Measuring Happiness of US Cities by Mining User-generated Text in Flickr.com: A Pilot Analysis

Sukjin You
School of Information Studies,
University of Wisconsin-Milwaukee
2025 E Newport,
Milwaukee, WI
yous@uwm.edu

Joel DesArmo
School of Information Studies,
University of Wisconsin-Milwaukee
2025 E Newport,
Milwaukee, WI
jdesarmo@uwm.edu

Soohyung Joo
School of Information Studies,
University of Wisconsin-Milwaukee
2025 E Newport,
Milwaukee, WI
sjoo@uwm.edu

ABSTRACT

This poster describes a methodology to numerically represent the happiness of a city by mining user generated terms in Flickr.com. As a pilot analysis, we collected 15,000 text records consisting of titles, tags, descriptions, and comments for the thirty most populous cities in the United States. Parsed text was utilized to calculate happiness scores (H-Score) by matching text extracted from Flickr.com with a happiness index dictionary. In addition, we examined the relationships between the calculated H-scores and real world phenomena including population, crime rate, and climate. Based on this pilot analysis, a future study is planed that involves a large dataset with prediction analysis.

Keywords

Flickr.com, social media mining, happiness index, text mining.

INTRODUCTION

Social media mining is an emerging field whose applications are widespread. Various text mining techniques make it possible to find secondary information from users' communications over social media sites. In particular, social media mining has been used to explain different social phenomena in the real world. For example, text from social media sites has been utilized to predict political outcomes (Lee et al., 2013), movie box office revenues (Asur & Huberman, 2010), disaster planning (Yin et al., 2012), disease outbreaks (Culotta, 2010), and to analyze public sentiment to predict the stock market (Bollen, Mao, & Zeng, 2011). Moreover, social media data imply user behavior and researchers have tried to model everyday behavior of a certain user group. For example, Twitter text was analyzed to describe the general behavior patterns of a group of people engaged in various ativities, such as when they are eating, working, or shopping (Lee, Wakimiya, & Sumiya, 2013).

Sentiment analysis has been also applied in the text analysis of social media. As social media data include users' real words, it typically contains sentiment information that represents the negative and positive status of specific objects. Researchers have tried to classify social media postings based on their sentiment characteristics. Popular text mining techniques, such as Naive Bayes and Maximum Entropy models, have been frequently employed in sentiment analysis of social media text (Parikh & Movassate, 2009). In a similar line, Wang et al. (2011) explored different relationships of Twitter hashtags at the sentiment level, and proposed a method of sentiment classification of topics.

In this way, social media mining has become a popular method to predict future outcomes or to gain a better understanding of the relationships that exists in various social phenomena. Sentiment analysis can be carried out since the nature of social media text involves users' sentiment. However, there is relatively less research that predicts real world phenomena based on sentiment analysis. This poster describes our methodological framework that intends to provide various prediction models based on social media mining. We plan on the analysis of user generated text of Flickr.com and suggest different prediction models. In this pilot analysis, we tested the calculation method of happiness score with a small dataset of user terms in Flickr.com. In addition, we tested an examination of the relationships between the sentiment information encoded within Flickr data and crime rates, population, and climate data.

METHODOLOGY

Data Collection

For the pilot test purpose, we generated a small set of data in this pilot analysis. Based on the US Census, we selected the top thirty most populated cities as of July, 2012 for the test dataset (US Census Bureau, 2012). Using the selected

ASIST 2013, November 1-6, 2013, Montreal, Quebec, Canada.
Copyright notice continues right here.

cities as a keyword, user text from Flickr.com was harvested. Two Flickr APIs were used to collect the following information: (1) titles, tags, and descriptions (flickr.photos.search); and (2) comments for each picture (flickr.photos.comments.getList). For each city, the first 500 results of the API, sorted by date posted after Jan 1, 2013, were used in this study. In total, 15,000 records of titles, tags, descriptions and comments were collected (500 records per city * 30 cities).

