Spring 2015

From surveys to surveillance strategies: a case study of life satisfaction

Chao Yang
University of Iowa

FROM SURVEYS TO SURVEILLANCE STRATEGIES: A CASE STUDY OF LIFE SATISFACTION

by

Chao Yang

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Computer Science
in the Graduate College of
The University of Iowa

May 2015

Thesis Supervisor: Professor Padmini Srinivasan

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Chao Yang

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Computer Science at the May 2015 graduation.

Thesis Committee: _____
Padmini Srinivasan, Thesis Supervisor

_____
Nick Street

_____
Alberto Maria Segre

_____
Philip Polgreen

_____
Kang Zhao

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Padmini Srinivasan for supporting me during these past five years. She motivated and encouraged me to start my research and gave me great guidance on various research problems.

I would like to thank my committe members, Professor Nick Street, Professor Alberto Segre, Professor Philip Polgreen, and Professor Kang Zhao for their valuable feedback and comments on my research.

I would like to thank the colleagues from the Text Retrieval and Text Mining group. I would especially like to thank Yelena Mejova for guiding me start the social media research, Hung Tran for introducing me Python and initial code of using Twitter API, Chris Harris and Sanmitra Bhattacharya for their help and support.

I would also like to thank my internship supervisors/mangers Miguel Ruiz, Girija Yegnanarayanan (Nuance) and Shimei Pan, Jalal Mahmud (IBM Research). I have learnt a lot from them.

Last but not least, I would like to thank to my family especially my wife Yixin Chen for their love and support.

ABSTRACT

Social media surveillance is becoming more and more popular. However, current surveillance methods do not utilize well-respected surveys, which were established over many decades in domains outside of computer science. Also the evaluation of the previous social media surveillance is not sufficient, especially for surveillance of happiness on social media. These motivated us to develop a general computational methodology for translating a well-known survey into a social media surveillance strategy. Therefore, traditional surveys could be utilized to broaden social media surveillance. The methodology could bridge domains like psychology and social science with computer science. We use life satisfaction on social media as a case study to illustrate our survey-to-surveillance methodology. We start with a famous life satisfaction survey, expand the survey statements to generate templates. Then we use the templates to build queries in our information retrieval system to retrieve the social media posts which could be considered as valid responses to the original survey. Filters were utilized to boost the performance of the retrieval system of our surveillance method.

To evaluate our surveillance method, we developed a novel method to build the gold standard dataset. Instead of evaluating all the data instances like the traditional way, we ask human workers to "find" as many of the positives as possible in the dataset, the rest are assumed to be negatives. We used the method to build the gold standard dataset for the life satisfaction case study. We also build three more

gold standard datasets to further demonstrate the value of our method. Using the life satisfaction gold standard dataset, we show that performance of our surveillance method of life satisfaction outperforms other popular methods (lexicon and machine learning based methods) used by previous researchers.

Using our surveillance method of life satisfaction on social media, we did a comprehensive analysis of life satisfaction expressions on Twitter. We not only show the time series, daily and weekly cycle of life satisfaction on social media, but also found the differences in characteristics for users with different life satisfaction expressions. These include psychosocial features such as anxiety, anger and depression. In addition, we present the geographic distribution of life satisfaction, including the life satisfaction across the U.S. and places around the world. This thesis is the first to systematically explore life satisfaction expressions over Twitter. This is done using computational methods that derive from an established survey on life satisfaction.

PUBLIC ABSTRACT

Social media surveillance is becoming more and more popular. However, current surveillance methods do not utilize well-respected surveys, which were established over many decades in domains outside of computer science. Also the current social media surveillance methods are not accurate enough. These motivated us to develop a general computational methodology for translating a well-known survey into a social media surveillance strategy. Therefore, traditional surveys could be utilized to broaden social media surveillance. The methodology could bridge domains like psychology and social science with computer science.

We use life satisfaction on social media as a case study to illustrate our survey-to-surveillance methodology. In addition, we developed a novel method to build the dataset to evaluate our surveillance method. We show the method of building the dataset is solid, and the performance of our surveillance method of life satisfaction outperforms other popular methods used by previous researchers.

Using our surveillance method of life satisfaction on social media, we did a comprehensive analysis of life satisfaction expressions on Twitter. We not only show the time series, daily and weekly cycle of life satisfaction on social media, but also found the differences in characteristics for social media users with different life satisfaction expressions. These include psychosocial features such as anxiety, anger and depression. In addition, we present the geographic distribution of life satisfaction. This thesis is the first to systematically explore life satisfaction expressions over Twitter.

This is done using computational methods that derive from an established survey on life satisfaction.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1

INTRODUCTION

In recent years, more and more people are using social media. The recent study by Duggan et al. [21] found that more than 70% of online adults use Facebook, and 50% of online adults use multiple social media sites. Therefore, passive surveillance of preferences, opinions and behaviors on social media is becoming increasingly common. The general goal is to make inferences from observations collected from the numerous posts publicly available in blogs, microblogs, and other social forums. Social media surveillance is different from the traditional surveillance. A traditional approach for collecting observations is to use surveys to query a random (or convenience) sample of individuals. The advantages of social media surveillance is that it can obtain more data very easily. It costs less time, has less bias, and it is cheaper. For example, we could passively surveil millions of social media posts every day. The posts are not only from one country but from all over the world. And the surveillance could be completely free (assuming the social media posts are public like Twitter).

Therefore, it is very important to have a method to build social media surveillance strategies. Since a wide variety of well respected survey instruments have been developed over many decades, especially in psychology and social sciences, a natural way of building social media surveillance is to directly use them to survey social media users. However, it is not currently practical to use surveys on social media users for different reasons. First, it is hard for surveys to reach social media users due to spam

detection algorithms on social media sites. Some websites like Facebook allow survey applications, but it is still hard to attract users to the apps. Second, the response rate of surveys from social media users is low (less than 1% response in our experiment, even we promised that they could win $50 gift card.) Typically online users do not want to take free surveys. The paid survey is too expensive to scale up to millions of participants. Additionally, surveys are static. They usually have fixed questions and cannot be changed easily.

While surveys are difficult to modify and difficult to administer to social media users, they are clearly well-respected instruments that carry decades of experience. Thus, we believe it is important to utilize such expertise and design surveillance strategies that are more appropriate for social media. The advantage in doing so is that the surveillance strategy will be founded on the principled approach that gave rise to the survey (assuming the survey is of good quality). The challenge faced is that it is not clear how to build a surveillance strategy from a survey of interest. Thus, the first question we ask in this thesis is: how does one "translate" a survey of interest into a surveillance strategy on social media? Specifically, how does one find the posts that could be interpreted as valid responses to the survey? Developing a general methodology for translating a survey into social media surveillance might further the inclusion of social media research into traditional social science research. In the methodological part of the thesis (Chapter 3), we develop a computational methodology that is able to translate a survey into social media surveillance. We use life satisfaction as a case study to illustrate our methodology.

To evaluate social media surveillance is not easy; the gold standard dataset for evaluation is hard to build because this kind of surveillance usually is large scale and has millions of data instances. Without a gold standard dataset for evaluation, one important metric "recall" is missing in previous related studies. A second major emphasis in this thesis is that we develop a novel method to build the gold standard dataset. We use the method to build the gold standard dataset for life satisfaction case study and three more gold standard datasets to further demonstrate the method (Chapter 4). Then we evaluate our surveillance method in terms of different metrics. We also compare our surveillance with other approaches which are popular in social media surveillance of happiness and show our surveillance outperforms them (Chapter 5).

The third contribution is that we design computational methods to observe expressions of life satisfaction on social media. Social media posts have been studied from many angles. For example, researchers explored different aspects of social media users. Those aspects include personalities [67], personal values (conservation, hedonism, etc.) [10], personal needs (challenge, love, etc.) [66], happiness [20], and depression [15]. One domain, which is our particular interest, is the area of Subjective Well Being (SWB). SWB is a well-recognized approach from social sciences to measuring happiness. It was formalized by Ed Diener [17] in the 1980s. SWB has two components: affect balance and life satisfaction [16]. Affect balance refers to the balance between positive and negative emotion, mood, or feelings of a person over short periods, such as a day. Life satisfaction is the stable and long term assessment

of one's life. We have observed computational methods proposed to study happiness, which is the first component of SWB. However, little to no work has been done with life satisfaction, which is also a key component of SWB. Our goal is to contribute to the stream of work studying SWB on social media.

Therefore, we combine our interest in designing surveillance strategies from surveys with our interest in studying life satisfaction. In particular, we take a well reputed survey on life satisfaction designed by Diener [18] and derive surveillance strategies using our translation method. In the life satisfaction analysis part of the thesis, we not only show the common statistics, such as time series, location, daily/weekly cycle of life satisfaction on Twitter (Chapter 8, 11), but also the differences of people who have different life satisfaction expressions (Chapter 9). In addition, we addressed the following questions: What factors associate with life satisfaction and life dissatisfaction (Chapter 10). We study especially psychosocial variables and associations over time.

CHAPTER 2

RELATED WORK

There are no previous studies developed a method to transform a traditional survey into a surveillance strategy. Therefore, we only introduce the related work about the Subjective Well Being (SWB) studies in this section.

People focus on their health more and more. Hospitals spend large effort monitoring patients' health status, also shifting the old paper-based records to Electronic Heath Record (EHR). The general public wears different devices like FitBit to track vital signs. Besides physical health, mental health is also an area that cannot be ignored. Bhutan's King Jigme Singye Wangchuck coined Gross National Happiness (GNH) for measuring countries' economy in 1972 because he was unsatisfied with the Gross Domestic Product (GDP) measurement. The UK government is also among the first countries to officially measure Subjective Well Being [55, 5]. The US constitution recognizes the pursuit of happiness as an unalienable right given to all human beings. Several organizations have been monitoring global happiness, well-being etc., such as with the Happy Planet Index[1], Gallup Well-Being Index [23, 24], and World Happiness Report [32]. Sites such as "Pulse of the Nation" present interesting visualizations of mood across the US over the course of a day[2].

Again, SWB has two components: affect balance and life satisfaction. Sur-

[1]http://www.happyplanetindex.org

[2]http://www.ccs.neu.edu/home/amislove/twittermood/

veys have been developed to surveil SWB. For example, Positive and Negative Affect Schedule (PANAS-X) has been used for evaluating affect balance. Satisfaction With Life Scale (SWLS) is a well-regarded survey for measuring one's life satisfaction (Details in Section 3.2).

There is a small stream of recent work studying the SWB on social media. However, these do not distinguish between affect balance and life satisfaction. Most of them used the term "happiness" generally to include both affect balance and life satisfaction. Some studies directly apply SWB surveys to social media users [60, 49]. Most of the research is at the general level, i.e. document level, not considering target subjects such as the happiness of the authors. Most of the studies used lexicon-based methods, most did not clearly assess the performance of their methods. A summary of happiness detection studies using online social network is shown in Table 2.1. In the rest of this chapter, we first introduce the studies that detect happiness using different datasets including blogs, Facebook, Twitter posts, etc. Then we discuss Dodds et al.'s text-based hedonometer which is most closely related to this thesis.

2.1 Happiness Studies With Blogs, Lyrics, etc.

In 2006, Mihalcea et al. [43] explored happiness and sadness in LiveJournal posts. Blog posts tagged with "happy" and "sad" moods in LiveJournal were collected. They built a Naïve Bayes classifier using unigram features to predict happiness and sadness. The accuracy of the method achieved 79%. The top most happy terms such as "yay," "shopping," "awesome;" top most sad words such as "goodbye," "hurt," "tears" were generated by the classifier. They did not use the classifier to

Table 2.1. Summary of happiness detection studies

Study	**Year**	**Corpus**	**Classes**	**Method**	**Super-vised Method**	**Performance**
Mihalcea et al. [43]	2006	Posts in LiveJournal	Happiness, Sadness	NB classifier using unigram features	Yes	0.79 (Accuracy)
Dodds and Danforth [19]	2010	Lyrics, Blogs, and State of the Union	Happiness	Utilized ANEW lexicon	No	N/A
Kramer [36]	2010	Facebook status	Positive, Negative sentiment	Word count using LIWC lexicon	No	r=0.17 with Gold standard
Bollen et al. [8]	2011	Tweets from Streaming API	Positive, Negative sentiment	Word count using OpinionFinder lexicon	No	N/A
Dodds et al. [20]	2011	Tweets from Streaming API	Happiness	Utilized labMT lexicon to get average valence	No	N/A
Quercia [50]	2012	London Twitter users and their tweets	Positive, Negative sentiment	Word count using LIWC and MaxEnt classifier	Both	0.66 (Precision)
Wang et al. [59]	2014	Tweets	N/A	POS and SentiWordNet lexicon	No	N/A
Hung et al. [34]	2014	Facebook status	7 emotions include happiness	SVM with PMI features	Yes	0.552 (F1)
Curini et al. [11]	2014	Italian Tweets	Happiness, Sadness	Statistical techniques	Yes	N/A

predict public happiness. Instead, they utilized the frequency of those salient words to surveil the public happiness. Then they examined users' seasonal cycles of happiness. They found that waking up appeared to be an unhappy moment, and that users are happiest at 3 AM and between 9 PM to 10 PM. Saturdays are the happiest and Wednesdays are the saddest.

Blogs were only one of the media in which researchers surveilled public happiness, they also examined other outlets of expressions. In 2010, Dodds and Danforth [19] measured happiness in song lyrics[3][4], blogs[5], and State of the Union messages (American Presidency Project[6] and British National Corpus[7]). To measure happiness, they used the ANEW lexicon, calculating the weighted average happiness score for all lexicon words in the text. Affective Norms for English Words (ANEW) is one of the most popular lexicons for sentiment analysis (Details in Section 5.2.1.2). They found that the average happiness for lyrics decreased after 1980 due to the loss of positive words and the gain of negative words. With blog data, there is generally an increase in happiness over the last part of each year. And happiness departs dramatically from the month's average on days like: Christmas Day; Valentines Day; September 11. Using the State of the Union address, they found President Kennedy, Eisenhower and Reagan have the highest average happiness scores. Hoover and Franklin have the

[3]http://www. hotlyrics.net

[4]http://www.freedb.org

[5]http://www.wefeelfine.org

[6]http://www. presidency.ucsb.edu

[7]http://www. natcorp.ox.ac.uk

lowest happiness scores, possibly because of the Great Depression and World War II respectively.

2.2 Happiness Studies With Facebook

Facebook is one important social media in which users tend to have their actual names and accurate information like gender, age, etc. We note that many happiness studies with Facebook were actually looking at life satisfaction (through the use of the SWLS survey), although the authors refer to happiness.

In 2010, Kramer [36] tried to model the GNH using Facebook. He utilized word count of LIWC lexicon for Facebook status updates to determine if they are positive or negative. Linguistic Inquiry and Word Count (LIWC) is a popular lexicon for sentiment analysis (Details in Section 9.3.3). Then he averaged all the status scores for US posts each day to get the GNH of that day in the United States. To validate their method, he had more than one thousand Facebook users participate in the SWLS survey. His method weakly correlates with the survey (r=0.17). In 2012, the Facebook Data Team released Facebook Gross National Happiness (FGNH[8]) [57] which is based on automated sentiment analysis utilizing the LIWC dataset. The FGNH index was available through the Gross National Happiness Facebook application for a short time but is currently not available. However, in 2012, Wang et al. [60] argued that the LIWC method could not accurately derive SWB from Facebook. They examined the validity of FGNH in measuring mood and well-being by comparing it with scores on their own SWLS surveys. They found the FGNH is actually slightly negatively

[8]https://apps.facebook.com/gnn_index

correlated with their survey results. Therefore, they concluded that FGNH is not a valid measure for SWB.

One dataset for analyzing users on Facebook is myPersonality [54]. It is a Facebook app which allowed users to take a series of tests including the SWLS survey, but it is closed in 2012. The number of users with SWLS survey results is not public.

In 2013, Quercia [49] explored the geography of happiness using Facebook. He collected SWLS test results from myPersonality. The survey results strongly correlated with the official well-being score of twelve rich countries. He found that the lower the happiness of a country, the greater the problems in homicide, obesity, drug use, mental illness, and anxiety. Interestingly those problems were not associated with absolute levels of income. He provided hints that social media could be used for data-driven social science research. In 2014, Hung et al. [34] built an integrated emotion regulation system (IERS) to detect 7 emotions including happiness from users' Facebook status. They utilized SVM classifier with Point-Wise Mutual Information (PMI) features. Most of the F scores of the 7 emotion detections are higher than 0.5.

2.3 Happiness Studies With Twitter

Although Facebook has the most accurate user information among social media providers, most of the data is not publicly accessible. However, Twitter has the largest publicly accessible social media posts (tweets). The social media site is becoming one of the most popular resources for text mining research including happiness studies. In 2009, Kim et al. [35] detected sadness about Michael Jackson's death on Twitter. They collected tweets one day before and about ten days after MJ's death using

search queries like "MJ," "Micheal Jackson," etc. They used ANEW lexicon to do the sentiment analysis. They showed that people tweeted a lot after they heard about Michael Jackson's death. Also they found that tweeting about MJ's death used many more negative emotion words. 3/4 of the tweets have the word "sad." They sampled more than three hundred tweets containing "sad," and found about 75% of them actually express sadness by human evaluation. In 2011, Bollen et al [8] found that happiness is assortative in Twitter. They collected tweets from streaming API for about half a year. They built a Twitter network using more than four million Twitter users with their following and followers. A tweet's happiness status was detected by calculating the number of positive and negative words using the OpinionFinder lexicon. One user's happiness status was assumed to be all his/her tweets' aggregated tweet happiness status. They found two connected users have similar happiness values. They also found an individual is influenced by the overall happiness of all of the people he/she interacts with.

In 2012, Quercia [50] studied the relationship between Gross Community Happiness in tweets and community socio-economic well-being by investigating Twitter users in London census communities. They crawled the accounts of more than 500 Twitter users in London neighborhoods. Also the Index of Multiple Deprivation (IMD) score of each of the 78 census areas in London was collected. IMD is a composite score based on income, employment, education, health, crime, housing, and the environmental quality of each community it indicates. They developed two methods: Word Count of words using LIWC lexicon and Maximum Entropy (ME) classifier

to measure tweet sentiment. The classifier was trained using another automatically labeled tweets dataset using "smiley" and 'frowny" emoticons. Tweets with smiley and frowny faces were considered as ground truth for the ME classifier. Precision for both methods were about 66%, Recall for Word Count was 38% and for MaxEnt was 68%. They calculated GCH score for communities and calculated Pearson correlation coefficient between IMD and GCH. They found the higher the sentiment score of a community's tweets, the higher the community's socio-economic well-being.

In 2014, Wang et el. [59] suggested a formula for calculating the Gross National Happiness using POS tagging and SentiWordNet lexicon on tweets. They suggest their method has higher accuracy than using lexicon only. In the same year, Curini et al. [11] examined the happiness of tweets from Italian users on a daily basis in all the 110 Italian provinces. Then they tried to find which variables affect the average level of happiness in Italian provinces. They found meteorological variables and events related to specific days have the largest impact.

2.4 Happiness Research by Dodds et al.[9]

One important work about happiness research was conducted by Dodds et al. [20] in 2011. They did a survey on Amazon Mechanical Turk (MTurk)[10] to obtain happiness evaluation of more than 10,000 selected words (common English words from Twitter, Google Books, etc.) in their labMT lexicon. For example, the average happiness score for "Laughter" is 8.5/10; the score for "Hat" is 2.34/10. For a given

[9]We put their study in a seperate section because it is most closely related to our interests

[10]https://www.mturk.com

text such as a tweet, they compute its average happiness score across all of its lexicon words. Then one day's happiness score for Twitter is the average happiness score of all the tweets on that day. They call their method a text-based hedonometer. They used the hedonometer to get the overall happiness score for Twitter in a 2-year-period, computing daily cycles and weekly cycles of happiness. The also explored the geography of happiness. Recently, Dodds' et al. show their results on their website[11].

In 2012, Frank et al. [22] adopted Dodds' labMT lexicon and hedonometer, utilized the GPS tag of tweets, and found that expressed happiness increases logarithmically with both distance from expected location and radius of movement. Individuals with a large radius use happier words than those with a smaller radius of movement. The same year, Bliss et al. [7] constructed and examined the reciprocal-reply networks in Twitter over the time scales of days, weeks, and months. They also adopted Dodds' labMT lexicon and hedonometer to investigate happiness expressions. They found the users' average happiness scores are significantly correlated with the scores of the neighbors who are one, two, and three degree of relationship away. They also found that users who have more connections write happier tweets. Therefore, they concluded that "happiness is assortative."

In 2013, Mitchell et al. [45] estimated the happiness levels of states and cities with Twitter using Dodds' labMT lexicon and hedonometer. They found that happiness levels correlated with most well-being measures such as: Behavioral risk factor

[11]http://hedonometer.org

survey score (BRFSS)[12], 2011 Gallup well-being index[13], 2011 United States peace index[14] and 2011 United Health Foundations Americas health ranking (AHR)[15]. They also found the happiness levels in Twitter were anti-correlated significantly with obesity rates and the number of shootings per 100,000 people in 2011. Thus, they concluded that social media may potentially be used to estimate happiness in real-time.

2.5 Limitation Of Happiness Studies With Social Media

More and more researchers are interested in the studies of surveillance of SWB using computational methods, which is also our interest. We have introduced the happiness studies with social media in the previous sections. The comprehensive survey for happiness studies in psychology and using computational methods is addressed in my comprehensive exam paper [64]. We found that most of the methods of happiness surveillance including Dodds' text-based hedonometer are not perfect. The limitations of current studies of happiness with social media are listed below. These motivate the current survey-to-surveillance methodology in this thesis research. Specifically, most of their research:

- focused on affect and ignore life satisfaction. The three exceptions are [36, 49, 60].
- did not focus on the author's happiness only. The analyzed text may only reflect

[12]http://www.cdc.gov/brfss/

[13]http://www.well-beingindex.com/files/2011CompositeReport.pdf

[14]http://www.visionofhumanity.org/info-center/us-peace-index/

[15]http://www.americashealthrankings.org/Reports

author's relatives, friends' happiness, even their pets' happiness.

- did not filter out the past, future tense expressions, or questions related to interrogations of happiness.
- did not have an efficient way to handle negation.
- did not fully evaluate the performance of their methods for detecting happiness.
- is not founded upon a formal and theoretically established survey of happiness (such as SWLS [18] and PANAS-X [62]).

Thus our goal in this thesis is to study life satisfaction expressions on Twitter with a methodology that overcomes the above limitations.

CHAPTER 3

SURVEY-TO-SURVEILLANCE METHODOLOGY

We have discussed many advantages of social media surveillance in the Introduction. We also discussed that using existing surveys directly for social media surveillance is difficult. In addition, the SWB surveillance studies on social media do not distinguish affect balance and life satisfaction, and the previous studies also have methodological limitations. Therefore, we were motivated to develop a survey-to-surveillance method to bridge the traditional survey to social media surveillance. We use life satisfaction surveillance as a case study to illustrate our survey-to-surveillance method.

