the most: Semantic analysis of preferential return locations of Twitter users”, Soliman and Wang (2015) present a framework of understanding how twitter users are using various urban resources of Chicago using only geo-tagged information by building on the work of Jurdak et al. (2015). In their work, urban resources are represented as various land classification types to denote various city services (e.g. residential, education, office) to understand daily urban mobility and social usage of the urban infrastructure of Chicago.

Soliman and Wang (2015) use land and building use information as a basemap to describe a range of city services. A city service is best described as a social resource; such as a workplace, education facility, home, etc. and should not be construed to be the layman term for utility, waste, and sewer services. The authors use the land use information basemap to assign semantics about the frequently visited locations of the Twitter users’ trajectories. Tweets were not trivially assigned as point-in-polygon because many tweets occur on sidewalks and streets. Instead, tweets were clustered for each user to define a frequently visited location using DBSCAN following the method proposed by Jurdak et al. (2015). Then the center point of the cluster was used to assign the tweets to a particular land use type. By enriching the Twitter dataset with land use information it became possible to describe the semantics of a Twitter user’s daily mobility (Soliman and Wang 2015). For example, moving from a frequently visited residential location to a frequently visited office location is indicative of a person going to work; the reverse path would be going home. This type of information is extremely powerful and very useful for city planners and policy makers to determine how citizenry interact with the city landscape and urban infrastructure.

Further, Soliman and Wang (2015) notes some remarkable properties of the Twitter users’ trajectories based on the enriched land use information. Previous assumptions, based on the work of Jurdak et al. (2015), led Soliman and Wang (2015) to develop simple hypotheses of human behavior that were over-turned. For example, it was assumed that most users move from home to work or that most users only travel to a single school on a regular basis. In fact, many users did not follow these patterns because many people do not work in office buildings. These findings are likely more pronounced in the Twitter dataset due to its unrepresentative sampling of a local population (i.e. Twitter users tend to be an over-sampling of those that have lower incomes than the real population).

Delving into the spatio-temporal trajectories of Twitter users as an investigation medium into human mobility is a relatively new approach to spatio-temporal knowledge discovery using social media. However, the current state of the research yields highly enriched data sets that are ascribed semantics based on geographic context and context based on the movement between frequently visited locations.
Research Data

Our research effort investigates the dataset produced by Soliman and Wang (2015). The dataset spans nearly 12 months of Twitter user activity (the 2014 calendar year) for the City of Chicago and has been cleaned to represent human users of Twitter. Month to month variations in were shown to be negligible. For our analysis we focus on the month of October, 2014 for further investigation which contains 3,232,116 spatially augmented tweets that have been ascribed land use attributes.

Sentiment analysis often requires the use of a lexicon to associate a sentiment score to each sentiment laden term in the message bodies. We use two different lexicons. Our first lexicon, from Kralj Novak et al. (2015), was used to ascribe sentiment scores to unicode emoticons. Our second lexicon, from Liu, Hu, and Cheng (2005), was used to classify the sentiment of words found in the tweets. We used the python emoji package (https://pypi.python.org/pypi/emoji/) to parse emojis in each tweet, and used the python Natural Language Toolkit Tokenizer (NLTK) to parse the content of each tweet (http://www.nltk.org/api/nltk.tokenize.html). The NLTK tokenizer was used because it provides conveniences for parsing twitter messages specifically.

Methods

We divide our dataset into two groups. Tweets made from an educational resource location: tagged by landuse codes 1321 and 1322; and non-educational resource locations. An educational resource location is any landuse category that has some educational function for K-12 or university level education. This allows for the comparison between the two groups to determine any observed differences. This effort uses two different algorithms to address the research questions. The first algorithm focuses on assessing any differences in the sentiment term density of our dataset. The second algorithm focuses on understanding to co-occurrence matrix of sentiment laden terms using an approach similar to that introduced by Sasahara (2014).

First out data is prepared and cleansed for further sentiment analysis following procedures discussed by earlier works. The NLTK tweet tokenizer is applied to tokenize each tweet message. All instances of web-urls are removed from the tweet message bodies as well as links to twitter users, which are preceeded by the “@” symbol followed by a username. Next, each token of the tweet is categorized as a hashtag, emoticon, or word.

Our lexicons were parsed into a file format that contained a term and a sentiment score. In the case of the lexicon provided by Liu, Hu, and Cheng (2005), the negative and positive scores for each term were summed to get a more generalized sentiment score. Some terms may express different sentiment connotations depending on how the term is used. The emoticon lexicon produced by Kralj Novak et al. (2015) was already in the desired format.

Our first algorithm was designed to determine if there was any notable differences in the frequency of the expression of sentiment by both groups. We divided our data into hour-of-the-day and day-of-week temporal windows to see how differently the groups may operate during the intra-day period and over the course of a week.