The R open-source software (www.r-project.org) was used to prepare the data and create a term-document matrix in order to determine a happiness index score for each city. The corpus was established by gathering the title, description, tags, and user comments together for each of the thirty cities in the study. Prior to creating the term document matrix, the data was cleansed by removing all punctuation, numbers, white space, and stopwords. The stopwords removed are a standard set of English stopwords derived from a built-in function of the R software "tm" module.

The other side of dataset consists of real world phenomena. As a pilot text, we selected crime rate, population, and climate data. As to crime rate data, the FBI website was consulted to collect violent crime data for the thirty cities in the study. Data from 2011 was the most recent data that was able to be obtained (FBI, 2011a). The FBI dataset provides data for both the Metropolitan Statistical Area as well as for the cities proper. For this study, the crime data for the city proper was used. The FBI defines violent crime as "composed of four offenses: murder and nonnegligent manslaughter, forcible rape, robbery, and aggravated assault. Violent crimes are defined in the UCR Program as those offenses which involve force or threat of force" (FBI, 2011b, para 2). For the climate data, we used data from the National Climate Data Center, which provides the annual average high and low temperature for US cities based on a the years 1981-2010. The dataset also contains precipitation data for cities, which includes the average number of days per year of precipitation as well as the average total inches of precipitation per year (National Climatic Data Center, 2011). Additionally, US Census data was used for population information (US Census Bureau, 2012).

Data Analysis

The most fundamental step of this study is to calculate the sentiment score of each city based on Flickr text mining. We decided to use a sentiment term dictionary to numerically represent a happiness level of a certain city. The labMT 1.0 dictionary of happiness index created by Dodds et. al (2011) was selected for our study. The labMT 1.0 dictionary was developed by using the crowdsourcing power of Amazon's Mechanical Turk. The researchers built a corpus composed of 10,222 unique words derived from a synthesis of the 5,000 most frequently occurring words in Twitter, Google Books (English), music lyrics (1960 to 2007), and the New York Times (1987 to 2007) (Dodds et. al, 2011). For each word in the corpus, the average of 50 independent 9-point scale sentiment evaluations was used to provide a numeric value.

The term-document matrix of Flickr text was matched with the happiness dictionary score. Since the size of collected terms varied by cities, the frequency was standardized to a percentage. The extracted terms were matched with the happiness scores of the dictionary and were then transformed to z-scores for standardization. We calculated the happiness score, H-Score, for each city by multiplying the standardized term frequency and the z-transformed happiness index. Then, the relationships between calculated H-Scores and selected associated variables were analyzed based on correlation coefficients (Pearson r). The selected associated variables are: population, population density, crime rate, and precipitation. Due to the limited sample size, prediction models, such as multiple regression or kernel regression, are not yet tested.

PRELIMINARY RESULTS

This poster presents preliminary results with the limited data as a pilot test. Table 1 shows the H-Scores calculated for each city. In this pilot analysis, two cities, Nashville and Phoenix, which had less than 2,000 terms matching with the dictionary, were excluded. The text collected from Charlotte, Louisville, Columbus, Portland, and Denver ranked 1st to fifth respectively in regards to positive terms. On the opposite end of the spectrum, New York City, San Antonio, and San Jose showed the lowest happiness term score. In particular, negative terms were most observed in user generated text related to the cities of San Antonio and San Jose.

Table 1. Happiness term scores for the selected cities

City	H-Score	Rank	City	H-Score	Rank
Charlotte	0.914	1	Memphis	0.536	15
Louisville	0.913	2	Philadelphia	0.532	16
Columbus	0.750	3	Fort Worth	0.503	17
Portland	0.713	4	Dallas	0.474	18
Denver	0.711	5	Chicago	0.469	19
Seattle	0.679	6	Houston	0.426	20
San Francisco	0.664	7	Baltimore	0.418	21
Austin	0.600	8	Detroit	0.405	22
San Diego	0.594	9	El Paso	0.399	23
Jacksonville	0.592	10	Washington D.C	0.391	24
Los Angeles	0.582	11	Indianapolis	0.369	25
Oklahoma City	0.580	12	New York City	0.334	26
Boston	0.554	13	San Antonio	-0.128	27
Milwaukee	0.551	14	San Jose	-0.313	28

More importantly, this study plans to compare happiness term occurrence with various variables related to cities. This study selected eight variables as examples. As shown in Table 2, Pearson r coefficients were computed between happiness term scores and the selected variables. Population and population density are negatively associated with happiness term scores. Also, crime rate has a slight negative relationship with H-Scores. Temperature showed negative association with happiness term scores, while precipitation exhibited a positive relationship.