In this chapter, we describe how we extended the traditional survey to surveillance strategy of life satisfaction on social media. We started with a famous survey in this domain and utilized information retrieval technology to extract the tweets which could be the valid feedback of the original survey. To the best of our knowledge, we are the first to build a method to move from surveys to social media surveillance.

3.1 Key Definitions

In this thesis, we use the term "LS tweet" to indicate a tweet that says something about the Twitter user's level of satisfaction with his/her life, including both the positive end, i.e., satisfaction with life (Class S tweets) and the negative end, i.e., dissatisfaction with life (Class D tweets). Tweets which are not about life satisfaction we refer to as "Class I" tweets (for irrelevant).

Table 3.1. Satisfaction with life scale.

Statement
1. In most ways my life is close to my ideal.
2. The conditions of my life are excellent.
3. I am satisfied with life.
4. So far I have gotten the important things I want in life.
5. If I could live my life over, I would change almost nothing

The users posting "LS tweets" is called "LS users." LS users also consist of two groups: Class S users and Class D users. Note that users can be in both classes, those are the users who changed their life satisfaction expressions. (Chapter 9.2)

3.2 Start With Traditional Survey

Our goal is to find all tweets that report on personal life satisfaction or dissatisfaction. We started in a principled way with a highly reputed self-assessment scale in psychology called the Satisfaction With Life Scale (SWLS) which was designed in 1985 by Diener [18] et al. The scale has been cited more than 9,800 times and used in many areas, including studies in psychology [2] and social media [37]. The scale has five statements as shown in Table 3.1. Notice that the SWLS survey deliberately steers away from specifying causes for happiness, such as pets, home, travel, spouse, etc. A respondent is asked to self-rate for each statement on a scale from 1 to 7 (Strongly Disagree (1); Disagree; Slightly Disagree; Neither Agree or Disagree; Slightly Agree; Agree; Strongly Agree (7)). Our goal is to find tweets can be seen as valid responses to the survey. Those tweets are the LS tweets of interest.

Unfortunately it is far from sufficient to search for these statements directly on Twitter. To illustrate, the expressions "*I am utterly content with my existence*" and "*I*

have achieved all my goals" share no non-trivial words with the scale statements. Yet both communicate satisfaction with life and are synonymous with statements 3 and 4 respectively. Given the many possible synonymous expressions, we have developed a process that takes expressions in the scale and generalizes them into expression templates. Each template is then transformed into a set of search queries that may be applied to a tweet collection. Finally, we use an information retrieval system to retrieve the LS tweets using the search queries. In this manner, the survey can be transformed into a surveillance strategy.

Fig 3.1 shows this process. While we focus on the SWLS, the process is also general enough to apply to other scales such as CES-D (Center for Epidemiological Studies Depression Scale) [51].

3.3 Synonymous Expressions

In the beginning of the translation process, we need to collect more expressions which express a similar sentiment to the survey statements. We obtained an initial set of synonymous expressions through crowdsourcing with MTurk. Ten workers were each asked to provide 20 alternate expressions (pseudo tweets) for each statement. They were given a few examples: for instance, for statement 3: "*I am very happy with my life*," "*I have a satisfying life*" and "*My life is the way I want it to be*." Thus, we obtained a maximum of 1,000 alternate LS expressions. It took less than five days to obtain this set and cost less than ten dollars. Any other crowdsourcing platforms could be sufficient for this.

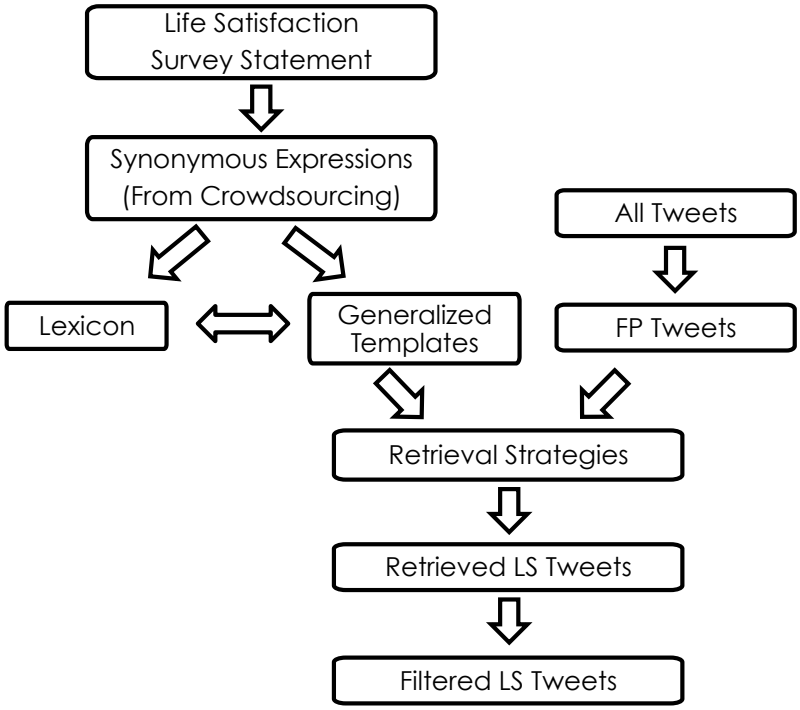


Figure 3.1. Flowchart of survey-to-surveillance method

3.4 Generalized Templates And Lexicons

Next, we manually generalize the statements and expressions into templates which can be utilized as search queries for the information retrieval system. A template is a sentence with optional variables referring to entries in a lexicon of functionally synonymous terms. The templates and lexicon were developed simultaneously. Our lexicon consists of 70 synonym sets. We also developed negative versions of the templates to retrieve Class D tweets as well. The total number of templates is 778. The full lexicon may be obtained by contacting the authors. Table 3.2 shows three example expressions, the generalization of each into a template and the corresponding lexicon entries. Phrase entries are given in parentheses.

For example, the expression "*I live a perfect life*" could be considered as three

Table 3.2. Sample expressions, templates, and lexicon entries

Phrase entries are in parentheses. Lexicon variable names carry no meaning.
Example 1: Expression: My life is perfect. → Expression Template: my X Y
Lexicon entry X: {life's (life is) (life has) (life has been) (life's been) (life has always been) (life's always been) etc...}
Lexicon entry Y: {amazing adorable awesome beautiful best (the best) blessed bliss blissful brilliant comfortable comfy contended delightful desired dream enjoyable exemplary excellent exciting fabulous fantastic fine flawless fulfilled fulfilling (full of joy) glorious good gorgeous grand gratifying great greatest happy heavenly ideal idyllic incredible joyous love (full of love) lovable lovely magical outstanding peaceful (picture perfect) perfect perfection pleasing super superb splendid etc...}
Example 2: Expression: I live a perfect life. → Expression Template: A Y B
Lexicon entry A: {(I have been living) (I've been living) (I am living) (I'm living) (I live) (I have been having) (I've been having) (I am having) (I'm having) (I have) (I have been leading) (I've been leading) (I am leading) (I'm leading) (I lead) (I have been getting) (I've been getting) (I am getting) (I'm getting) (I get) (I have got) (I've got) (I have gotten) (I've gotten) etc., ..}
Lexicon entry Y: {Shown Above.}
Lexicon entry B: {life existence}
Example 3. Expression: I love my life. → Expression Template: D E F
Lexicon entry D: {(I've (I have) (I have been) (I've been) (I've been having) (I have been having) (I am) I'm Im (I'm having) (I am having) I...}
Lexicon entry E: {like liking love loving adore adoring enjoy enjoying etc.. }
Lexicon entry F: {life existence (my life) (this life) myself (my existence) (this existence) etc...}

parts: "*I live*," "*perfect*," and "*life*." Therefore, we could generalize this expression into template: A Y B. (Definitions are in the table.)

3.5 Templates To Retrieval Strategies

Our next step is to build a set of retrieval strategies (search queries) from templates. Table 3.3 illustrates the 12 kinds of queries we built from each template. The rows indicate the word gap, which is the number of other words allowed between the words of a template. The columns indicate the number of words (W) allowed

Table 3.3. 12 retrieval strategies derived from a expression template.

Entries in [] and {} may be substituted by lexicon entries for set X and set Y respectively.

Template: **My [X] {Y}** Expression: [My life is] {perfect}.	**W1** **(0 or 1 word allowed before and after)**	**W2** **(2 words allowed before and after)**	**W3** **(3 words allowed before and after)**
A **(word gap = 0)**	[my life is] {perfect}!	I feel [my life is] {perfect}!	I can say [my life is] {perfect} and no less.
B **(word gap = 1)**	[My life is] really {perfect}, truly.	[My life is] really {perfect} so far.	No one doubts [my life is] really {perfect}.
C **(word gap = 2)**	Happily [my life is] really quite {perfect}.	For sure [my life is] really quite {perfect}.	See [my life is] completely and truly {perfect} for good.
D **(word gap = 3)**	See [my life is] completely and truly {perfect}!	See [my life is] completely and truly {perfect} for good.	[My life is] completely and truly {perfect} thats the truth!

before or after the text segment that satisfies the template. Limiting W is a conservative strategy designed to limit the length of the retrieved tweets. For example, B-W2 allows one word gap in the statement and two words before and after the statement. Using one example "*My life is perfect.*" from the template, we could retrieve expressions like "*My life is really perfect so far.*"

These 12 strategies are designed as a "cascading" set. That is, if a tweet is retrieved by multiple strategies, then it is considered retrieved by the most restrictive query. The strategies are less restrictive (which increases recall but potentially risks precision) as one moves down a column or to the right on a row. To compensate for possible losses in precision, we applied a set of quality filters to the retrieved set as explained next.

Table 3.4. Filters

Filter	**Method**	**Example Tweets Filtered**
Remove tweets ending with irrelevant words	Words referring to pets, and other objects are in a stoplist	I am happy with my cat.
Remove tweets referring to past or future tense.	Words referring to specific times such as "will," "was," "now," ... are in a tense stoplist	party! I was happy
Don't cross sentence boundaries when filling a template	Check for sentence boundaries	I am here. Happy birthday!
Remove tweets referring to 3rd person.	"you," "his," "her," ... are in a stoplist	I think you are a happy person
Remove tweets asking a question	Check for question mark	I am happy?

3.6 Quality Filters For Retrieved Tweets

We used a set of filters to a) remove irrelevant tweets and to b) identify an important subset of Class D tweets. For example strategy A-W3 could also retrieve "*My life is amazing because of my cat.*," which is not relevant. As said earlier, the SWLS survey does not survey the causes for happiness, such as pets, home, travel, spouse, etc. We are also not interested in tweets that refer to the future or the past ("*I so used to be happy*"). Most Class S tweets with negations are automatically added to our set of Class D tweets. More straightforward examples for this are: "*My life is not perfect*" and "*I am not happy.*" Table 3.4 shows the filters used. Also notice that we applied the method only to first person (FP) tweets (details are in Chapter 8). The FP tweets dataset is naturally another filter in order to find expressions that approximate self-ratings.

In summary, we start with a well-know survey and transform it to a surveillance

strategy. The surveillance strategy is based on an information retrieval technology. It only surveils the author's life satisfaction. It has filters to deal with different tenses, question, negation, etc. Next, we need to build a gold standard dataset (Chapter 4) and test the performance of our surveillance strategy (Chapter 5).

CHAPTER 4

BUILDING A GOLD STANDARD DATASET

To evaluate an automatic detection algorithm, we need to apply the algorithm to a dataset in which all the data points are labelled. Then we check the difference between the prediction of the algorithm and the gold standard label in the dataset for each data point in order to evaluate the performance of the automatic algorithm. Such a completely labelled dataset is called a gold standard dataset.

It is very important to get the gold standard dataset right. There are different ways of building the gold standard dataset. Some researchers use hand labeling by experts [25, 42]. Some researchers use crowdsourcing services like Amazon Mechanical Turk [46, 14, 13]. Using crowdsourcing, researchers can put the data online, and request workers to label the data for them and pay them. Some researchers utilize emoticons or emotion hashtags to automatically label data [26, 14, 68, 13, 3]. Some others utilize blogs already tagged with emotion by authors, or movie & product reviews rated by customers [38, 4, 44, 43, 47, 33].

For some datasets, the positive signals are very rare. It is then hard to adopt the traditional ways to build a large gold standard dataset. This is because it is impractical to ask annotators to label all the data instances in the dataset. One of the approaches call "Pooling" is what the microblog track in TREC (Text REtrieval Conference) does. There the task is to find tweets related with some topics automatically from millions of tweets. Instead of labeling the full dataset they pool all

the tweets submitted from different participant groups and annotate them by highly paid workers. Then the precision-based metrics can be used to evaluate performance of the algorithms from different groups. However, since they don't know the total number of topic related tweets in the whole dataset, they are not able to calculate "recall" for the algorithms. Therefore, this is still only a partial solution.

To deal with problems involving very large datasets with extremely rare positive instances we propose a novel way to build gold standard datasets in this chapter. We start by introducing our method of building the gold standard dataset. Then we use the method to build the gold standard dataset for our life satisfaction problem. In order to further evaluate the method, we build three more gold standard datasets (TREC Microblog Track datasets). Finally we compare our method with the pooling method used in TREC.

4.1 Find vs. Label

If the dataset is very large (e.g. millions of instances), it is practically impossible to manually annotate every data instance. It is time consuming and could be expensive to pay annotators. One traditional way is to annotate a random subset. However, this method does not work when the positive signals are very rare since positive instances may easily be missed by a random selection process. Moreover the random selection process is done only once.

The novel method we propose can involve any type of annotators including hired experts and crowdsourcing workers. The idea is that we do not need to annotate every instance since most of them are not the positive signals! Instead, if we

can"find" the positives, or at least close to all the positives, then we may safely assume that the remaining instances are negatives. To find those positives, we propose using an information retrieval engine to help annotators. We ask annotators to become searchers and find the positives in the dataset using the retrieval system. The searchers are welcome to use any query that they can design. The detailed procedure is listed below:

- Delete the duplicates in the dataset, but keep a record of deleted data instances.
- Index the whole dataset using the retrieval engine.
- Let annotators search for the positive signals and submit them to us.
- Remove the submitted items from the dataset so that they are not searchable by other annotators.
- Run the process for a period of time until it becomes very hard for annotators to find more positive signals.
- For the submitted data instances, obtain more judgements so that we can decide with a majority vote.
- Then assume almost all the positives were found, and the remaining instances are negatives.
- Inherit the labels for deleted duplicate data instances using their labeled original.

This method of building a gold standard dataset is one of the contributions of this thesis. The method is different from traditional methods in three aspects: 1) It does not require us to label all the data instances, which is impossible for large datasets. (The traditional method needs to label all the data instances or a random sample subset.) 2) We ask the annotators to find the data instead of labeling them, because of their diverse thinking, we are able to find the needles in the haystack. 3) Our method supports more accurate evaluations (i.e., more precise recall calculations). In addition, the cost could be much lower than traditional ways, since we don't need to label all the data instances, only the positive ones that are found and submitted.

In the previous chapter, we used life satisfaction as the case study to demonstrate our survey-to-surveillance method. We also found the percentage of tweets (less than 0.1% in the later analysis) indicating life satisfaction is very small. Therefore, building a life satisfaction gold standard dataset can not only be used to evaluate our surveillance method, but also demonstrate our method of building a gold standard dataset. To further validate our method of building a gold standard dataset, we annotate datasets from TREC Microblog track. The advantages of these datasets are that they are publicly accessible, and most importantly they are very large (millions of tweets). They cover a variety of topics ranging from U.S. to international news. The relevant tweets are also very rare. In addition, TREC provided their own annotation of relevant tweets for each topic, although this was done using the pooling method described earlier, and the annotation is not complete. Therefore, we could use our

method to build gold standard datasets for selected topics and compare them with TREC's annotation. So next, we discuss how we adopted the method to build different gold standard datasets including the life satisfaction tweets dataset, and TREC Microblog track topic datasets.

4.2 Life Satisfaction Dataset

We randomly selected two days: 2012-12-30 and 2013-01-11. We would like to build a gold standard dataset for life satisfaction using all the two day's first person (FP) tweets. The tweets in the dataset should be assigned to one of the three labels: Class S, Class D, and Class I. Class S and Class D are the positives we would like to find. The FP tweets were obtained from Twitter Streaming API[1] with filters. Each collected tweet must have "I," "me," "my," or "mine" in order to be classified as a FP tweet. The total number of FP tweets recorded on the two days is about 8.5 million.

As we discussed in the previous section: firstly, in the corpus of FP tweets from the two days, we deleted duplicates. We have a record of tweets which were deleted, so we can use them in a later step. Then we used the Indri retrieval system [56] to index the remaining tweets. We built a website and asked MTurk workers to use whatever queries they want to search our corpus to find LS tweets. By the end of the experiment, they had tried 936 different queries. The top 20 queries they used are shown in Table 4.1. The frequency of most queries were less than five. They marked the tweets belonging to Class S or Class D, while ignoring Class I, and submitted them. The submitted LS tweets were deleted from the index, so that other users were

[1]https://dev.twitter.com/docs/streaming-apis

not be able to retrieve them again. The submitted LS tweets were re-evaluated by our team and assigned a final label (Fleiss' Kappa = 0.78). Workers were paid $0.05 for every correct LS tweet submitted. Workers were paid a fixed amount of money for the re-evaluation since we needed majority vote for each submitted tweet. Finally duplicate tweets inherited the labels that are assigned to the original.

The task of obtaining the life satisfaction gold standard dataset using MTurk started on May 12th, 2013, and ended on Dec. 14th, 2013. By the end of 2013, 327 workers had participated in this effort and they found 1,954 and 2,801 Class S and Class D tweets respectively. Total cost was $237.75 (search)+$90 (re-evaluation)=$327.75. We paid $30 for each of the three annotators for the re-evaluation of all the submitted tweets. Figure 4.1 shows the number of correct LS tweets (after our re-evaluation) found per week by MTurk workers. (We had a minor issue between week 15 and 20.) Figure 4.2 shows the number of cumulative correct LS tweets found per week by us and MTurk workers. (We initially searched thousands of tweets by ourselves, so the number is not 0 at week 1.)

We can see that there was a cold start in the beginning. In the first two weeks, MTurk workers found relatively small number of LS tweets. Then some more successful MTurk workers joined the experiment. In the first two months, MTurk workers helped us find the majority of the LS tweets. Then the number of LS tweets found dropped dramatically. By the end of the experiment, the MTurk workers were still submitting tweets, but only a tiny number of them were approved. We cannot say we found all of the LS tweets from the two day tweet collection of the FP tweets,

Table 4.1. Most frequent queries used by crowd workers to find LS tweets

Frequency	Query	Frequency	Query
140	Life is good	52	I am satisfied with life
118	Life	48	Happy life
75	My life sucks	38	My life
67	Life is great	37	Love my life
67	life sucks	35	I hate my life
63	I love my life	29	My life is perfect
62	I am happy	24	my life is awesome
60	Happy	24	I am satisfied
58	Happy with my life	24	I am satisfied with my life
57	My life is great	24	good life

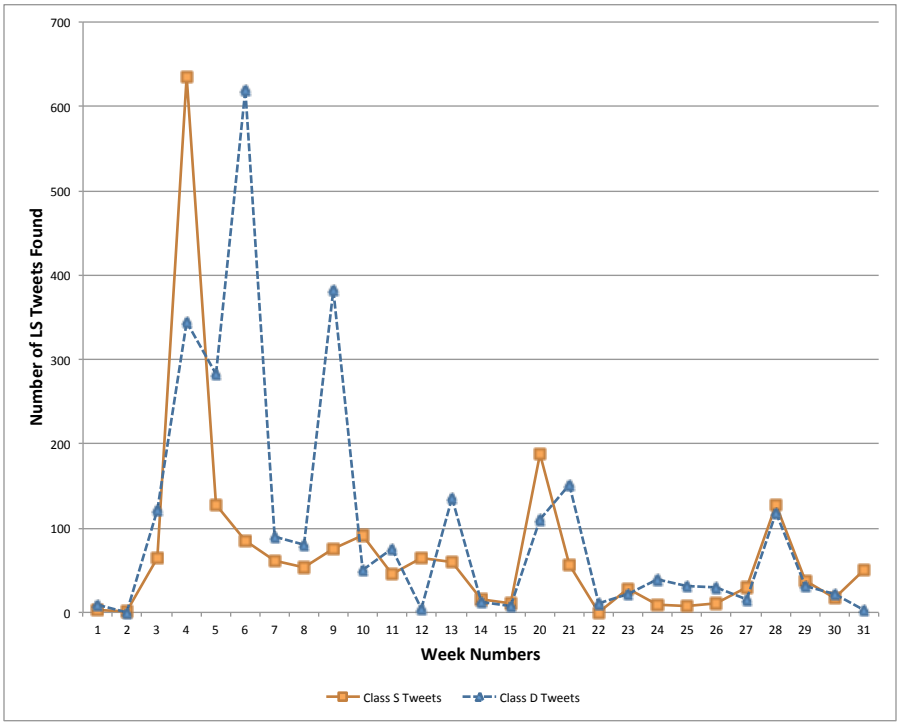


Figure 4.1. Number of LS tweets found each week

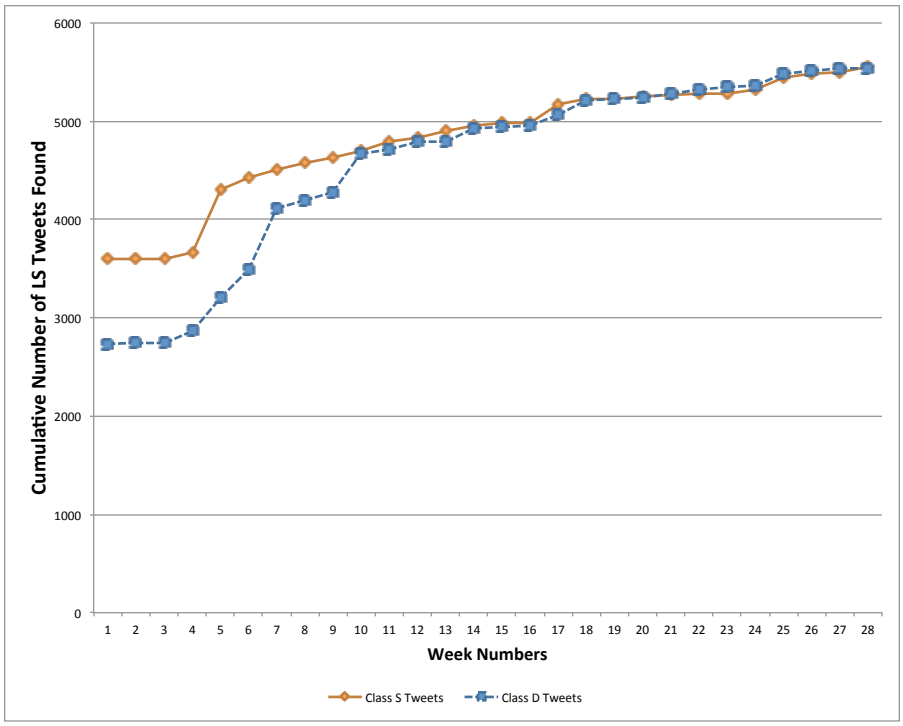


Figure 4.2. Cumulative number of LS tweets found

but we claim that we found close to all the positives (Class S and Class D tweets). Table 4.2 summarizes the characteristics of this two-day gold standard dataset for life satisfaction. From the table, we can see that MTurk workers' submissions were not highly reliable. The accuracy for finding LS tweets was only 0.38. Another important finding (though this is consistent with what we expected) is that the distribution of LS tweets is highly skewed. Both Class S and Class D tweets only account for less than 0.1% of the total number of FP tweets. It indicates that machine learning methods are likely to face a serious challenge, which we explore in Section 5.2.2. Our two-day gold standard dataset is the biggest collection of data on life satisfaction research using Twitter. Our method for producing the dataset is also novel.