Our approach was to parse each twitter message, \( t \), in the set of messages, \( T \), into a set of tokens. The count of sentiment laden tokens, \( s \), in each \( t \) was divided by the total number of tokens in the tweet \( n \). This provides a metric of sentiment density for each tweet. Sentiment laden terms are defined as the sentiment terms, including emojis, found in the lexicon, \( L \). The mean of the sentiment densities for each tweet within a temporal window was then calculated for the education and non-education user groups. \( N \) is the total number of tweets within a temporal window.

\[
\frac{\sum_{t=0}^{N} \text{sum}(s_t)}{\text{length}(t)}
\]

Our second algorithm was similar to the approach used by Sasahara (2014), but we expand our co-occurrence matrix to be a larger set of sentiment related terms. For every tweet, \( t \), within the set of tweets, \( T \), within a temporal window, we count the number of times pairs of terms in our lexicon co-occur. For example, if the “:-)” term occurs with the “love” term in 100 tweets, this co-occurrence would be given a value of 100 within our matrix. The sum of the scores of each sentiment term pair was computed. Then a weighted average was computed for every pair score in the matrix, with the weight being defined by the total number of occurrences of the pair divided by the total number of occurrences of all pairs in the
Taking $s_i$ to be the sentiment of word $i$ and $w_{ij}$ to be the normalized co-occurrence of words $i$ and $j$ for all tweets $t \in T$ we construct the co-occurrence matrix as follows:

$$
\begin{pmatrix}
(s_1 + s_1)w_{11} & (s_1 + s_2)w_{12} & \ldots & (s_1 + s_n)w_{1n} \\
(s_2 + s_1)w_{21} & (s_2 + s_2)w_{22} & \ldots & (s_2 + s_n)w_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
(s_m + s_1)w_{m1} & (s_m + s_2)w_{m2} & \ldots & (s_m + s_n)w_{mn}
\end{pmatrix}
$$

Results

The results of our first metric presented an exciting and meaningful discovery. In Figure 1, the proclivity of educational services users to express more sentiment on twitter is quite pronounced during the normal school time hours. There is an obvious difference between the willingness to express sentiment related content between the two groups of individuals. This suggests a spatiotemporal social process is occurring. It is uncertain whether or not this is more related to the freedom to tweet sentiment, perhaps school attendees are less censored than professionals, or if this is an age related process.

Sentiment Density by Hour of Day

Further, Figure 2 plainly shows the differentiation between sentiment expression and day of the week. Notice that the weekdays associated with education service use have greater sentiment expression on twitter than non-educational services users. It is notable that in both cases, the sentiment densities come into alignment when the educational services are the least active within the Chicago: the evening and night-time hours, and weekends.

Sentiment Density by Weekday

Now that we can establish that sentiment is more commonly expressed during certain temporal periods by users of the educational service, we can begin to assess the sentiment of the polarity expressed. Sentiment polarity is a quantification of how positive or negative the sentiment is.

Figure 1: Mean sentiment expression density between education (blue) and non-education (red) landuse types for each hour of the day during the month of October.

Figure 2: Mean sentiment expression density between education (blue) and non-education (red) landuse types for each weekday during the month of October.
Figure 3: Differences in the expression of Polarity between education (blue) and non-education (red) landuse types for each hour during the month of October.

Figure 4: Differences in the expression of Polarity between education (blue) and non-education (red) landuse types for each weekday during the month of October.

Figure 5: A plot of the sentiment occurrence between education (blue) and non-education (red) landuse types is also plotted.

The co-occurrence approach used in the second metric did not yield substantial differences in sentiment polarity. While education related polarity does skew slightly more negative, the difference is very minuscule. We cannot say for certain that the sentiment expressed actually means the sub-population of the society using the educational service has a more negative mood or not. Also, the slight systematic negative skew in our data is likely due to the lexicon used, which has almost twice the number of negative terms than positive terms. However, it is also known that people tend to complain and discuss misery on Twitter, perhaps more frequently than happy activities is discussed.

The final plot of Figure 3 is the polarity score for each temporal window between the two groups. There appears to be some diurnal cycling of the educational service related windows. Also, the noise level is shown to be extremely high; which corroborates earlier sentiment related studies of Twitter data. The slight peak in values at the end of the temporal window may be due to the observance of Halloween related celebrations.
Discussion

Results

Based on the data presented and how it was analyzed no strong conclusion can be drawn about the reason of the differences in sentiment. Referring to Figure 1, there is a noticeable difference from 7am to 3pm, the hours that school is usually in session, between the users who are in areas of education versus those who are not. What is noticeable, as well, is that the sentiment tends to re-normalize to the same level of sentiment of non-education areas in the later hours of the day. The reasons that education users have higher sentiment density during these hours could be due to a number of reasons: younger people having more sentiment to be expressed in general, teaching being a more sentiment-involved occupation, students using emoticons and sentiment-heavy words more frequently than the average user, stresses associated with education leading to polarizing emotions, or even a different type of vocabulary that reflects sentiment differently being used by a different age group. However, it could be related to a reduction in the use of sentiment related terms by those who are not current users of the educational services. More investigation would be required to answer the question as to whether one or many of these characteristics were responsible for the difference in sentiment displayed. What is notable about this finding is that, clearly, spatio-temporal trajectories are providing a viable mechanism to segment the Twitter users into different user communities that are expressing themselves differently.