Table 2. Relationships between happiness term scores and population, crime rate, and climate

Correlation with H-Score	Pearson r
Population	-.173
Population density	-.128
Crime rate	-.008
Average high temperature per year	-.215
Average low temperature per year	-.367
Average temperature per year	-.295
Inches in year precipitation	.191
Number of days of precipitation per year	.357

CONCLUSION

This study is an initial attempt to investigate user generated text in Flickr.com to reveal relationships between social media text and real world phenomena based on sentiments. This pilot test shows that happiness term scores are calculable based on the analysis of user terms collected from Flickr.com. As happiness level can be represented by a numerical index, we can further analyze the relationships between social media text and other real world data, particularly using prediction models.

This study establishes a methodology for using social media mining and sentiment analysis to develop an understanding of the factors that contribute to the well-being of urban cities. This information could be valuable to people seeking to relocate, such as retirees or job-seekers, as well as to policy makers working to improve the well-being of their communities. By unobtrusively using user generated text, hidden relationships between aspects of city life such crime and climate and the happiness of those places can be revealed that may not always be intuitive.

Because of its limited data size, the preliminary results presented in this poster are not yet reliable. Again, the purpose of this poster is to introduce the methodology to produce prediction models using Flickr data.

FUTURE STUDY

This pilot study describes the methodology we are going to use for a larger study. Our major research idea is to empirically prove whether user generated text from social media sites would be useful to estimate various phenomena in the real world. In the future study, of course, our sample will be enlarged to better represent user wordings in social media. We plan to collect about 30,000 records per city including titles, description, tags, and comments. More importantly, prediction models will be employed, especially multiple regression and nonparametric regression. Additionally, a wide variety of text mining techniques will be applied such as clustering, semantic analysis, and topic modeling.

REFERENCES

Asur, S., & Huberman, B. A. (2010). Predicting the Future with Social Media. In 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (Vol. 1, pp. 492–499). Presented at the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT).

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1–8.

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. In Proceedings of the First Workshop on Social Media Analytics (pp. 115–122). New York, NY, USA: ACM.

FBI. (2011a). Crime in the United States: by Metropolitan Statistical Area, 2011:Table 6. FBI. Retrieved June 26, 2013, from http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2011/crime-in-the-u.s.-2011/tables/table-6

FBI. (2011b). Offenses Known to Law Enforcement. FBI. Retrieved June 26, 2013, from http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2011/crime-in-the-u.s.-2011/offenses-known-to-law-enforcement/offenses-known-to-law-enforcement

Lee, J., Ryu, H., Mon, L., & Park, S. J. (2013). Citizens' Use of Twitter in Political Information Sharing in South Korea. In iConference 2013 Proceedings (pp. 351-365)

Lee, R., Wakamiya, S., & Sumiya, K. (2013). Urban area characterization based on crowd behavioral lifelogs over Twitter. Personal Ubiquitous Computing, 17(4), 605–620.

National Climatic Data Center. (2011). NOAA's 1981-2010 Climate Normals. Retrieved June 26, 2013, from http://www.ncdc.noaa.gov/oa/climate/normals/usnormals.html

Parikh, R., & Movassate, M. (2009). Sentiment analysis of user-generated twitter updates using various classification techniques. CS224N Final Report.

US Census Bureau. (2012). City and Town Totals: Vintage 2012. Population Estimates. Retrieved June 19, 2013, from http://www.census.gov/popest/data/cities/totals/2012/index.html

Wang, X., Wei, F., Liu, X., Zhou, M., & Zhang, M. (2011, October). Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 1031-1040). ACM.

Yin, J., Lampert, A., Cameron, M., Robinson, B., & Power, R. (2012). Using Social Media to Enhance Emergency Situation Awareness. IEEE Intelligent Systems, 27(6), 52–59.