Table 4.2. Summary of LS gold standard dataset

Total Two Days' FP Tweets	8,640,007
Total Two Days' FP Tweets Without Duplicates	7,437,551
Start/End Date of Labeling experiment	May 12th, 2013 / Dec. 14th, 2013
# MTurk Workers Participated	327
# Tweets Labeled By Us Initially	Class S: 3,594 Class D: 2,735
# Tweets Submitted By MTurk Workers	Class S: 6,001 Class D: 6,557
# Correct Submissions by MTurk Workers	Class S: 1,954 Class D: 2,801
Accuracy of MTurk Workers	38%
Final Unique Tweets in Labeled Set	Class S: 5,547 Class D: 5,536 Class I: 7,426,468
Final Tweets (considering duplicates) in Labeled Set	Class S: 8,181 Class D: 6,250 Class I: 8,625,576
Class Distribution in Labeled Set	Class S: 0.095% Class D: 0.072% Class I: 99.833%

4.3 TREC Microblog Topics Datasets

Our goal is to further explore the potential of this method of building gold standard datasets. We would like to use some other datasets which also have very small number of positives. We found the datasets in TREC Microblog Track 2011 are good fits.

The task in the Microblog track was to design algorithms to retrieve tweets relevant to given topics from a large tweet collection. Each topic is given a time stamp and the relevant tweets should have a time stamp earlier than the timestamp of the giving topic. To evaluate the performance of the algorithms from participants, TREC pools all the submitted tweets of all participants and hires annotators to annotate the relevance of the submissions. The relevance of submitted tweets were stored in a file named Qrel. Then precision-based measures were used to evaluate the performance of participants' retrieval algorithms. However, since pooling of submissions was done,

the total number of relevant tweets existing in the dataset for each topic is still unknown. Therefore, another important metric "recall," was not calculated.

In this section, we would like to use our method to build gold standard datasets for three selected topics. We would like to see if we could find all the positives. How long will it take to build the gold standard datasets? Can we find more relevant tweets for each topic than the Qrel from TREC? How much money do we need to build gold standard datasets?

We manually selected three topics for TREC Microblog track 2011: "Pakistan diplomat arrest murder (Topic 07)," "Assange Nobel peace nomination (Topic 12)," and "US unemployment (Topic 26)." We intentionally selected these three different types of topics. Based on the TREC Qrel, "Assange Nobel peace nomination (Topic 12)" has very small number of relevant tweets (Just 4 out of 8.2 million). "Pakistan diplomat arrest murder (Topic 07)" and "US unemployment (Topic 26)" have relatively more relevant tweets. (122 out of 16 million for topic 07; 144 out of 11.8 million for topic 26). Also, topic 07 and 12 are international news and topic 26 is US national news.

Similar to what we did for the life satisfaction dataset, we indexed the dataset for each topic using Indri and ran three separate crowd-based experiments to find the relevant tweets. Based on the timestamp for each topic, we created its dataset since we only need to find the relevant tweets before the topic time stamp. We asked MTurk workers to use whatever queries they wanted to search the dataset for each topics. We ran the experiments for about three months.

At the end of the experiments, MTurk workers had tried 18, 13, and 124 different queries to search relevant tweets for topic 07, 12, and 26 respectively. The most frequent queries for searching are shown in Table 4.3. Similar to the result of building life satisfaction dataset, the frequencies for most of the queries are less than five. All the submitted tweets were re-evaluated again, the final labels were obtained using majority vote from three annotators. Fleiss' kappa scores are 0.94, 0.90, and 0.69 for Topic 07, 12, and 26 respectively. Therefore, the annotation agreement for Topic 07 and 12 is very good, but the agreement for Topic 26 is just moderate. It indicates "US Unemployment" is a harder topic than the other two.

The results with the three datasets are shown in Table 4.4. The number of total relevant tweets is the combination of TREC Qrel and our evaluation of submitted tweets. Note that a submitted tweet has to have at least a majority vote (2 of 3) to be considered relevant. We see that our method found much more relevant tweets than the TREC Qrel. Especially for Topic 26, the TREC Qrel has 144 relevant tweets, while the total relevant is 459. Therefore estimating recall on the TREC pooled data would be highly likely to differ from using more complete information about the prevalence of relevant items. Notice that the positives are even less than the life satisfaction dataset. Especially for Topic 12, there are only 11 relevant tweets out of 8.2 million tweets. However, it is surprisingly easy to find relevant tweets for this topic, as the MTurk workers found all the relevant tweets. The results are also good for the other two datasets, 93.09% and 96.30% of total relevant tweets were found for Topic 07 and Topic 26 respectively. In addition, the cost is very low, all the topics

cost about $50 (Details in Table 4.4). We also notice that the average accuracy of MTurk workers is also low as for building the life satisfaction dataset.

Table 4.3. Most frequent queries used by crowd workers

Topic 07		**Topic 12**		**Topic 26**	
Freq.	**Query**	**Freq.**	**Query**	**Freq.**	**Query**
20	pakistan	14	assange nobel peace nomination	100	us unemployment
18	pakistan diplomat	10	assange nobel	89	unemployment
14	pakistan diplomat arrest murder	8	assange	20	unemployment snippets
10	us diplomat pakistan	8	tweets	18	american unemployment
10	pakistan diplomat murder	6	nobel	18	unemployed
8	arrest diplomat	6	nobel peace	16	jobless
8	pakistan murder	4	woman	16	us
8	pakistan diplomat arrest	4	oscar	14	united states unemployment
6	cia murder	4	peace	12	unemployment rate

Figure 4.3 shows the cumulative number of participants in the three month experiments. Topic 07 and Topic 12 had similar number of participants, and Topic 26 had much more participants. One possible reason why Topic 26 has more participants is that the topic is more general than the other two, the workers may think this topic has more relevant tweets. We would like to see how many people did a good job. So we count the number of people who have accuracy of more than 0.5. We found that 57.89%, 25%, and 71.79% of the participants have good accuracy for Topic 07, 12, and 26 respectively.

Table 4.4. Summary of TREC microblog datasets

Topics	Pakistan diplomat arrest murder (Topic 07)	Assange Nobel peace nomination (Topic 12)	US unemployment (Topic 26)
# Docs in dataset	16 million	8.2 million	11.8 million
Experiment Time	3 months	3 months	3 months
# Total Relevant Tweets	188	11	459
# Total Relevant Tweets identified by TREC	122	4	144
# Participants	19	16	39
# Total submissions	216	86	593
# Relevant submissions	175	11	442
# Relevant tweets not found	13	0	19
Accuracy of MTurk Workers (Std.)	0.58 (0.49)	0.25 (0.45)	0.68 (0.35)
Percentage of relevant tweets found	93.09%	100%	96.30%
Cost	\$8.75 (search)+ \$9 (re-evaluation) =\$17.75	\$0.55 (search)+ \$9 (re-evaluation) =\$9.55	\$22.1 (search)+ \$9 (re-evaluation) =\$31.1

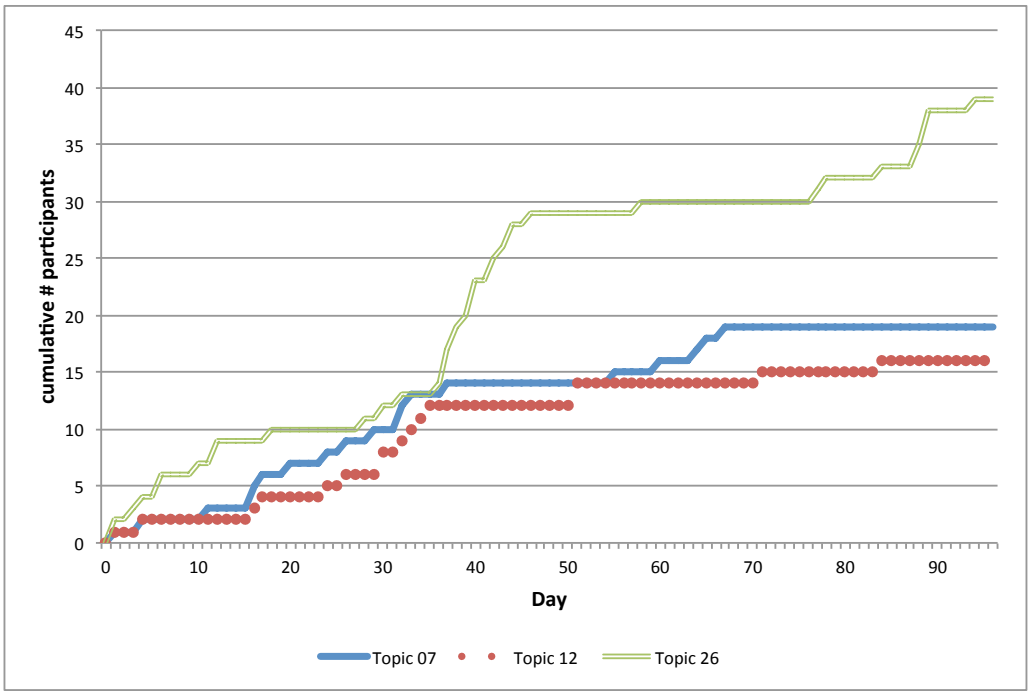


Figure 4.3. Cumulative number of participants

The cumulative number of relevant tweets retrieved for topic 07, 12, and 26 are shown in Figure 4.4, 4.5, and 4.6 respectively. For each figure, the green dashed line represents the cumulative number of tweets submitted from the MTurk workers. The solid blue line represents the cumulative number of relevant tweets submitted. The red double line represents the number of relevant tweets in TREC Qrel. Overall, MTurk workers were able to find most of the relevant items in about 1 month. By the end of the three months, we see that it is very hard for the MTurk workers to find more relevant tweets. So we could reasonably claim that we found close to all the positives for the datasets, and that the remaining tweets are negatives. Compared to TREC's pooling process, we find 54.10%, 175%, and 218.75% more relevant tweets for Topic 07, 12, and 26. In addition, we found for some topics like 12 and 26, there

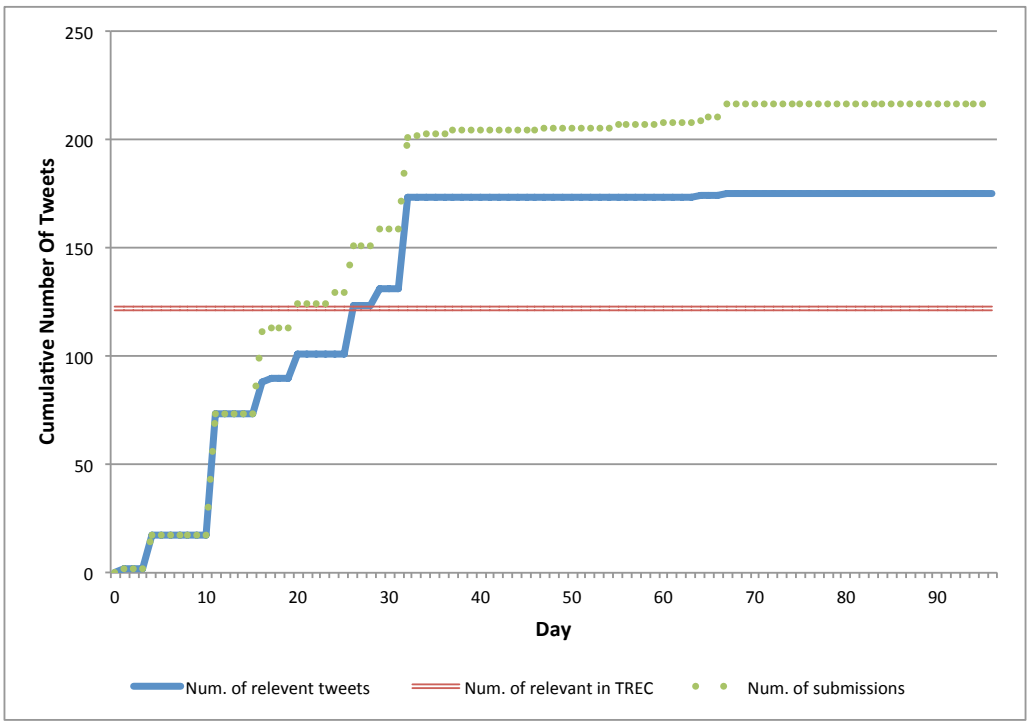


Figure 4.4. Cumulative number of relevant tweets found for Topic 07

are overlaps in the green and blue lines for the first 10 days. In the overlapping region everything submitted was declared relevant by majority vote. This region represents positives that were easy to find compared to non overlapping regions. Also for topic 12 the overlapping blue green region is higher than the red line representing TREC decisions. In other words the TREC process seems to have missed some relevant items that were easy to find for our crowd workers. This is also true to a slight extent for topic 26.

In conclusion, we developed a method to build a gold standard dataset. The method is best for a dataset which has a large number of instances and a small ratio of positives. Different from the traditional way of building a gold standard

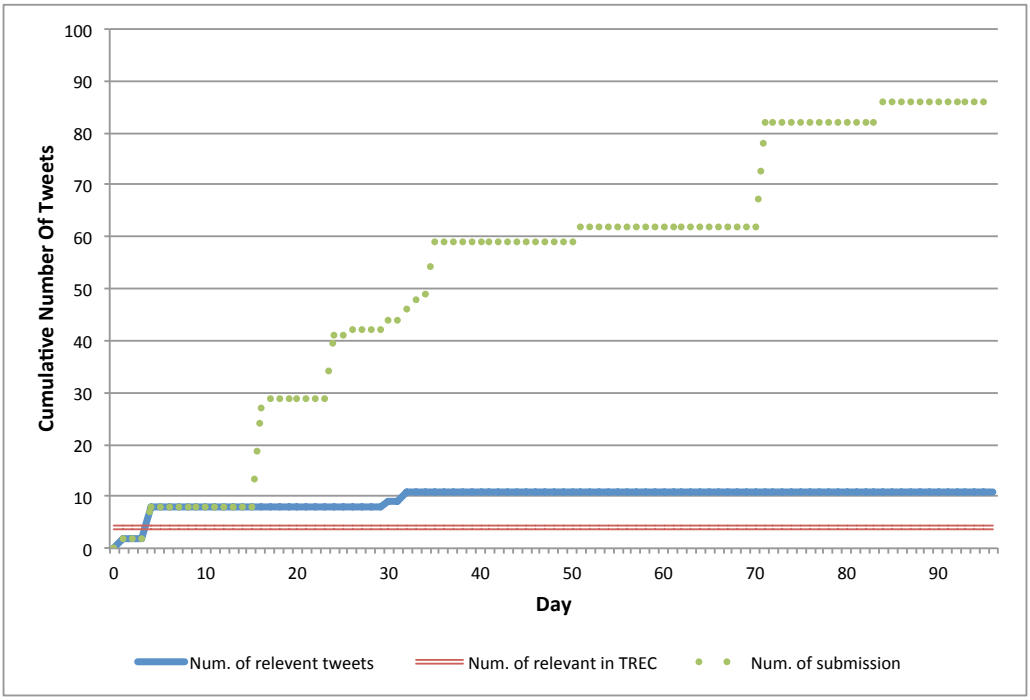


Figure 4.5. Cumulative number of relevant tweets found for Topic 12

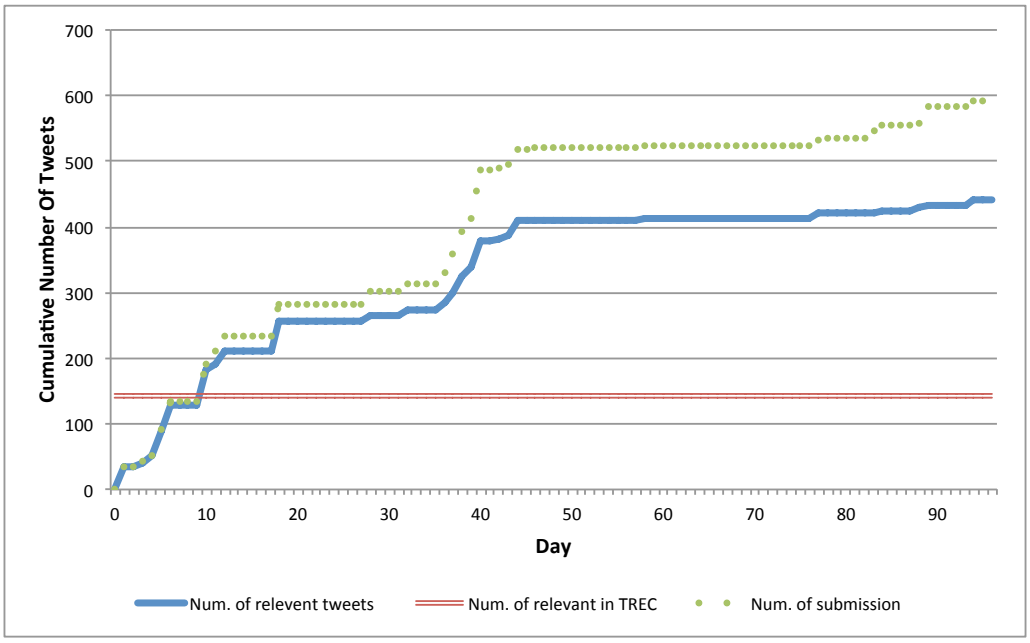


Figure 4.6. Cumulative number of relevant tweets found for Topic 26

dataset, our method doesn't require labeling of all the data instances. Also we ask the annotators to find positives in the dataset instead of labeling a given set or subset. The method supports computing recall, since we are able to find close to all the positives. In addition, the cost is lower for building the gold standard dataset than traditional expert based ways. We used the method to build one life satisfaction dataset, and three TREC Microblog track datasets. We showed our method can be used for different problems and the method is solid. In the comparison of our method and TREC's pooling method, our method obtained many more positives.

The limitation of our "annotate by finding" method is that when the number of positive instances is large, the cost to pay for submitted positive instances could be high. We also notice that for all the four gold standard datasets we built, the average accuracy of submission from MTurk workers is low. To further improve this method, we could improve the accuracy of submission by using Master workers only (Masters are one type of workers who have demonstrated accuracy on previous tasks). Additionally, we could have a penalty for workers who submitted irrelevant data. The penalty may encourage the workers to be more careful and increase the accuracy of their submission. In addition, we currently have a secondary process to obtain more judgements for the submitted data in order to have the majority vote. We could obtain the majority vote for each data instance by allowing more than one submission for each instance submitted. For example, if one data instance is submitted twice, then it can be consider as positive.

CHAPTER 5

EVALUATION OF THE TEMPLATE BASED SURVEILLANCE METHOD

We introduced our survey-to-surveillance method in the Chapter 3. We introduced how we build a gold standard dataset in the previous chapter. Next we use our life satisfaction gold standard dataset to evaluate our surveillance method. We ask several questions. How good are our filters? What evaluation metrics could be used for evaluating the performance of surveillance of life satisfaction? In addition, since the surveillance method is based on information retrieval technology and filters, can we approach the same task using other popular technologies? Specifically, can we use lexicon based or machine learning based methods to do surveillance on social media? If yes, are they better than our surveillance method based on information retrieval technology? In this chapter, we are going to answer these questions.

5.1 Evaluation Of Survelliance Method

Since we have almost all the LS tweets identified in the gold standard dataset, we could use precision and recall as the metrics to evaluate our method. Ours is the first study to use a very large dataset (here we use all the tweets in two days) to evaluate a detection method. It is also the first study which is able to provide recall scores. Traditionally, only precision based metrics are provided.

Again, the life satisfaction surveillance method uses templates to retrieve Class S & D tweets. We also use filters to filter out the irrelevant tweets. First we evaluate

Table 5.1. Accuracy of filters on 2012-12-30 data (Filtered 7,035 tweets)

Filter	**# Tweets Filtered Out**	**Accuracy**
Remove tweets ending with irrelevant words	773	0.78
Remove tweets referring to past or future tense.	3,222	0.95
Don't cross sentence boundaries when filling a template	2,138	0.95
Remove tweets referring to 3rd person.	518	0.98
Remove tweets asking a question	384	0.99

the accuracy for each filter in our method. Table 5.1 shows the accuracy of each filter on the 2012-12-30's gold standard dataset. Here accuracy means the percentage of tweets which are correctly filtered out. E.g., the filter "Remove tweets asking a question" filters out the 384 tweets which are questions. 99% of the filtered out tweets are correctly identified as questions. Most of the filters have very good results. Removing tweets ending with irrelevant words is harder than the other filters, but the accuracy is still reasonable. Therefore, overall our filters are effective.

Next we examine the effectiveness of our methods for detecting LS tweets. Table 5.2 and 5.3 show the result for detection of Class S and Class D tweets respectively using 2012-12-30's gold standard dataset. Again, the columns indicate the word gap, which is the number of other words allowed between the words of a template. A indicates the word gap is 0 and D indicates the word gap is 3. The rows indicate the number of words (W) allowed before or after the text segment that satisfies the tem-

Table 5.2. Precision for detection of Class S tweets on 2012-12-30 data

#Retrieved (Precision)	A	B	C	D
W1	404 (0.93)	463 (0.68)	245 (0.34)	167 (0.3)
W2	266 (0.83)	320 (0.77)	180 (0.57)	99 (0.34)
W3	202 (0.73)	494 (0.89)	151 (0.48)	54 (0.17)
Total: 3,045 (0.69)				

Table 5.3. Precision for detection of Class D tweets on 2012-12-30 data

#Retrieved (Precision)	A	B	C	D
W1	419 (0.98)	208 (0.89)	184 (0.67)	123 (0.55)
W2	260 (0.95)	166 (0.66)	139 (0.57)	107 (0.33)
W3	112 (0.69)	132 (0.75)	76 (0.63)	78 (0.13)
Total: 2,004 (0.74)				

plate. The number of Class S and Class D tweets retrieved by each retrieval strategy are shown in the two tables. Also the precision of each retrieval strategy is shown in parentheses.