If we refer to Figure 2 and Figure 4, looking at weekday sentiment density and sentiment polarity, we see a general trend across the week with a heightened sentiment and an increase in sentiment polarity, expressed as a decrease in polarity magnitude, midweek with a peak on Tuesday. This trend is magnified in the tweets from educational services. In Figure 2 there is a normalization of sentiment density on the weekends as well. This could be significant in that it is another time where the education services are not being used that re-normalizes to the levels of other areas. Sentiment polarity, however, does not show this trend.

Figure 3-5, presents interesting information about sentiment polarity. The early hours have to be disregarded as the number of tweets at that time is minimal (often in the double digits) and does not lend itself to analysis. The later hours, however, are just as unpredictable with sentiment polarity by hour seeming to edge ahead in magnitude after 8 am in education services and stay that way but the amount of difference is not consistent between the two. Taking into account Figure 4 as well it seems that the majority of tweets have a more negative sentiment polarity in education services during work hours. However, they have a higher sentiment density. This is indicative that those in education areas are more likely to express sentiment in general. More specifically they are more likely to express negative sentiment.

This claim is reinforced by Figure 5 which shows a much less stable sentiment in the education services areas. However, this instability could also be accounted for by the fact that there are fewer people in the education services areas than everywhere else. This can be seen as sentiment score stabilizes towards the middle of the graph (closer to mid-afternoon). This technique would be more useful in comparing two similarly populated areas than one area (education) to the general population.

Future Work

There are many avenues of future work for both the dataset and the application presented. Comparing other areas of interest, as was mentioned previously, would be a good use of the methods devised, particularly if the two sets of tweets are similar populations.

On top of this, given more computational power and time, a classifier could be used in order to improve the results, using the lexicon as a starting point for sentiment. The lexicon itself could also be iteratively refined and better related to terms used on Twitter (due to the many abbreviations, hyphenations, misspellings, and acronyms used). The location and time of day could be used as inputs to the sentiment to determine if there was anomalous activity for the area. Implementing a stream mining architecture on top of the co-occurrence matrices could lead to real-time detection in sentiment changes.

Following users from area to area might be interesting as well. Specifically, of users that move from within education services areas to non-education services areas might have their sentiment change dependent on factors other than the area of interest. However, if these users stop going to these areas it might be interesting to measure their changed sentiment against the established pattern.
Going further, sentiment about topics could be used to determine spatio-temporal changes in sentiment based around events. This could be very useful for political topics and the like.

The work we have built on and the analyses we have performed have assisted in allowing many different possible avenues of future research of the Twitter data collected.

Conclusion

Here we classify the sentiment of Twitter users using trajectory based attributes to generate the aggregate moods of different sub-populations living similar lives. We demonstrate that spatio-temporal trajectories can be used as a suitable mechanism for generating sub-communities of Twitter users to further evaluate.

By investigating the spatial and temporal changes in emoticon and word use of these communities, future research could help civic leaders understand the impacts specific events are having on large groups of individuals.

For example, closing a school permanently may yield higher usage of negative emoticons related to the event for trajectory based communities using education resources in the city; but, individuals who are not users of the educational resource may not have any expressed change in sentiment. We uncover a general trend between general population and those who use education services that may indicate that certain sub-populations may express sentiment differently. In particular, those in education areas tend to express more sentiment in their tweets (positively and negatively) than those who do not take advantage of these services. In addition to this, the absolute sentiment polarity tends to be more negative in education areas as compared to the areas not using these services.

In conclusion, although general trends have been uncovered, new work may unlock latent sentiment structure by incorporating metadata ascribed by spatio-temporal analysis. Further research is required for more specific analysis of these sub-populations in order to determine reasoning behind the sentiment differences uncovered here.

Acknowledgments

The authors would like to thank Junjun Yin and Aiman Soliman for providing a suitable dataset for this investigation. Also, the authors thank the Cyber Infrastructure and Geospatial Information Laboratory and CyberGIS Center for Advanced Digital and Spatial Studies for their help as well. Also, we acknowledge the instructors of the CS512 course to whom this report was prepared for.

References