We can see that with one exception precision scores are generally highest for column A and decrease along each row. Comparing rows there is less consistency, for column A in both tables and for column B in the second table, precision decreases as we go down the rows. In the first table the middle row seems to have a bump. A-W1, A-W2 are some of the strategies that make high contributions to recall without sacrificing precision.

The final precision and recall scores of our surveillance of life satisfaction on the two days' gold standard dataset are shown in Table 5.4. We see that the precision scores are good for Class S and Class D on 2012-12-30. Recall scores on 2012-12-30

Table 5.4. Precision and recall on two-day gold standard dataset

	2012-12-30			2013-01-11			Two Day Combined		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Detection Of Class S	0.69	0.46	0.55	0.60	0.65	0.62	0.65	0.53	0.58
Detection Of Class D	0.74	0.43	0.54	0.59	0.60	0.59	0.66	0.50	0.57

are lower, above 0.4. The combined F1 is above 0.5. On 2013-01-11, F1 is around 0.6 due to increasing of recall scores. The explanation could be that 2012-12-30 is the end of the year. People may express their life satisfaction more and use more varieties of expressions. We did the analysis in order to find out why we missed about 50% of LS tweets. One reason for a large portion of the misses is tweet length; we deliberately avoided retrieving long tweets (using W $\leq$ 3) hoping to keep precision reasonable. But we missed, for example, "*I have a perfect life, well its perfect in my eyes. I'm not complaining by any means.*" We will focus on long tweets in future work.

5.2 Comparison Between Our Method and Other Methods

Our surveillance method has good precision and reasonable recall performance. Since our method is based on information retrieval technology, we would like to know what other approaches we could adopt for the same task? If we can use other algorithms to do the surveillance on social media, how good are they? Therefore, we built surveillance of life satisfaction approaches using two other popular categories of methods and compared with our information retrieval based approach in this section. Because there is no existing method for detecting life satisfaction from the previous studies on social media, we built the methods by modifying the existing methods used for detecting "happiness" expressions. Two kinds of surveillance methods based on

lexicon and machine learning were built separately.

5.2.1 Lexicon-based Methods

Most of the previous studies use lexicon-based methods to detect happiness. Therefore, we built lexicon-based method to detect life satisfaction. We explored two lexicons here: labMT which was built specifically to detect happiness, and ANEW which is popular for detecting sentiment.

5.2.1.1 labMT Lexicon

The closest study to ours on detecting LS is Dodds et al.'s work on "Happiness," which has been mentioned in Section 2.4. Again, they built a 10,000 word happiness lexicon named labMT. Then they calculated the happiness score for each tweet using the lexicon. They defined the "happy tweets" as tweets which have score above 6.0. (They only define the "happy tweets," not the "sad tweets.") Dodds' method has been adopted in several studies [45, 7].

Since happiness and life satisfaction are both aspects of subjective well being we explore if it is possible to apply their method with small modification to find LS tweets. In detail, we randomly sampled 100,000 tweets from 2012-12-30's FP tweets. We only kept the tweets containing the word "life" in order to favor tweets that might talk about life satisfaction. 1,613 tweets have the word "life." After applying Dodds et al's method, there are 1,093 "happy tweets" (score above 6, defined by Dodds et al.) among them. Only 55 tweets have scores below 5 and 2 tweets have scores below 4.

We manually judged the top 100 "happy" tweets and top 100 "sad" tweets

Table 5.5. Top 10 "happy" tweets (with the word "life") ranked by LabMT lexicon

Rank	**Tweet**	**Happiness Score**	**Class S?**
1	So funny"@iam dheji: The love of ur life"@Teh mi: Sule"@iam dheji: @Teh mi @dacutest me u that mouth that u are usin to hiss, I destr"	8.42	No
2	@Louis Tomlinson I love your zest for life ;3 Xx	7.87	No
3	@babblingbrat But do I have a life? *laughs*	7.75	No
4	RT @MeekMill: I aint tripping life Is great!	7.60	Yes
5	@DaniCim Dani you're super! when you smile at me your smile forces also smile ! your life is Paradise:) youre the best	7.47	No
6	RT @phamBAAAM: a haiku about my love life: hahaha- haha. hahaha- hahahaha. hahahahaha	7.46	No
7	RT @DLdemons: I find it hard to trust, hard to love, hard to determine what's real in this life. Some days it just gets too hard.	7.44	No
8	RT so true @MeekMill: I ain't tripping life Is great!	7.43	Yes
9	@postsofagirl: I'll look back on this and smile, because it was life and I decided to live it. #postsofagirl	7.42	No
10	I'm back in the relationship life And I must say it's great...	7.41	No

Table 5.6. Top 10 "sad" tweets (with the word "life") ranked by LabMT lexicon

Rank	**Tweet**	**Happiness Score**	**Class D?**
1	@ASAP Crosby: Im mad Nelson tried to hit the slick move with the challenge flag...lol he done stashed some shit before in his life..lol	3.88	No
2	I hate crying and being unhappy with parts of my life.	3.99	Yes
3	@MsKgoribby I dnt knw lenna bt my witch doctor said ontshelletse in a drink(extream) so ill folow u 4 life.I ws lik ah vele I was gona do dt	4.13	No
4	RT @ColdYellow Bone: "@xLilBOOTYJudy : - same bitches screaming we gone be bitches for life , switched up & became bitches I don't l ..."	4.24	No
5	I fear for Luck's life with his o-line against the Ravens	4.35	No
6	RT @DanaJones : I dont understand why teens waste their teenage years on drugs and alcohol? Its the most pivotal moment in your life..wh ...	4.37	No
7	If she does then im gonna be really mad and just hate my life and die.	4.39	Yes
8	RT @JayZClassicBars: "This is why I be so fresh / I'm tryna beat life, 'cause I can't cheat death" #Pray	4.40	No
9	@RayRay215 lmaooooo I'm crying Ray. I quit life smh	4.42	Yes
10	"@societybarbie: Missing people is one of the things I hate most in life"	4.43	No

that have the word "life" ranked by happiness score. In the top 100 "happy" tweets, only 16 correctly indicate the author's life satisfaction, while in the top 100 "sad" tweets, only 20 correctly indicate the author's life dissatisfaction. In other words, the precision@100 for Class S tweets is 0.16, while the precision@100 for Class D tweets is 0.2. Table 5.5 and Table 5.6 show the top "happy" tweets and "sad" tweets which have the word "life" respectively.

This shows that the method using labMT lexicon even with modification maybe not good for detecting life satisfaction on Twitter. For the sake of completeness we would like to see if the lexicon is good for detecting "happiness," which is the original purpose of labMT lexicon. We conducted our own evaluation of Dodds' method for detection happiness in tweets by manually evaluating the top 100 FP tweets from the 100,000 random selected tweets ranked by happiness score. We find that it is challenging to identify a tweet that conveys happiness. Consider the following two examples: "I am in love," "I love his truck." One can observe the difference, the second requires a more loose definition of happiness compared to the first. Adopting a broad definition we found precision@100 to be 0.66. Note that precision is likely though this is not a certainty to drop as we move further down the ranks (below 100).

5.2.1.2 ANEW lexicon

We explored another popular lexicon from sentiment analysis research which is Affective Norms for English Words (ANEW). It was developed by Bradley and Lang [9] in 1999. It provides a set of normative emotional ratings for about 1,000 words in the English language. Three major dimension scores have been rated for every word in

ANEW using survey results. Affective valence: ranging from pleasant to unpleasant. Arousal: ranging from calm to excited. Dominance: ranging from dominance to submissiveness. For example, both "fear" and "anger" are unpleasant emotions in valence dimension, but "anger" is a dominant emotion, "fear" is a submissive emotion in dominance dimension. To do the sentiment analysis, we could calculate the Valence scores for the text. We utilized ANEW lexicon for the same 1,613 tweets having the word "life" from 2012-12-30. We computed a mean score of valence for each tweet, and ranked tweets by this score. Table 5.7 and Table 5.8 show the top tweets and bottom tweets which have the word "life," ranked by ANEW valence score, respectively. We see the lexicon-based method using ANEW has similar results with the one using labMT lexicon.

Table 5.9 shows the performance of the two lexicon-based methods for detecting LS. Our surveillance method does not rank the LS tweets, but our best strategies could achieve the precision of more than 0.9 (Table 5.2, 5.3). The precision is significantly better than the lexicon based method. In conclusion, we used two lexicons: labMT which is used to detect happiness and ANEW which is a more general sentiment lexicon. Neither of two lexicons work well for detecting LS. We decided not explore more lexicons since there is no lexicon designed specifically for detecting LS. We conclude from the experiments that using either of the two lexicons creates a big risk with precision for detecting life satisfaction on Twitter.

Table 5.7. Top 10 tweets (with the word "life") ranked by ANEW valence

Rank	**Tweet**	**Valence Score**	**Class S?**
1	So funny"@iam dheji: The love of ur life"@Teh mi: Sule"@iam dheji: @Teh mi @dacutest me u that mouth that u are usin to hiss, I destr"	8.72	No
2	@tattedbikelover yes. i swear the love of my son has opened me up to real feelings I didnt knw I was only faking before in my life.now i knw	8.72	No
3	@hippiemollz99 the true love of my life 3	8.72	No
4	RT @ltsTyga: I am who I am. I like what I like. I love who I love. I do what I want. Get off my back and deal with it. Its my life, not ...	8.24	No
5	@DanielllePeazer Hay.How are you ? I love you, you are my inspiration and life ! I love you, please follow me 4	8.24	No
6	RT @ltsTyga: I am who I am. I like what I like. I love who I love. I do what I want. Get off my back and deal with it. It's my life, not ...	8.24	No
7	RT @EndBullyinNow: You are so amazing. You are talented. You are beautiful. And you have saved so many lifes, including mine. I love yo ...	8.16	No
8	Dear God, I know you have plans.....and i am settled and done in your notebook of life!..looking forward for what is installed for me	8.15	No
9	I drink every day and music is my #life. X	8.13	No
10	Honestly can't wait for tomorrow. Movies all day long with the love of my life, then my first real New Years kiss #newyearseve #2013	8.08	No

Table 5.8. Bottom 10 tweets (with the word "life") ranked by ANEW valence

Rank	**Tweet**	**Valence Score**	**Class D?**
1	RT @Flopz17: I wish I could have done better things my life....well gotta keep trying I ain't dead yet so who knows.	3.88	No
2	"@iFriendships: I really really really really wanna be friends with Fat Amy in real life." you already are #itsme @SavRoell	2.28	No
3	@ASAP Crosby: I'm mad Nelson tried to hit the slick move with the challenge flag...lol he done stashed some shit before in his life..lol	3.39	No
4	@SorryNotSoory: it's not my fault your mean and nobody likes you story if my life	3.43	No
5	@Aurellie xo: No lie this time tomorrow I'll probably be sleeping. My very very sad life not sad, on point trust me!!	3.60	No
6	#Maybe2013WillBring justic fr women, death of hatred n wif in my life	3.62	No
7	I hate crying and being unhappy with parts of my life.	3.65	Yes
8	RT @HayleeJaike: Failure is my ultimate fear in life..	3.91	No
9	If she does then im gonna be really mad and just hate my life and die.	3.94	No
10	If you knew how lonely my life has been & how long I've been so alone.	3.95	Yes

Table 5.9. Performance of lexicon-based methods

Lexicon Method	labMT	ANEW
Precision@100 (Class S)	0.16	0.11
Precision@100 (Class D)	0.20	0.23

5.2.2 Machine Learning Method

Another approach for detecting "happiness" or other emotions is using machine learning algorithms. (This is not popular in happiness studies probably because the gold standard dataset is hard to build). However, since we have built the gold standard dataset, we can test machine learning methods. We employed SVM (SMO in Weka) [29] which had good performance in our previous work [65] and other text mining problems [1].

For detecting LS, we conducted several different experiments. We not only did 10 fold cross validation on our datasets, but also tried to train using one day's data and test using another day's. As far as we know, this is the only work testing the ML method for detecting life satisfaction using a very large dataset (here we use all the tweets in one day). In detail, we used gold standard data from 2012-12-30 to build and test the classifiers with 10 fold cross validation. 1, 2, and 3-gram features were used. The classifiers distinguish three classes: Class S, Class D, and Class I.

We did three experiments using SVM. Experiment 1 used a balanced training set: the numbers of tweets from the three classes are balanced. In this case, SVM classifier should achieve best performance in 10 fold cross validation. But this design is unrealistic as the in reality LS tweets are rare. Experiment 2 and 3 used unbalanced training sets. For these two experiments, we increased the number of tweets for Class I. Experiment 2 has about one million Class I tweets, while Experiment 3 has 3.7 million (all the non-duplicate Class I tweets on 2012-12-30). These two unbalanced experiments simulated the real data distribution better than the balanced experiment.

Table 5.10. Class distribution of training set from 2012-12-30

	Experiment 1: Balanced Training Set	**Experiment 2: Unbalanced Training Set**	**Experiment 3: Unbalanced Training Set**
Total # FP Tweets	4,901	1,000,001	3,746,340
# Class S Tweets	1,559	1,559	1,559
# Class D Tweets	1,744	1,744	1,744
# Class I Tweets	1,598	996,698	3,743,036

Table 5.11. 10 fold cross validation on training set from 2012-12-30

	Experiment 1: Balanced Training Set			**Experiment 2: Unbalanced Training Set**			**Experiment 3: Unbalanced Training Set**		
	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
Class S	0.853	0.899	0.875	0.669	0.313	0.427	0.923	0.023	0.045
Class D	0.866	0.917	0.891	0.72	0.334	0.457	0	0	0
Class I	0.884	0.78	0.828	0.998	1	0.999	0.999	1	1
Accuracy	86.65%			99.73%			99.91%		
Running Time	30 secs			1.6 days			8.6 days		

Table 5.10 and Table 5.11 show the class distribution of the training set and performance of 10 fold cross validation for the three experiments respectively. From the tables we can see that in the 10 fold cross validation, the balanced experiment has the best performance as expected. The F scores achieved more than 0.8 for the three classes. But in Experiment 3 which is the most realistic, we cannot detect Class D.

Next, we trained the classifier using LS tweets from 2012-12-30, and tested using all the LS tweets from 2013-01-11. We didn't use the classifier of Experiment 3 because it cannot detect Class D tweets. The result is shown in Table 5.12. We can see that for the balanced experiment, the F scores for Class S and D are about 0.01, because the precisions of Class S and Class D are less than 0.01. Experiment 2

Table 5.12. Experiments (traing on 2012-12-30, test on 2013-01-11)

	Experiment 1: Balanced Training Set			**Experiment 2: Unbalanced Training Set**			**Template-based Surveillance Method**		
	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
Class S	0.0054	0.9104	0.0108	0.34	0.35	0.35	0.6	0.65	0.62
Class D	0.0066	0.9061	0.0131	0.38	0.32	0.35	0.59	0.60	0.59
Class I	0.9999	0.8969	0.9456	0.99	0.99	0.99	N/A	N/A	N/A

achieved better F score (0.35). But it still performed much worse than our template-based surveillance method. (Since we only focus on finding the Class S and Class D tweets, we didn't calculate the scores for Class I for our method)

In conclusion, we built surveillance methods for life satisfaction on social media using both lexicon-based and machine learning approaches. We observe that the lexicon-based method using labMT and ANEW have low precision. ML method has better performance, but our template-based surveillance method is still the state-of-art method for detecting life satisfaction expressions on Twitter.

CHAPTER 6

ANALYSIS OF RETRIEVAL RULES

We introduce our surveillance method in Chapter 3. We have 430 Class S templates and 346 Class D templates. For each template, we have 12 retrieval strategies (Table 3.3). Here we define one "retrieval rule" as one combination of a template and a strategy. Thus we have a total of 5,160 Class S rules and 4,152 Class D rules. In Chapter 5, we evaluate the overall performance of our method. It is also important to know the performance of each rule independently. It not only shows a better understanding of our method, but also it unveils the usages of different life satisfaction expressions on social media. In this section, we did the analysis of the thousands of retrieval rules. We focused on questions like: how many rules actually retrieved LS tweets? Which rules are the most important in terms of different metrics? If we only can use n rules, which ones we would use? Note that the rules were developed essentially using one day's FP tweets. We also answer the questions like: are the results likely to be when using the retrieval rules on different dates?

6.1 Performance Of Retrieval Rules

To answer those questions, we used all the 5,160 Class S rules and 4,152 Class D rules to retrieve LS tweets on our two day gold standard dataset for life satisfaction. We treated those rules as totally independent. i.e. the LS tweets retrieved by rule 1 can also be retrieved by rule 2. Therefore, we can evaluate different aspects of the retrieval rules more equitably. Also filters (Table 3.4) were applied to boost

the precision of rules. We calculated four metrics for each rule: Number of tweets retrieved, Precision, Recall, and F score. Table 6.1 summarizes the result of all rules. We see that Class S and Class D rules have similar performances. As expected, not all the rules retrieved LS tweets. Slightly more than 20% of rules retrieve LS tweets. The rule which retrieved most tweets retrieved more than 1,500 tweets.

Table 6.1. Performance of retrieval rules

	Class S Rules		**Class D Rules**	
	2012-12-30	2013-01-11	2012-12-30	2013-01-11
# Rules Retrieving Tweets	1,184 (22.95%)	1,226 (23.76%)	972 (23.41%)	885 (21.32%)
Highest # Tweets Retrieved By One Rule	1,679	1,812	1,554	1,891
Highest Precision Achieved By One Rule	1.0	1.0	1.0	1.0
Highest Recall Achieved By One Rule	0.234	0.133	0.094	0.124
Highest F Achieved By One Rule	0.356	0.186	0.160	0.216

To represent the four metrics, we show the rules in 3D scatter plots. The three axes in the 3D scatter plots represent number of tweets retrieved, precision, and recall. Class S and Class D rules on 2012-13-30 are shown in Figure 6.1 and Figure 6.2 respectively. We select example points in Figure 6.1 to explain Class S rules in detail. Because the rules are too long and too complex to show on paper, we show the tweets that could be retrieved by the rules instead.

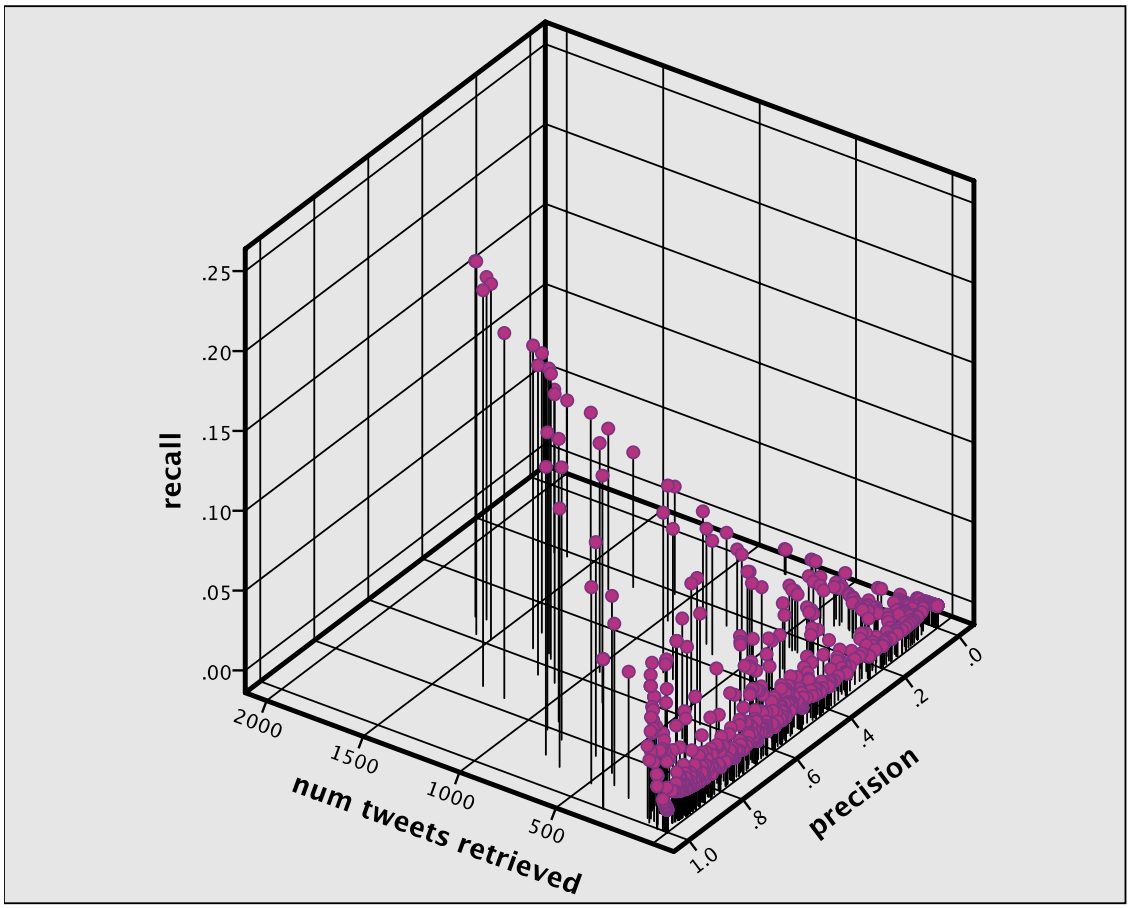


Figure 6.1. Performance of Class S rules on 2012-12-30

On the precision axis, most of the rules (near 80%) retrieve no tweets, so the precision is 0. For example, all the 12 retrieval strategies of "I am good with my achievements" have no tweet retrieved. About 20% of rules can retrieve tweets. And there are plenty of rules which have moderate precisions. For example, allowing 3 words in the template and 2 words around the template for "I have a perfect life" has the precision of 0.515, and retrieves 66 tweets. There are some rules which do not retrieve many tweets but have high precision. For example, allowing 2 words in the template and 2 words around the template for "I am satisfied the way it is." It has a precision of 1.0, but only retrieves 5 tweets. Some rules have high recall but the precision is not exciting, for example, allowing 3 words in the template and 2 words around the template for "I am happy." It has high recall of 0.2 (relatively high for one Class S rule), it retrieves 1500 tweets and has precision 0.582. Some rules have middle range of precision and recall. For example, allowing 4 words in the template and 1 word around the template for "I am satisfied" retrieves 973 tweets, with precision of 0.556 and recall of 0.129. Besides, some rules retrieve high number of tweets but the precision is low. For example, allowing 4 words in the template and 2 words around the template for "I have everything" retrieves 774 tweets, but precision is only 0.009.

For Class D rules, we explain some example points in Figure 6.2. A similar number of rules (20%) retrieve tweets. For example, all the 12 retrieval strategies of "I am far from satisfied in my life" have no tweet retrieved. Notice that this is a reasonable sentence construction to express life dissatisfaction. We just don't observe it in this dataset. There are also plenty of rules which have moderate precisions. For

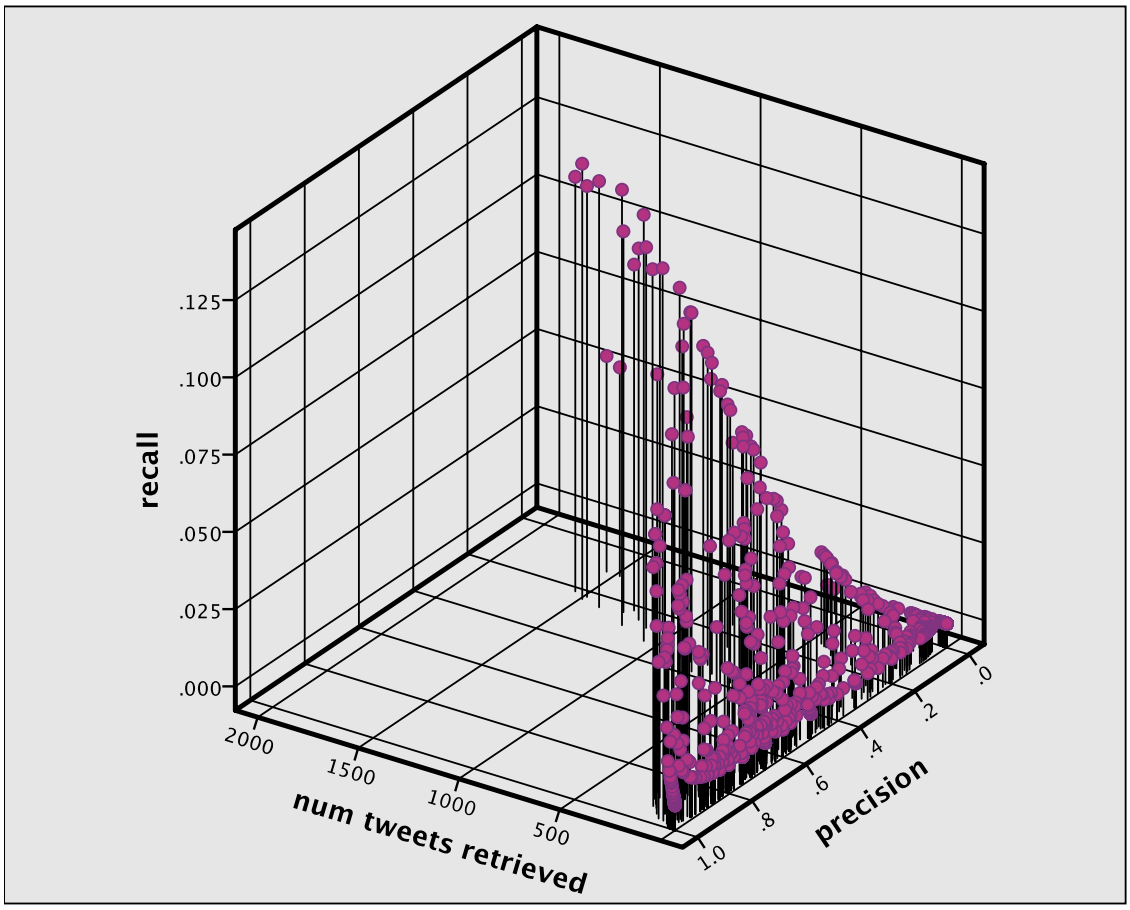


Figure 6.2. Performance of Class D rules on 2012-12-30

example, allowing 4 words in the template and 2 words around the template for "I have dissatisfaction in my life" has precision of 0.568, and retrieves 81 tweets. There are some rules which retrieve not many tweets but have high precision. For example, allowing 3 words in the template and 2 words around the template for "I need change in my life" has precision of 1.0, but only retrieves 6 tweets. The Class D rules do not have as high recall as the Class S rules. The highest recall achieved is 0.094. Some rules have middle range of precision and recall. For example, allowing 4 words in the template and 3 words around the template for "I hate my life" has high recall of 0.09 (relatively high for one Class D rule), retrieves 384 tweets and has precision 0.764. Some rules retrieve a high number of tweets but the precision is low. For example, allowing 4 words in the template and 3 words around the template for "I have depression" retrieves 1,544 tweets, but precision is only 0.028.

In addition, Class S and D rules on 2013-01-11 are shown in Figure 6.3 and Figure 6.4 respectively. We noticed that the shape in Figure 6.1 seems different from Figure 6.2, Figure 6.3, and Figure 6.4. The reason could be the date is close to New Year's Eve, the Class S tweets may be more than usual. Therefore, we can see that there are more templates which have high precision and number of tweets retrieved. Also, in Table 6.1, we can see the highest recall achieved is much higher in 2012-12-30 than in 2013-01-11.

6.2 Best *n* Retrieval Rules

In the previous section, we explore how many rules retrieved tweets, and we show the three evaluation metrics for all the rules. Now we want to know how to

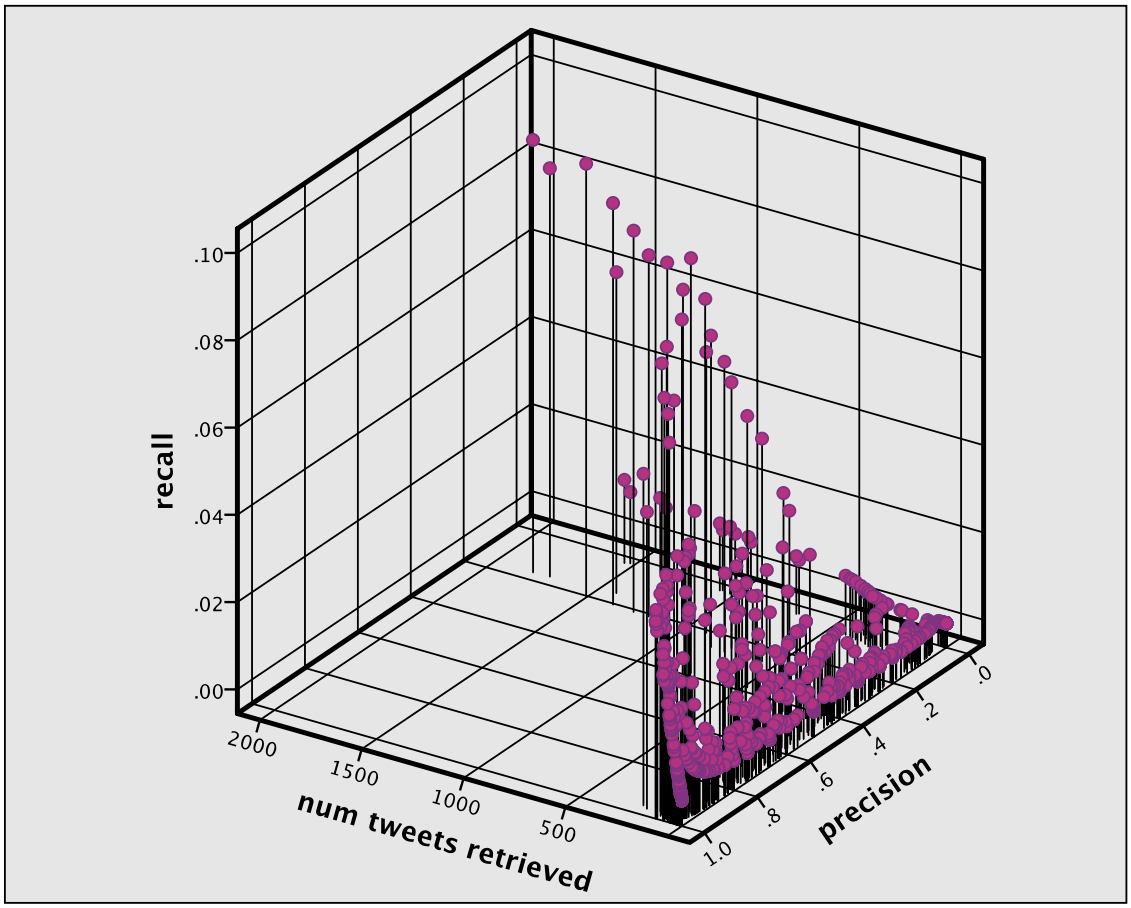


Figure 6.3. Performance of Class S rules on 2013-01-11

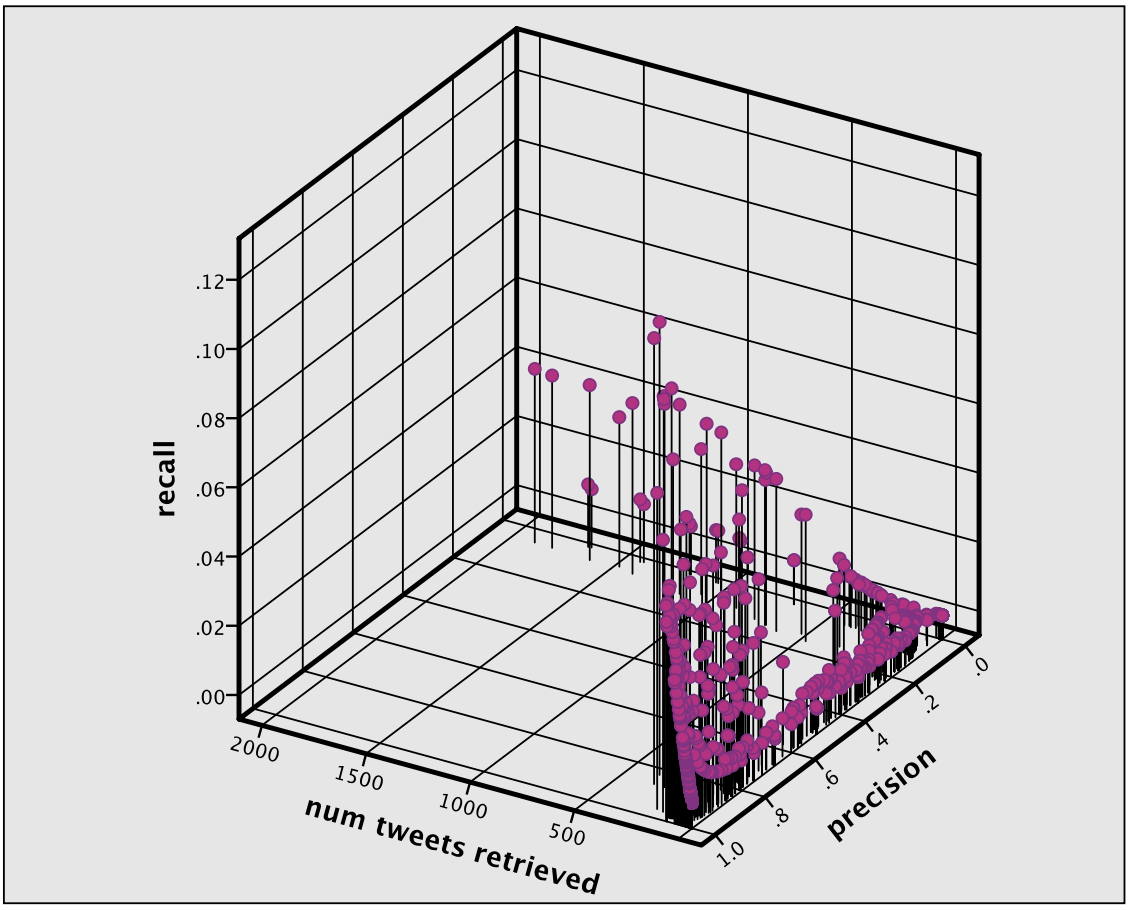
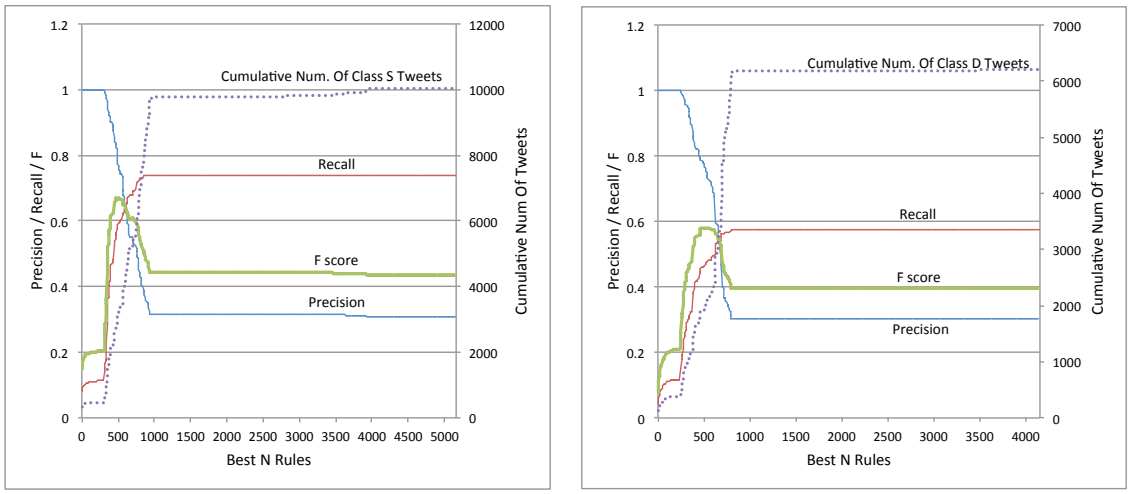


Figure 6.4. Performance of Class D rules on 2013-01-11

select the rules since we have thousands of rules. More specifically, we ask how to select the best n retrieval rules in this section. The selection of rules is important. It could reduce the redundancy between the rules. It could also boost the performance of the system by deleting the rules that have bad performance. In order to select the best rules, we first sort the rules by precision, then by number of tweets retrieved. For example, to select the best 1 rule, it should be the rule which has the highest precision and retrieved the most tweets. So we start from the top 1 rule to calculate the number of tweets retrieved, precision, recall and F scores. We add more rules to get the cumulative number of tweets retrieved (dropping duplicate tweets), and also calculate the precision, recall and F scores using the cumulative number of tweets retrieved. In our expectation, the cumulative number of tweets retrieved and recall should be increasing with increasing n. Then they should be stable, as it will become difficult to retrieve more relevant tweets. The precision should start from 1.0, then decrease with increasing n until stable.

Figure 6.5a and Figure 6.5b show the four metrics using top n rules on date 2012-12-30. Blue, red, and green solid lines represent precision, recall, and F score respectively. The purple dashed line represents the number of tweets retrieved. We found similar patterns from the figures. As we expected, precision starts from 1.0 and stays stable for about the top 300 rules, but the number of tweets retrieved is small. Those are the highest precision rules. For n from 300 to 500, the number of tweets retrieved and recall increase dramatically, while precision decreases suddenly. When n equals about 500, F score reaches its peak. When n reaches about 600-700,

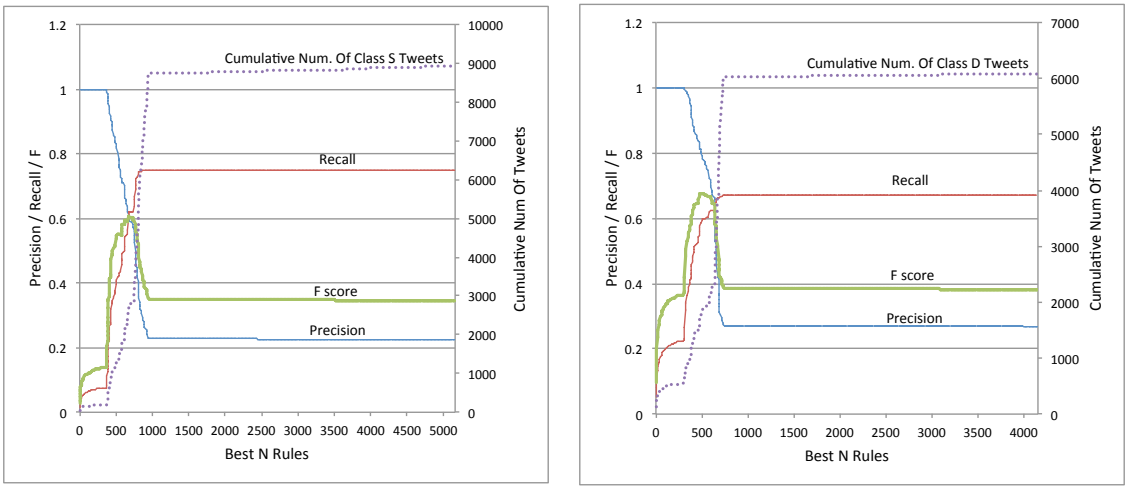


(a) Class S rules (b) Class D rules

Figure 6.5. Performance for best n rules on 2012-12-30

precision, recall and F score are the same number. At one point (about top 1,000 for Class S and top 900 for Class D), almost no more tweets are retrieved by more rules, and all the metrics become stable. Figure 6.6a and Figure 6.6b shows the four metrics using top n rules on date 2013-01-11. They also have the similar patterns with the figures on 2012-12-30.

Therefore, we conclude that to achieve the best performance, we could use roughly the top 600 Class S rules and top 650 Class D rules based on the two day's data. It is about 11% to 16 % of rules out of all the Class S or D rules. However, we could not simply remove the rules that retrieved nothing. The reason is that on different days, the same rule may retrieve a different number of tweets. Therefore, we would like to see if the rules are likely to provide consistent results on other days.



(a) Class S rules (b) Class D rules

Figure 6.6. Performance for best n rules on 2013-01-11

6.3 Consistency Of Retrieval Rule Results

Finally, we want to know if the rules are likely to yield consistent results from different days. As just mentioned the best n rules for Class S and Class D is about 600 and 650 respectively. But the top 600 rules for two days could be completely different. Therefore, we calculated the number of common rules in the top 600 for 2012-12-30 and 2013-01-11, then drew the Venn diagrams shown in Figure 6.7 and Figure 6.8. The graphs show most of the rules are common. For Class S rules, about 65% of the rules are common for the two days. For Class D rules, even more rules are common for the two days (75%). These results are encouraging.

To further analyze consistency of the retrieval rules, we randomly selected two more days: 2013-06-07 and 2014-03-05. We used all the rules to retrieve LS tweets from the two new days. Since we don't have the gold standard labels of the LS tweets on the new two days, we only analyze the number of tweets retrieved. We would

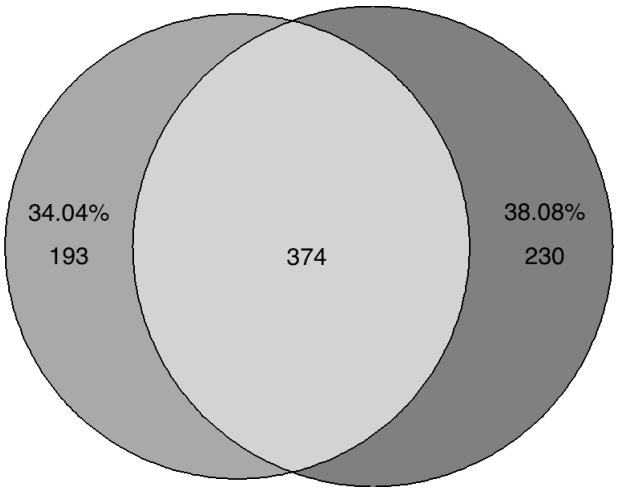


Figure 6.7. Common Class S rules for 2012-12-30 and 2013-01-11

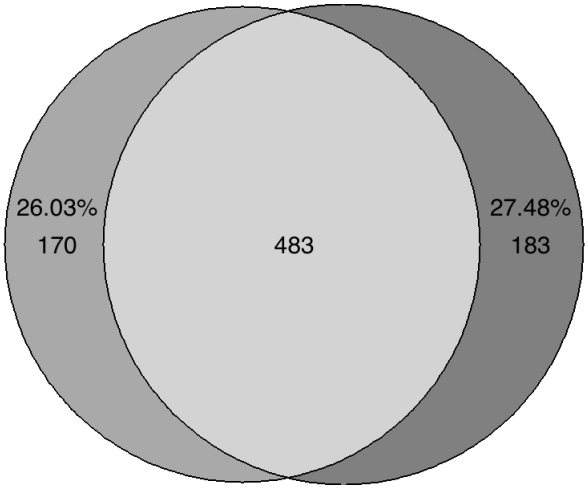


Figure 6.8. Common Class D rules for 2012-12-30 and 2013-01-11

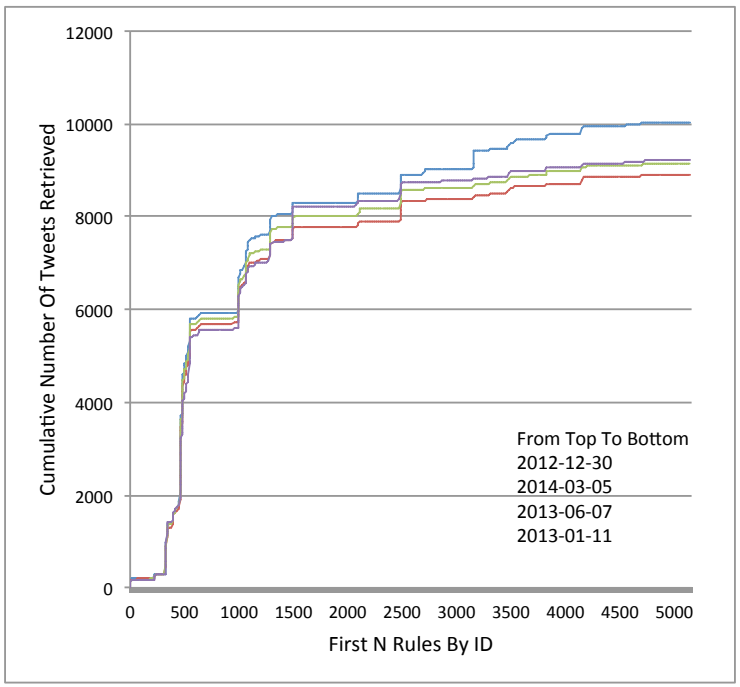


Figure 6.9. Cumulative number of tweets retrieved by Class S rules

expect that relatively speaking each rule retrieves a similar number of LS tweets on different days. The total numbers of LS tweets retrieved on different days using all the rules should also be similar. So we sorted the rules by rule ID, then we calculated the cumulative number of tweets retrieved by using the first 1 rule, the first 2 rules, to the last rule rules. The results from two gold standard days and two new days are shown in Figure 6.9 and Figure 6.10 for Class S and Class D respectively. The X-axis shows the first n rules used, the Y-axis shows the cumulative number of LS tweets retrieved. E.g. the first 500 Class S rules retrieve 4,000 tweets for each of the four days. The first 1,000 Class S rules retrieve close to 6,000 tweets for each of the four days. Therefore, we found that the total number of tweets retrieved varied little for different days, and the overall cumulative trends are also similar.

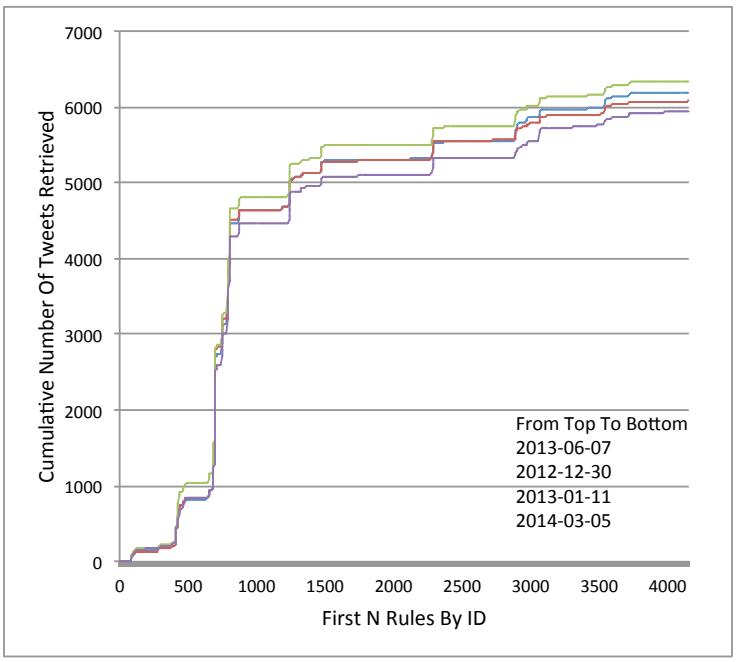


Figure 6.10. Cumulative number of tweets retrieved by Class D rules

In conclusion, the results show that about 20% of our retrieval rules actually retrieve LS tweets. From the two-day gold standard datasets, we found that using about top 600 rules (about 11%-16%) could achieve the highest F score. We also found our retrieval rules have consistent results in number of tweets retrieved for the four days of data.

CHAPTER 7

VALIDITY OF SURVEILLANCE STRATEGY

We introduced our survey-to-surveillance method and its evaluation in the previous chapters of this thesis. Surveillance on social media is just the first step for finding interesting information for social media text. In the later part of the thesis, we use life satisfaction as a case study to illustrate how did we conduct a comprehensive analyses on social media.

Before we discuss the analysis of LS tweets and LS users, we would like to discuss an assumption which most researchers seldom test. Most social media research assume that the post reflects the social media users' status, except some studies about sarcasm [27, 12].

The problem is that to do the analysis of the text on the internet is tricky. It is different from analyzing users with a survey. In a traditional survey, participants give answers according to the questions. Or in the biomedical domain, the patients' medical status are diagnosed by doctors. However, we know that people are anonymous online and they may post social media posts that they may never say in the real life, or the posts may not reflect the user's real status accurately. There is no guarantee that the social media posts really convey users' true status. This is less of an issue with answers from survey or doctor's diagnosis, e.g., if someone posted he was depressed but doesn't mean he really was depressed in the medical sense. If users posted they were satisfied or dissatisfied with their lives, are they really satisfied or

dissatisfied with their lives in a formal sense? Therefore, since we don't ask questions in passive social media surveillance, we have to make sure the social media posts can be used to infer users' life satisfaction. Therefore, we answer the questions in the chapter: Can we infer users' life satisfaction from their social media posts? If a user posted a Class S or D tweet, is the user satisfied with his/her life or not?

Previous studies haven't addressed this issue, as researchers usually assume that the users' social media posts can directly indicate their psychological status. Therefore, no study has tried to test this assumption. In this thesis, we are the first to attempt this kind of testing of assumption. Since different aspects (life satisfaction, depression, personality, etc) may need different testings, here we only show if we can use users' social media posts to infer their life satisfaction. In particular, we first detected LS tweets from Twitter, then we asked the users who posted those LS tweets to participate in the SWLS survey. We make this request within 24 hours of the post date and time. Then we see if our Class S users have significant better life satisfaction scores from formal survey than Class D users.

7.1 Dataset

We applied our survey-to-surveillance method to detect LS tweets every day. For each day, we selected the Class S and D tweets extracted using our high precision retrieval strategy. Then we randomly selected 120 Class S users and 120 Class D users as our target users. We were not able to survey more users because our Twitter accounts were blocked by Twitter if we sent a lot of invitations.

7.2 Survey Experiment

We wanted to conduct SWLS for the target Class S and D users. To avoid large emotional changes from the time they posted tweets and the time they participated in the survey, we sent the survey invitations one day after they tweeted LS tweets. To stimulate the target users to participate in the survey, we promised the chance to win a $50 Amazon gift card. The invitations were sent by our six Twitter accounts. To avoid Twitter blocking our accounts, the time interval for sending invitation tweets was about 15 mins. We also made sure the invitation tweets were slightly different from each other to prevent being detected as robot accounts. In the invitation tweet, the user clicked our survey introduction webpage[1]. The introduction webpage shows the instructions of our survey and will verify if the participant is a valid twitter user, then the user will be directed to the survey webpage at Surveygizmo[2]. In our survey webpage, we had 5 statements from SWLS as the required questions. The estimated time for answering the survey was about 3 min. The design of survey experiment has IRB approval. Figure 7.1 shows the snapshot of one part of the survey webpage.

7.3 Results

The experiment started from Oct. 27th, 2014, and ended on Mar. 16th, 2015. In the nearly five months, we sent nearly 29,890 invitations. We received only 137 survey answers. The response rate is 0.46%. The response number for each day is shown in Figure 7.2.

---

[1]http://lifesatisfaction.herokuapp.com

[2]http://www.surveygizmo.com

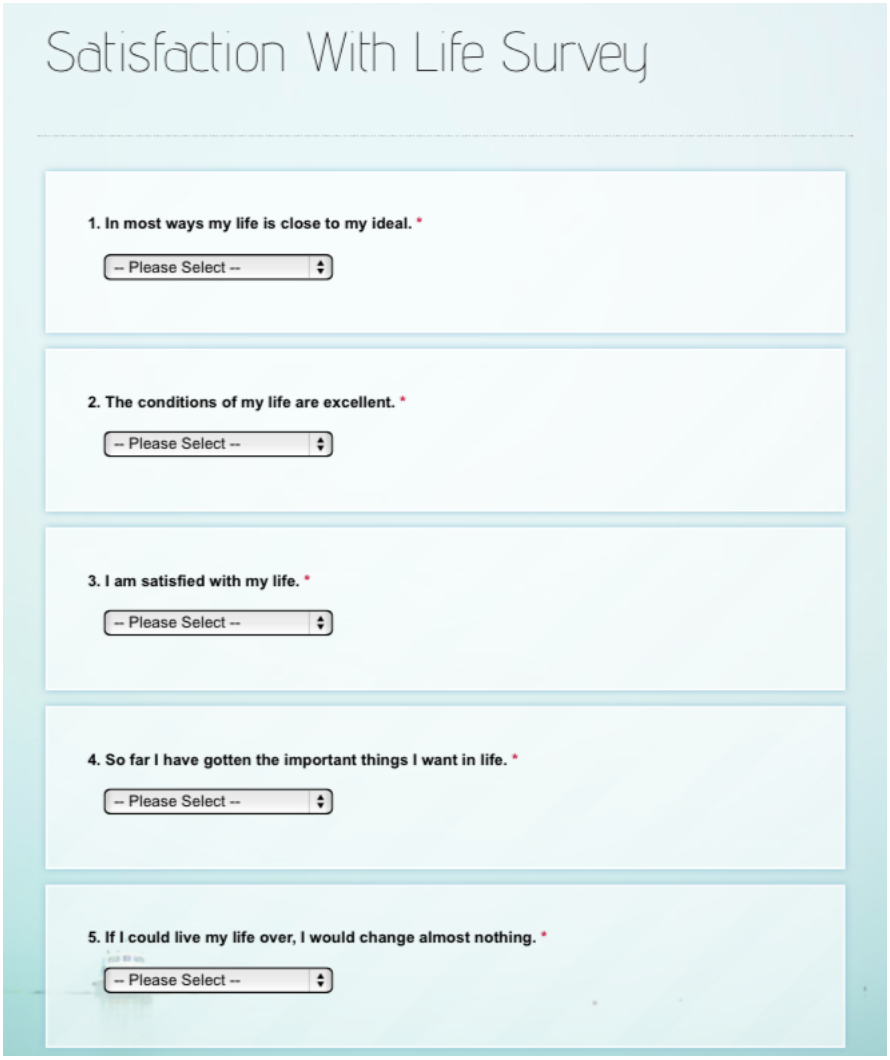


Figure 7.1. Snapshot of survey webpage

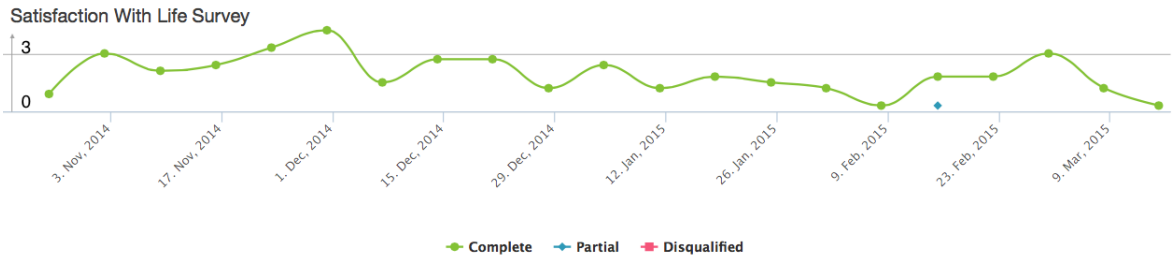


Figure 7.2. Survey response activity

Table 7.1. SWLS score interpretation

SWLS score	**Interpretation**
31 - 35	Extremely satisfied
26 - 30	Satisfied
21-25	Slightly satisfied
20	Neutral
15-19	Slightly dissatisfied
10-14	Dissatisfied
5-9	Extremely dissatisfied

Due to the retweet of some users, non-target users still can see the URL. Therefore, we filtered out the non-target users and the final number of valid answers is 104. The SWLS is a score ranging from 5 to 35. The interpretation of the SWLS scores is shown in Table 7.1

We calculated the means and standard deviations of SWLS scores for Class S and D users. We also did independent t-test for the SWLS scores for the two classes of users. Summary of the results is shown in Table 7.2. We found the SWLS scores for two classes are significantly different with p value < 0.0001. Therefore, the result shows that the two classes of users have significantly different SWLS scores. Because the SWLS score mean of Class S users is higher than the mean of Class D users, we could conclude that the predicted Class S users have better life satisfaction than predicted Class D users. In addition, we show the selected users which have contradictions between their LS tweets and SWLS scores in Table 7.3. The contradictions warn us there are situations that the users' life satisfaction cannot be inferred from their LS tweets. In addition, we found generally Class S tweets can infer Class S users better than Class D tweets.

Table 7.2. Summary of survey results

	Predicted Class S Users	**Predicted Class D Users**
# users invited	14,945	14,945
# participants	48	56
Response rate	0.32%	0.37%
SWLS score mean	24.35	19.63
SWLS score std.	6.29	6.67
t test (p value)	**<0.0001**	

Table 7.3. Selected contradictions

User Tweet	**Predicted Class**	**SWLS score**
I hate myself	Class D	21
I have no life	Class D	27
My life is sad	Class D	27
I'm stressed out	Class D	28
I love my life. #LivingLifeToTheFullest	Class S	12
My life is great	Class S	17

In conclusion, the result supports that we could infer uses' life satisfaction using their tweets. Also it is the further validation of life satisfaction surveillance on social media using our method. Again, this testing has not been done by the previous studies. The findings also warn us it is risky to assume that the social media posts written by users necessarily reflect reflect the users' real meaning or that social media posts can be used directly to infer user information before any testing.

CHAPTER 8

ANALYSIS OF LS TWEETS

Now that we have developed a surveillance method for tracking expressions of life satisfaction on Twitter, we would like to analyze the output from such surveillance. In this chapter we present our observations regarding trends in such expressions extracted from a 24 month time period. In the next chapter we present analysis of the users behind the life satisfaction posts.

8.1 First Person Twitter Dataset

We first collected first person (FP) tweets from Oct. 2012 to Oct. 2014 (about two years) using the Twitter Streaming API. Again, each collected tweet must have "I," "me," "my," or "mine" in order to be a FP tweet. We collected all metadata such as "tweet id," "text," "utc_time," "utc_offset," "geo," "place," "user id," "user name," etc., and user details such as "number of followers," "number of friends (followings)," "number of tweets," "location field," etc. We call this dataset FPTweets2Years. Table 8.1 shows a description of FPTweets2Years.

8.1.1 Validity Checks For FP Tweets Dataset

Twitter documentation states that the tweets from the Streaming API are a real-time random sample from the complete Twitter corpus. We did validity checks using intuitive notions. Similar checks have been performed by Dodds et al. [20] These were done on the January 2013 subset of our data. We call the subset FPTweetsJan13.

Figure 8.1 shows that the peak occurrences for the words "morning," "noon"

Table 8.1. Description of FPTweets2Years dataset

Time span	2012-01-03 to 2014-09-30
# tweets	About 3 billion
# tweets per day	About 4 million
# unique users per day	About 3 million
# tweets with geo tags per day	About 100,000
# replies per day	About 600,000
# retweets per day	About 1 million
# URLs per day (non-unique)	About 1 million

Table 8.2. Description of LSTweets2Years

Time Span	2012-10-03 to 2014-09-30
# Class S tweets	1,945,198 (0.065%)
# Class D tweets	1,908,571 (0.064%)
# LS tweets per day	about 4,750

and "evening" are as expected. Peak occurrences for "breakfast," "lunch," and "dinner" are also as expected, see Figure 8.2 . These checks raise our confidence in our dataset.

8.1.2 Surveillance Results And Description Of All Datasets

We ran our surveillance method to detect LS tweets from the two year dataset. The collection of all LS tweets detected is called LSTweets2Years. Table 8.2 describes this dataset.

As we expected, the proportion of LS tweets is very low. From three billion FP tweets, we only detected about four million LS tweets (about 0.13%). The number of Class S tweets is slightly more than the Class D tweets.

We list all the datasets in Table 8.3 before the analysis which should be

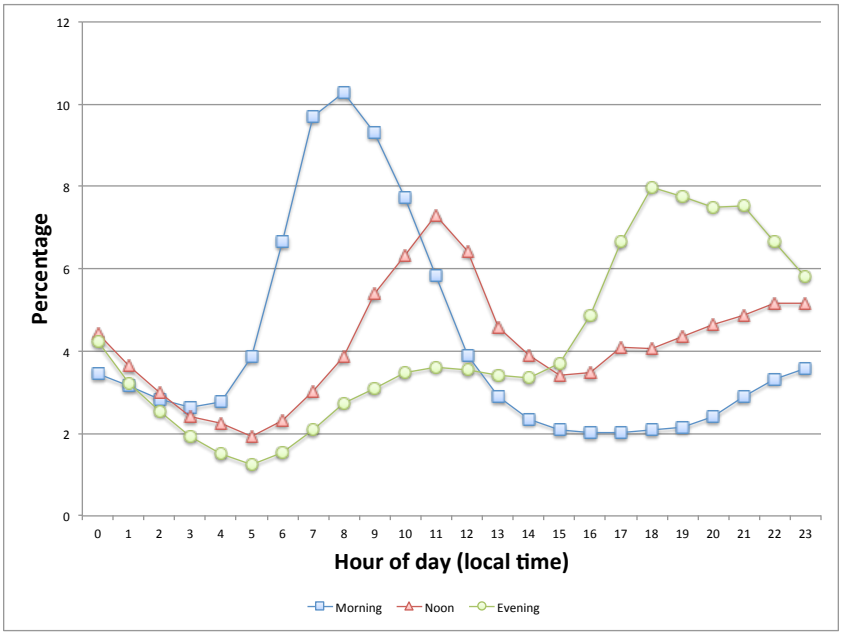


Figure 8.1. Hourly distribution for the words "morning," "noon," and "evening" (January 2013 subset)

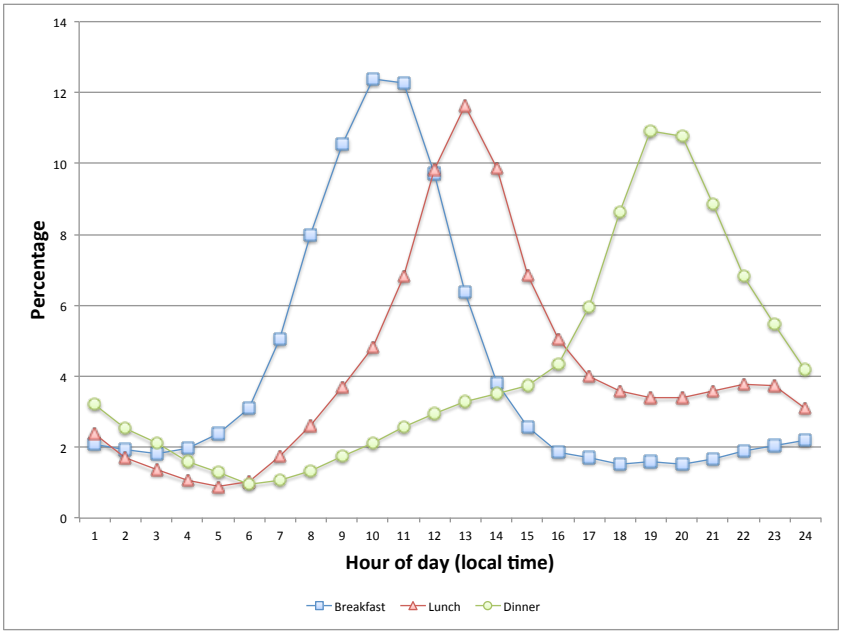


Figure 8.2. Hourly distribution for the words "breakfast," "lunch," and "dinner" (January 2013 subset)

more convenient for the readers. Overall, GS2Days is our gold standard dataset we have introduced before. FPTweets2Years and FPTweetsJan13 are the datasets that have FP tweets from Twitter Streaming API. LSTweets2Years and LSTweets1Year are the datasets that have LS tweets detected from our method. LSUsers2Years and LSUsers1Year are the datasets that have LS users. LSUsersMarApr04 and LSUser2YearsSelect are the datasets that have LS users, and all their tweets were crawled in order to do user level analyses. The table provides links to the relevant chapter or section. Next, we present our analyses of the life satisfaction tweets.

8.2 Time Series Analysis Of LS Tweets

To understand life satisfaction expressions on Twitter, we would like to know: What does the time series of LS tweets look like over two years? Does the time series show distinct patterns? Do holidays, local and world events influence the time series? We used two years' LS tweets dataset (LSTweets2Years) to do the analysis. We calculated the percentage of Class S or Class D tweets over all the FP tweets for each day for two years and plot their distribution in Figure 8.3. The orange and blue lines represent the time series of Class S and Class D tweets respectively.

Overall we observe that the two time series only show random fluctuations. Events such as political, economic events, and seasons do not appear to influence these distributions, with very few exceptions. There are significantly more Class S tweets than Class D tweets posted on Christmas days: 2012-12-25 and 2013-12-25. Also on Valentine's days, we found more Class S tweets than Class D tweets. We cannot explain the dip of Class D on 2013-01-26 and the peak of Class D on 2014-06-16. Also

Table 8.3. Description of all datasets used in thesis

Name	**Desc.**	**Time**	**Size**	**Usage**
GS2Days	The gold standard life satisfaction dataset	2012-12-30, 2013-01-11	8.5 mil.	Evaluation of surveillance method (Chap. 5)
FPTweets2Years	Tweets from Streaming API	2012-10-03 to 2014-09-30	3 bil.	Surveillance of life satisfaction on Twitter (Sec. 8.1)
FPTweetsJan13	Subset of FPTweets2Years	2013-01	125 mil.	Validity check of FPTweets2Years (Sec. 8.1.1)
LSTweets2Years	LS tweets detected by surveillance method from FPTweets2Years	2012-10-03 to 2014-09-30	1.9 mil. for each Class S & D	Time series of life satisfaction on Twitter (Sec. 8.2)
LSTweets1Year	Subset of LSTweets2Years	2012-10-03 to 2013-09-30	0.9 mil. for each Class S & D	Daily and weekly cycle of life satisfaction on Twitter (Sec. 8.3)
LSUsers2Years	LS users from LSTweets2Years	2012-10-03 to 2014-09-30	1.9 mil. for each Class S & D	Chap. 9
LSUsers1Year	LS users from LSTweets1Year	2012-10-03 to 2013-09-30	0.9 mil. for each Class S & D	1) Distribution of num. of LS tweets posted by each user (Chap. 9) 2) CCDF of followers and followings (Sec. 9.3.1) 3) Location of LS users (Chap. 11)
LSUsersMarApr04	LS users randomly detected in Mar. & Apr. 2014 (All tweets were crawled for each user)	2014-03 to 2014-04	15k for each Class S & D	Characteristics Analysis (Sec. 9.3)
LSUser2YearsSelect	One part of LS users in LSUsers2Years were selected (All tweets were crawled for each user)	2012-10 to 2014-09	30k LS users	Change of life satisfaction (Sec. 9.2)
LSUser4Groups	Subset of LSUser2YearsSelect	2012-10 to 2014-09	1.3k for each of 4 groups of users	Factors Associated With Life Satisfaction (Chap. 10)

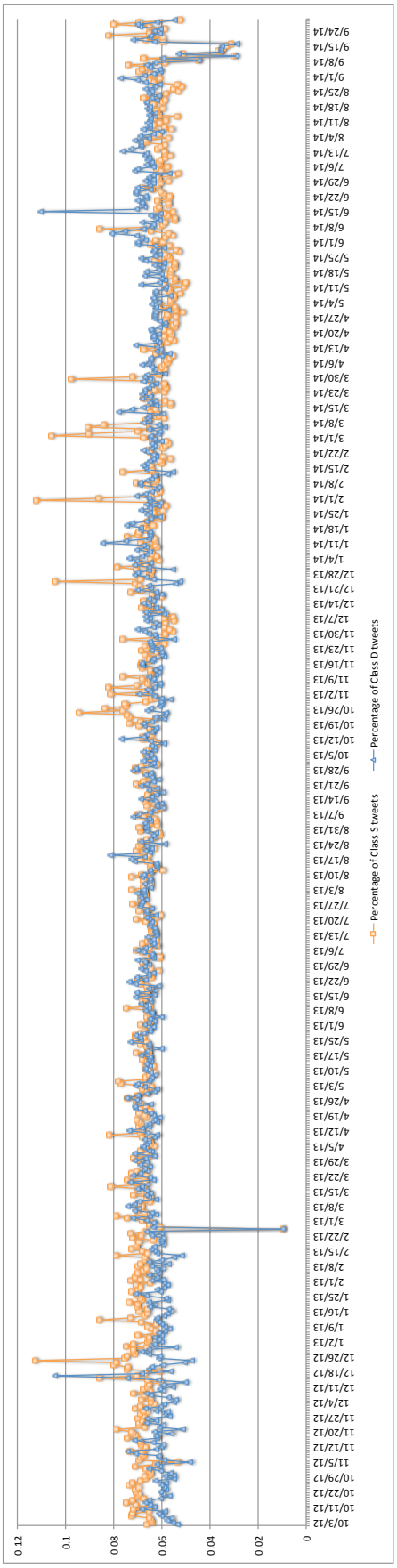


Figure 8.3. Two year's time series for percentage of LS tweets $\frac{\#\text{ Class S/D Tweets}}{\#\text{First Person Tweets}} \times 100\%$ per day

we cannot explain the peaks of Class S on 2013-10-25, 2014-02-01, 2014-03-03, and 2014-03-30. In addition, the time series from 2013-03 to 2013-10 is slightly different from the time series after 2014-04. This could have several explanations. For example, the algorithm of streaming tweets from Twitter changed. The Twitter streaming API returns a random sampled 1% of tweets. At some points the streaming API returned many more tweets. For example, in 2014, the streaming API returned about 4 million tweets a day. After sometime in March 2014, the streaming API suddenly returned 16 million tweets a day. To make the data comparable, we selected 4 million tweets from the 16 million tweets. For every 4 tweets, we selected the first one. Therefore, the time series after March 2014 could be slightly different from the date before. We observed the streaming API returned 4 million tweets again after Feb. 2015.

We present a cumulative time series for LS in Figure 8.5. Similar to the previous figure, the orange and blue lines represent cumulative time series of Class S and Class D respectively. The green line represents the cumulative time series of LS tweets (Class S + Class D tweets). From the figure, we clearly see that the 3 lines are almost straight. It means that the number of LS tweets detected in Twitter is generally stable and it is not easily affected by the seasonal change or other major events.

In summary, except for these two holidays, Christmas and Valentines day, the percentage of LS tweets didn't seem to be influenced by major events. This is considerably different from previous research on "happiness" trends where the researchers were trying to find the correlation between happiness and events. Dodds et al. found

that the average happiness on Twitter is affected by major holidays and other events. Their figure 8.4[1] shows their time series of "happiness" though the same period, from Oct, 2012, to Sep, 2014. They tried to find out what events caused the fluctuation of the "happiness" on Twitter. For example, they found "happiness" peaked on Mother's day, Father's day, Independence day, Christmas, etc. It dipped on the days which have events like the bombing at the Boston marathon, the arrest of Justin Bieber, Germany beats Brazil in World Cup, etc. Our results indicate that LS is more resilient to events than happiness as measured by Dodds et al.

8.3 Daily And Weekly Analysis Of LS Tweets

Next, we would like to know when people express their opinions of life satisfaction the most. In particular, which hour in a day and which day of the week has the most Class S and Class D tweets posted? The analyses in this section used the first year's LS tweet dataset: LSTweets1Year.

We first calculated the local time for all LS tweets and FP tweets. To get the daily cycle (over hours) of LS, with each day's data, we calculate the percentage of Class S, Class D, and all FP tweets posted in each hour. We then calculate the average of the percentages for each hour across the dataset. We show them as the daily cycle (over hours) of LS in Figure 8.6. Since we did a macro average, the error bars in the figures show the standard error. The green line represents the average tweeting trend for FP tweets since it is calculated using all the FP tweets. It peaks around 9 - 10 pm and then gradually drops over the early morning hours, reaching

[1]The figure is obtained from http://hedonometer.org

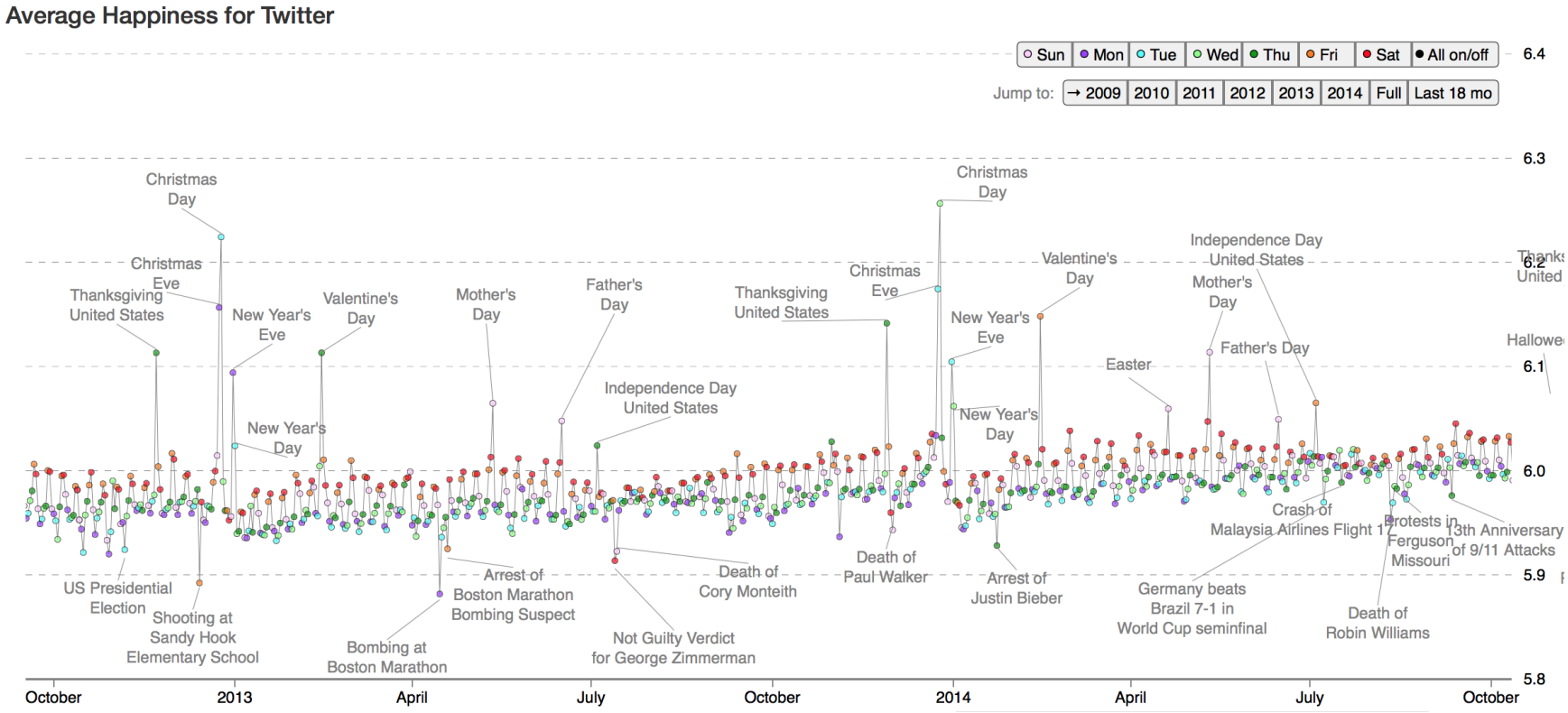


Figure 8.4. Time series for average happiness by Dodds et al. [20]

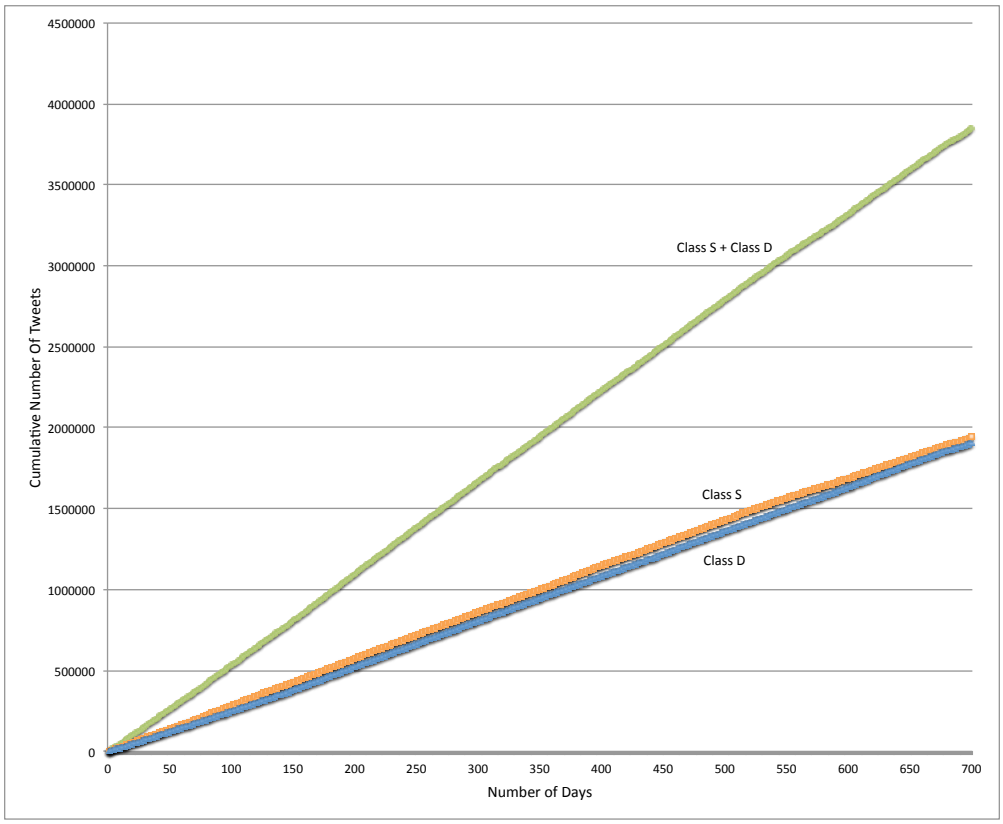


Figure 8.5. Cumulative number of LS tweets

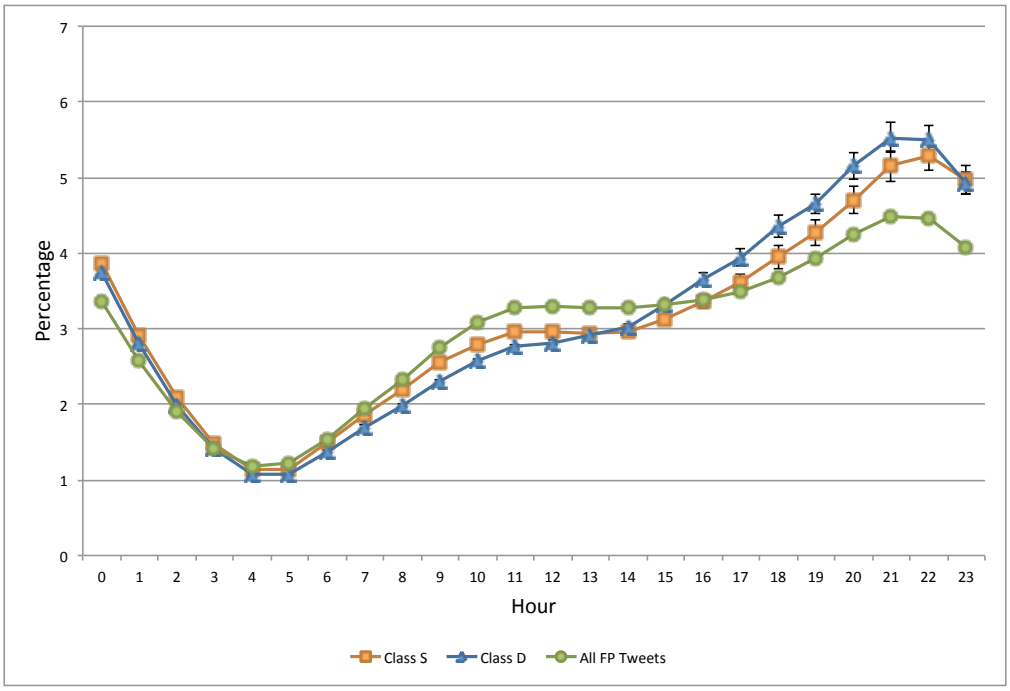


Figure 8.6. Daily cycle (over hours) of Class S, Class D, and FP tweets (macro avg.)

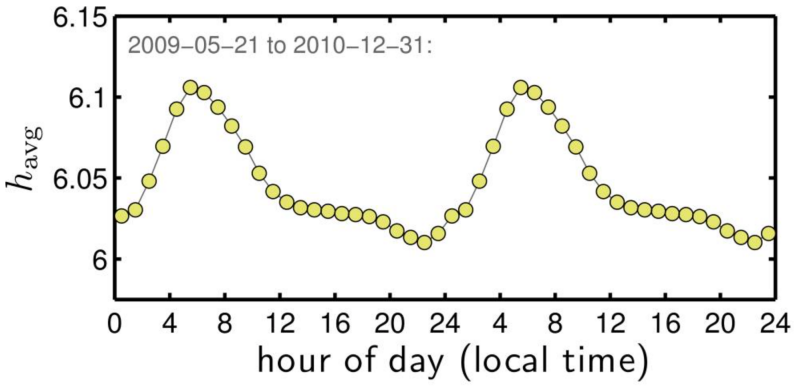


Figure 8.7. Daily cycle (over hours) of happiness on Twitter by Dodds et al. [20]

the lowest posting point in the early morning hours of 4 to 5 am. After this, the number of posts gradually rises, with leveling starting around the late morning hours (11 am to 2 pm roughly). This is within our expectation. The orange and blue lines which represent Class S and Class D tweets respectively also follow a similar trend. However, we found that LS tweets were posted less at noon and much more at night. For example, more than 5.5% of the Class D tweets were posted at time 22 (10 PM), while about 4.5% all FP tweets were posted at that time. If we only compare Class S and Class D, Class S tweets were posted more at noon while Class D tweets were posted more at night. It seems people express their opinion about their life satisfaction more at night, especially for the users who are not satisfied with their life. It is reasonable because most people are busy at working hours; they may think about their life more after work. As a point of comparison, Dodds et al.'s calculation in which they found 6 AM is the most happy hour and 11 PM is the least happy hour. Dodds et al.'s daily cycle of "happiness" on Twitter is shown in Figure 8.7.

To get the weekly cycle of LS, we calculated the number of Class S, Class D,

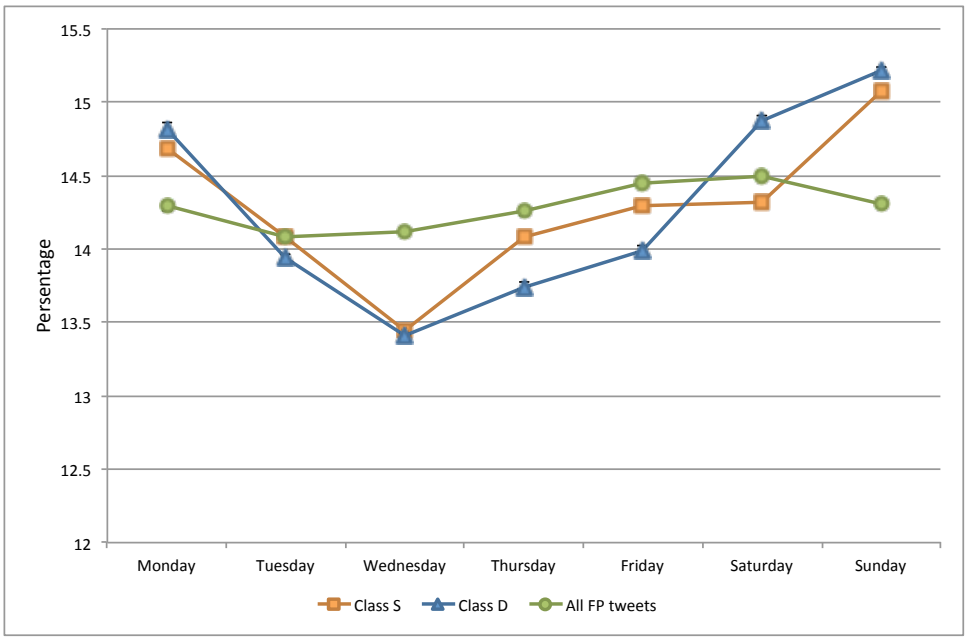


Figure 8.8. Weekly cycle of Class S, Class D, and FP tweets (macro avg.)

and all FP tweets posted in each week in the first year, and number of 3 types of tweets posted in each day. Then for each week, we have the percentage of Class S, Class D, and all FP tweets posted in each day. For example, for all the Class S tweets in week 1, 14.5% of them are posted in Monday of week 1, and for all the Class S tweets in week 2, 14% of them are posted in Monday of week 2. Finally, we have the macro average of percentage of Class S, Class D, and all FP tweets in different week days. We show them as the weekly cycle of LS in Figure 8.8.

Again, the green line which represents the average tweeting trend for LS tweets shows people tweet slightly more on Saturday and less on Tuesday. There is a low point on Wednesday for both Class S and Class D tweets, while the high points are on Sundays and Mondays. The rise is somewhat steady from Wednesday with a slight

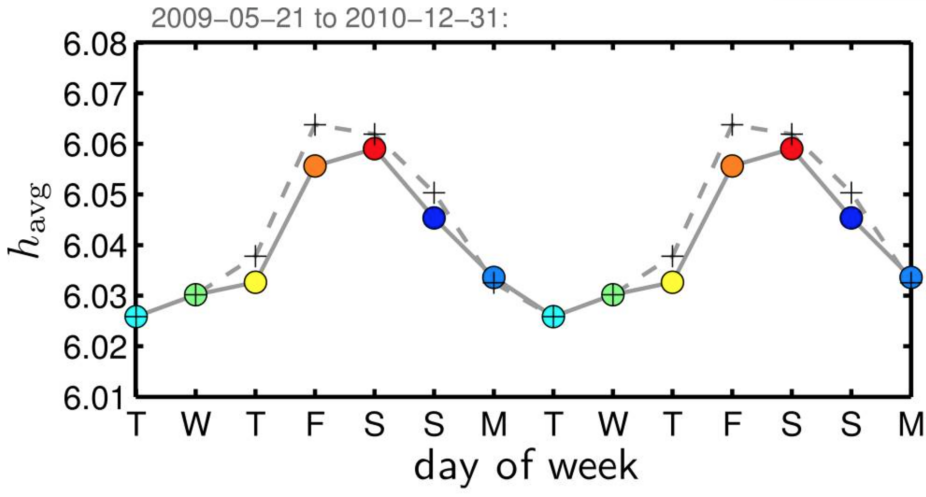


Figure 8.9. Weekly cycle of happiness on Twitter by Dodds et al. [20]

leveling off from Friday to Sunday. The activities for Class S and Class D are similar, except for Saturdays. About 15% Class D tweets were posted on Saturdays, while only about 14.3% Class S tweets were posted on that day. As a point of comparison, Dodds et al. found that Saturday is the most happy day in the week and Tuesday is the least happy day. The weekly cycles of happiness on Twitter by Dodds et al. is shown in Figure 8.9.

In conclusion, although some days like Christmas have more LS tweets, overall the time series of life satisfaction seems to fluctuate randomly. In other words the prevalence of LS tweets does not appear to be affected by global or local events. These include elections, New Year Eves, and U.S. Independence Days. This is very different from the observations of researchers like Dodds et al.[20], Kim et al. [35], and Mitchell et al. [45]. They found that happiness on Twitter is correlated with events like Michael Jackson's death, U.S. Independence Days, etc. Thus we note

that on the same social medium, these two notions of subjective well being: affect (which Dodds et al. call happiness) and LS exhibit different patterns of occurrence. LS exhibits resistance to world events while affect does not. This is consistent with the definitions of these two components of SWB. Life satisfaction is a longer-term cognitive assessment of one's achievements of life goals. Affect reflects daily mood that can be shaped by daily events. Consistency with the literature on life satisfaction [52] boosts our confidence in our surveillance methods that derive from the SWLS survey.

For the daily and weekly cycle of life satisfaction, we found the LS tweets were posted less at noon and much more at night. Especially, more Class S tweets were posted at noon, while more Class D tweets were posted around 10 PM. We also found Wednesdays have the least LS tweets, while Sundays have the most LS tweets. Overall our surveillance method contributes to tracking one major component of subjective well being, which is life satisfaction, the longer term, more stable, and cognitive, self-assessment of one's life.

CHAPTER 9

ANALYSIS OF LS USERS

In the last chapter, we presented life satisfaction analysis at the tweet level. In order to better understand these expressions on social media, we explore life satisfaction at the user level in this chapter. We ask several questions. How many Class S and Class D users are in our dataset? Also, we would like to know if there are users posting both Class S and Class D tweets. If so how much time separates the posts of satisfaction and dissatisfaction. In addition, do Class S and Class D users differ in any respect? For example, could it be that Class S users write more positive words while Class D users write more negative words?

9.1 Description Of LS Users

We identified LS users using the LS tweets in LSTweets2Years dataset and built a LSUsers2Years dataset (and its subset LSUsers1Year) for this analysis. Table 9.1 shows the description of the LS users in LSUsers2Years dataset. We find that usually an LS user in our dataset tweets only one LS tweet. A tiny portion of LS users (0.067%) have two or more LS tweets. One reason for this may be that even though we received 4 million tweets a day from Twitter's Streaming API, it is still a very small sample. Even if one user posts an LS tweet every day, we could easily have missed all of the posts because of the 1% API limitation. We show how we overcome this limitation for some of the analysis in this chapter. Given this API limitation, we find that in the data we have collected that the overlap of Class S and Class D users

Table 9.1. Description of LS users for LSUsers2Years dataset

Time Span	2012-10-03 to 2014-09-30
# unique Class S users	1,943,832
# unique Class D users	1,907,363
# overlap between 2 user groups for the first year	143,833 (at least 15%)
# overlap between 2 user groups for the two years	327,627 (at least 17%)

is at least 15% for the first year (LSUsers1Year) and at least 17% for the whole two years (LSUsers2Years).

Figure 9.1 shows the distribution of the number of LS tweets posted per user using the LSUsers1Year dataset. The orange and blue lines represent Class S and Class D users respectively. The Y-axis is log scale. We see that most users only post one LS tweet. Only about 400 Class S and 600 Class D users have five LS tweets. For the users who have less than eight LS tweets, the distributions are similar for the two classes. Again, this analysis is limited to the constraints of the 1% API sample.

9.2 Users Who Changed Their Life Satisfaction

We are going to focus on the users who posted both Class S and D tweets in this section. Again, we do find there are about at least 17% LS users who changed their opinions about their life satisfaction. Since life satisfaction is defined as somewhat stable, we won't expect users change their minds in a short time. To better understand this phenomenon, we would like to answer the following questions: 1) How much time separates the posts of satisfaction and dissatisfaction. 2) Are there users who change their minds more than once and if so how frequently?

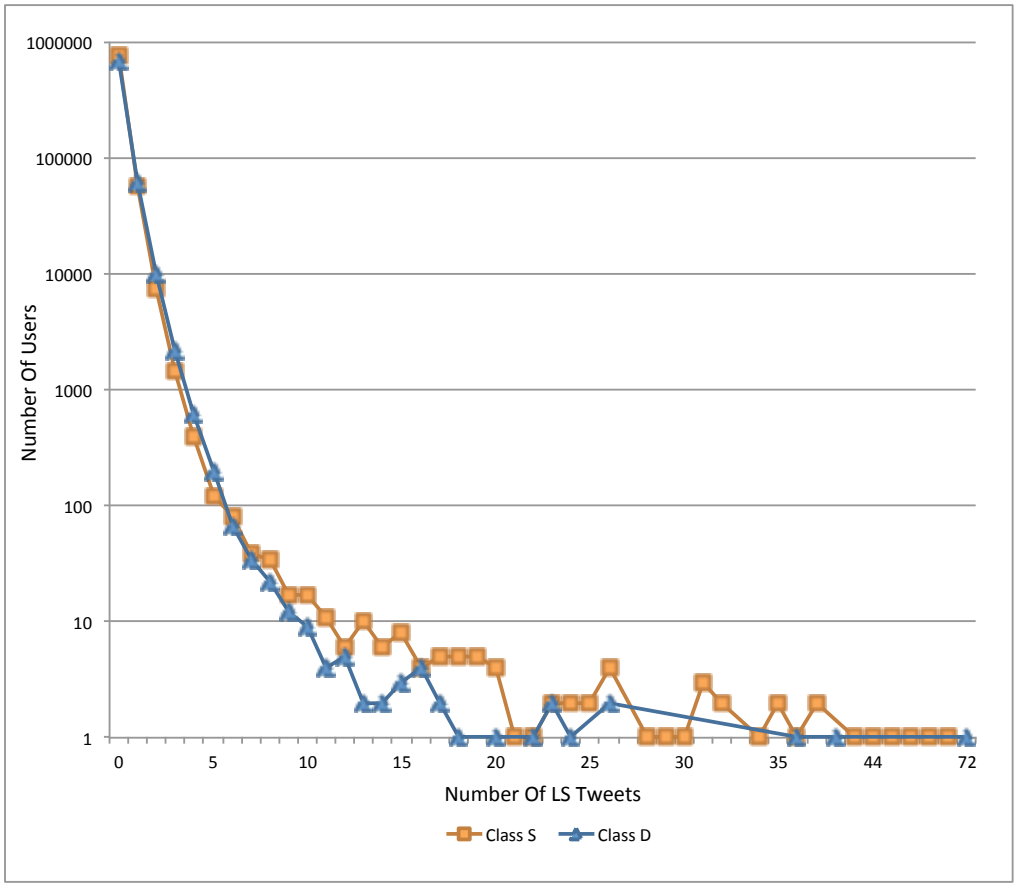


Figure 9.1. Distribution of LS tweets per user

In our LSUsers2Years dataset, we only have LS tweets for the users. So even if we found one user that posted one Class S tweet and one Class D tweet, we cannot directly analyze the time of changing opinion, because FPTweets2Years is a random sample from Twitter Streaming API. Therefore, we built a new dataset called LSUser2YearsSelect. We selected all the high precision users who changed their life satisfaction in LSUsers2Years. We also randomly selected a subset of high precision Class S and Class D users in LSUsers2Years. Then we crawled all their tweets for the selected users. High precision users are those detected by high precision retrieval strategies (A-W1 in Figure 5.2 and 5.3), so that we can be confident in their LS status. Then we indexed all theirs tweets and applied our surveillance method to find out all the LS tweets posted by those selected users. In summary, the total number of users we crawled is 31,706. The total number of tweets we crawled is 79,443,500 for those users. We calculated how often do users change their status in life satisfaction in LSUser2YearsSelect dataset and summarized the result in Table 9.2.

We define several types of users. Group S: Class S users who never changed to Class D. Group D: Class D users who never changed to Class S. Group S⇒D: LS users who were Class S then changed to Class D. Group D⇒S: LS users who were Class D then changed to Class S. The rest are the LS users who changed their mind more than one time. For example, the users who changed 2 times could have 2 situations. One is they began with S, then changed to D, and finally changed to S again (S⇒D⇒S). Another one is they were D, then changed to S, and finally changed to D again (D⇒S⇒D). For Group S⇒D and D⇒S, the average number of

Table 9.2. Description of users who changed life satisfaction

User Type	**Num. Users**	**Change Interval (Mean)**	**Change Interval (Std.)**
S	13,528	N/A	N/A
D	11,923	N/A	N/A
S⇒D	2,186	82	116
D⇒S	2,175	85	110
change 2 times	1,332	54	74
change 3 times	347	37	47
change 4 times	143	26	30
change 5 times	34	12	12
change 6 times	18	7	6
change 7 times	12	11	8
change 8 times	8	10	10
change 9 times	1	3	0.00
change 10 times	1	5	0.00

intervening days is about 80, and the standard deviation is about 110. It makes sense since life satisfaction should be a stable variable, so we won't expect users to change their minds in a short time. We checked the person who changed his life satisfaction 10 times. It turns out he posted multiple times of tweets like "I love my life," "I hate myself" at different times.

Then we would like to see if different groups of users have similar life length in Twitter. E.g., if the average Twitter life of Group S and D users is less than 80 days, they may change their life satisfaction later. Therefore, we calculated the average Twitter life (days) for the 4 groups of users and show the results in Table 9.3. We found the 4 groups of users have similar life length in Twitter and importantly they are well above 1 standard deviation of the mean change interval.

Since most of the users who changed LS only change once, we calculated the

Table 9.3. Average Twitter life for four group types

User Type	Number Of Users	Average Life In Twitter (Std.))
S	13,528	1,180 (457) days
D	11,923	1,104 (467) days
S⇒D	2,186	1,111 (452) days
D⇒S	2,175	1,134 (448) days

distribution of change interval for these two types of users and show them in Figure 9.2 and Figure 9.3 respectively. The change intervals were more than one month for more than half of the Group S⇒D / D⇒S users. However, there were a lot of users who changed their life satisfaction in one month.

9.3 Differences Between Class S And D Users

We would like to know what are the differences between the Class S and Class D users. In this section, we first explore the number of followers and followings since these two are important metadata in social media analysis. Then we do a comprehensive comparison of other characteristics of the two classes of users. Finally, we show the differences of their topics of interest. Users who posted both Class S and D tweets are excluded in this section.

9.3.1 Followers And Followings

We did this analysis using LSUsers1Year dataset, and present the Complementary Cumulative Distribution Function (CCDF) in Figure 9.4 and Figure 9.5. The orange dots represent Class S users and the blue dots represent Class D as before. The X- and Y-axes are log scale. To interpret the figures, we take Figure 9.4 as example. The orange and blue dots (overlap at coordinate (10^0, 10^0)) at the top left

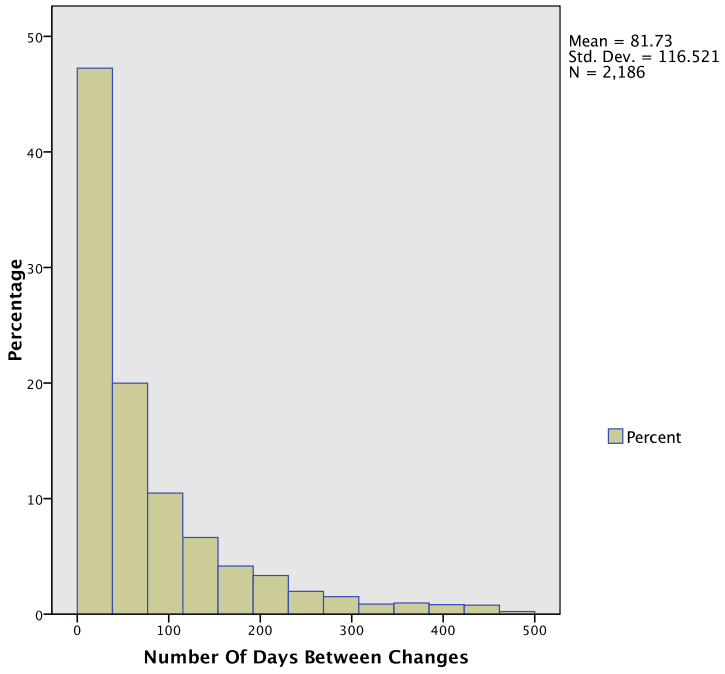


Figure 9.2. Distribution of change interval (Group S⇒D)

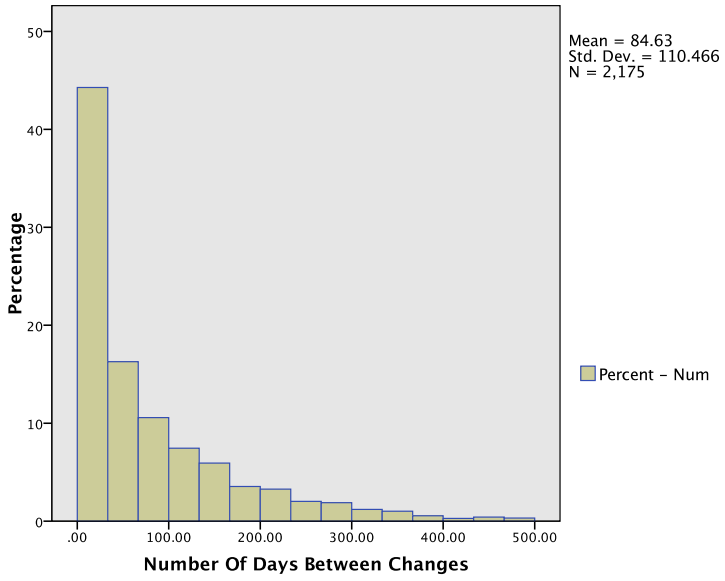


Figure 9.3. Distribution of change interval (Group D⇒S)

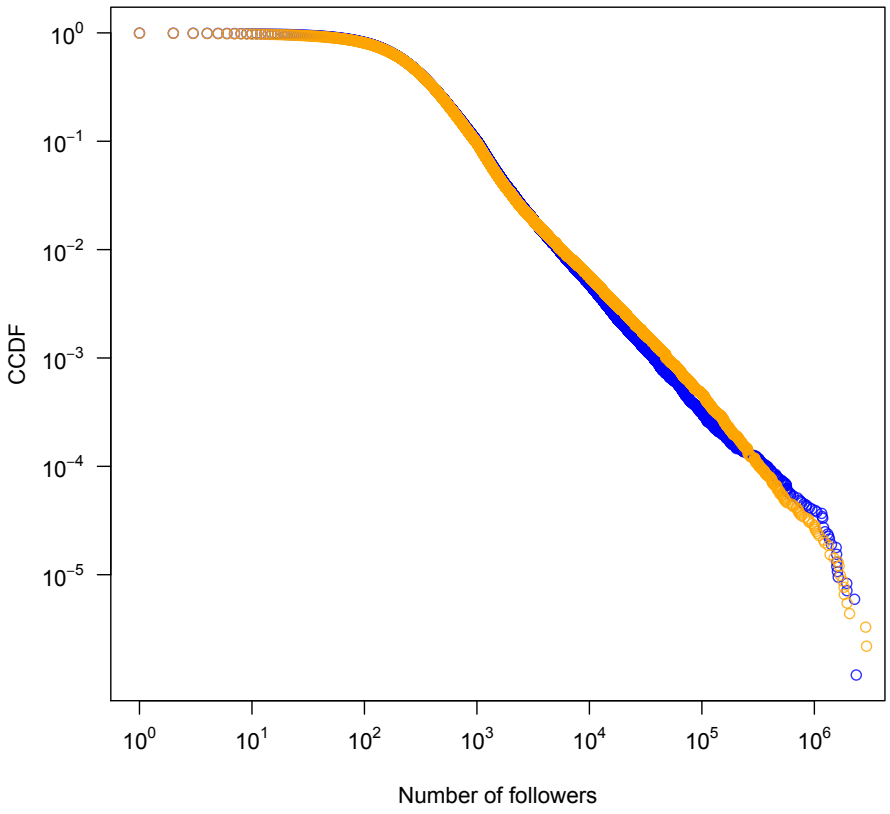


Figure 9.4. CCDF for number of followers

show that the probability that LS users have greater than or equal to 1 followers is 1. It means all the LS users have at least one follower. The dots at X-axis 10^2 show that most of the users have more than 100 followers. The probability of LS users having more than 1000 followers is about 0.1. From the two figures, we see that there is little difference between the two groups for the number of followers and followings.

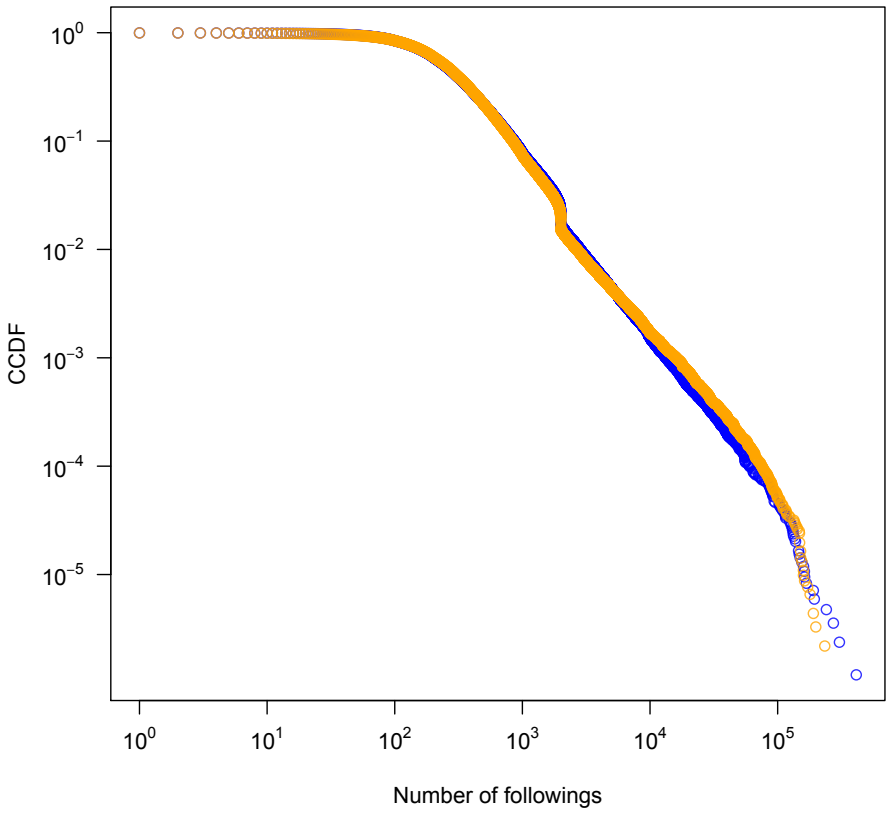


Figure 9.5. CCDF for number of followings

9.3.2 Twitter Metadata

There are many metadata for Twitter users including number of tweets, hashtags, etc. We want to see whether the pattern of occurrences differs between the two classes of users. It is insufficient to use LSUsers1Year or LSUsers2Years dataset, because most of the users only have one tweet in this dataset as we mentioned before. Therefore, we need to build a new dataset in which we have all the tweets for at least a sample of the users. Then we could extract their characteristics from their text and do the analysis.

To build the new dataset, we first selected two classes of users detected in March and April 2014. In the two months worth of data, we randomly selected 14,506 Class S users and 14,743 Class D users out of 360,000 LS users for the two months. (The number of users crawled depends on the rate limits of Twitter API.) We crawled all the tweets from each user. (The maximum number of tweets that can be crawled using Twitter API is 3,200.) This dataset is called LSUsersMarApr04 which is described in Table 8.3. Then we aggregated all the tweets for each user as one pseudo text document. Finally, we compared users' characteristics including Twitter metadata information and other characteristics calculated from their documents.

We explored 12 Twitter metadata, including number of tweets, user active days (number of days between the first and last tweet), the hour which has the most frequent tweets, number of URLs, number of hashtags, etc. The comparison is shown in Table 9.4. The table includes means of the values. We also tested the significance of the differences of the two means using independent t test [53] and the p values are

Table 9.4. Comparison of Twitter metadata for Class S and Class D users

Categories having more than 10% differences are highlighted.

Metadata	**Class S Avg.**	**Class D Avg.**	**Sig. of t test**	**Diff (%)**
# tweets	2,301.082	2,401.436	0.000	-4.361
User active days	***249.086***	***197.983***	***0.000***	***20.516***
Most frequent hour	10.897	11.161	0.007	-2.422
# urls	***347.645***	***383.490***	***0.000***	***-10.311***
# hashtags	***205.923***	***158.456***	***0.000***	***23.051***
# unique hashtags	***107.750***	***84.550***	***0.000***	***21.531***
# retweets	638.166	674.499	0.000	-5.693
Average tweet size	9.857	9.830	0.247	0.268
# emoticons	1,173.975	1,265.817	0.000	-7.823
# pos emoticons	159.638	160.453	0.752	-0.510
# neg emoticons	306.096	292.569	0.011	4.419
# unique words	5,876.278	5,903.063	0.373	-0.456

shown in the table. Most are significantly different (p value < 0.05). We italicize the metadata with more than 10% difference. We found that Class S users have more active days and number of hashtags, while Class D users have a higher number of URLs. We checked particular hashtags and URLs from the two classes, and we found they are similar for the two classes of users, e.g., the frequencies for "#sad" in both classes are about 1,000.

9.3.3 Linguistic and Topic Differences

We also want to know: What topics do the two classes of users tweet about (this question is not limited to LS tweets)? Are there topic differences? There are different ways to detect topics; one of the most popular methods now is LDA [6]. It has been used in different areas for text data [28, 63, 61]. But the drawback of this method is that we need to define how many topics we want and we need to manually examine the frequent words for each topic before we label the topics. Therefore, we

decided to use Linguistic Inquiry and Word Count (LIWC) to find the word usage of different pre-defined categories for the two classes of users. LIWC is a lexicon developed by Pennebaker et al.[48], and is one of the most popular lexicons used to calculate the degree to which people use different categories of words across a wide array of texts. It has about 80 categories such as "Positive/Negative emotion," and topical categories such as "Friends," "Work," and "Money." Some of the categories have hierarchical structures. We calculated the word count for the 80 categories for each user in LSUsersMarApr04 dataset using the LIWC lexicon.

Table 9.5 shows the comparison of LIWC linguistic categories for Class S and Class D users. Notice that LIWC categories are arranged hierarchically. Most categories are significantly different for the two classes. Again, we marked the categories which are 10% different for the two classes. Class S users have significantly more exclamations than Class D users. Class D users have significantly more personal pronouns, present tense, adverbs, conjunctions. The most interesting category is the "Swear words" which Class D users used significantly more than Class S users.

9.3.4 LIWC Psychological Processes

One important component in LIWC categories is "Psychological Processes." It includes "Social processes," "Affective processes," "Cognitive processes," "Perceptual processes," "Biological processes," and "Relativity." "Affective processes" has the positive emotion and negative emotion words which are very popular in sentiment analysis studies. Table 9.6 shows the comparison of LIWC Psychological Processes categories for the Class S and Class D users. Most of the categories are significantly

Table 9.5. Comparison of LIWC linguistic categories for Class S and Class D users

Categories having more than 10% differences are highlighted.

Categories	**Example**	**# of Words**	**Class S Avg.**	**Class D Avg.**	**Sig.**	**Diff (%)**
1 Word count			9.5	9.6	0.2	-1.0
2 QMark			211.8	225.5	0.0	-6.5
3 ***Exclam.***			***475.3***	***401.8***	***0.0***	***15.5***
4 Total function words		464	9,401.1	10,099.5	0.0	-7.4
4.1 ***Total pronouns***		116	***3,539.2***	***3,893.1***	***0.0***	***-10.0***
4.1.1 ***Personal pronouns***	***I, them, her***	70	***2,575.2***	***2,873.8***	***0.0***	***-11.6***
4.1.1.1 ***1st pers singular***	***I, me, mine***	12	***1,559.5***	***1,790.3***	***0.0***	***-14.8***
4.1.1.2 1st pers plural	We, us, our	12	113.9	115.7	0.1	-1.6
4.1.1.3 2nd person	You, your, thou	20	620.1	673.0	0.0	-8.5
4.1.1.4 3rd pers singular	She, her, him	17	181.8	195.0	0.0	-7.3
4.1.1.5 3rd pers plural	They, their	10	99.9	99.8	0.9	0.2
4.1.2 Impersonal pronouns	It, it's, those	46	964.0	1,019.3	0.0	-5.7
4.2 Articles	A, an, the	3	762.3	783.7	0.0	-2.8
4.3 Common verbs	Walk, went, see	383	3,010.1	3,292.5	0.0	-9.4
4.4 Auxiliary verbs	Am, will, have	144	1,741.3	1,895.4	0.0	-8.9
4.5 Past tense	Went, ran, had	145	440.7	466.8	0.0	-5.9
4.6 ***Present tense***	***Is, does, hear***	169	***2,159.7***	***2,392.9***	***0.0***	***-10.8***
4.7 Future tense	Will, gonna	48	187.6	203.3	0.0	-8.4
4.8 ***Adverbs***	***Very, really***	69	***940.2***	***1,047.6***	***0.0***	***-11.4***
4.9 Prepositions	To, with, above	60	1,686.7	1,718.9	0.0	-1.9
4.10 ***Conjunctions***	***And, but***	28	***838.6***	***957.2***	***0.0***	***-14.1***
4.11 Negations	No, not, never	57	461.4	476.3	0.0	-3.2
4.12 Quantifiers	Few, many	89	413.8	425.3	0.0	-2.8
4.13 Numbers	Second	34	110.1	118.6	0.0	-7.7
5 ***Swear words***	***Damn, piss***	53	***220.4***	***246.6***	***0.0***	***-11.9***

different for the two classes. Consistent with expectations, Class D users have significantly more Negative emotion words, especially Anger and Sadness words. Also Class D users mentioned significantly more Sexual words. Discrepancy category includes words like "should," "would," "expect," "hope," and "need." They may express determination and aspirations for the future. [58, 41]

One issue when using LIWC category is that we just know the word count without knowing the sentiment. For example, Class D users had more words in the "friends" category. However, we don't know if they have expressed negative or positive opinions to their friends.

To overcome this limitation, we did sentiment analysis for each category. We tested several sentiment analysis tools, and we found that the "pattern.en" module[1] for Python has the best performance at identifying the sentiment (positive, negative, and neutral) of tweets. It can also detect negation. In this experiment, for each user, we find the number of positive and negative tweets for each category (ignoring the neutral one). Then we calculated the percentage difference of positive and negative for each category. The formula is:

$$\frac{\#\ pos\ tweets_{in\_category\_x} - \#\ neg\ tweets_{in\_category\_x}}{\#\ total\ tweets_{in\_category\_x}} \times 100 \tag{9.1}$$

Finally for each category, we calculate the average percentage difference of positive and negative for Class S and D users and did the t test of mean difference. Figure 9.6 shows the categories which are more than 10% percentage difference of Class S and Class D users for each psychological processes category. There are some interesting

[1] http://www.clips.ua.ac.be/pages/pattern-en

Table 9.6. Comparison of LIWC psychological processes for Class S and Class D users

Categories having more than 10% differences are highlighted.

Categories	**Example**	**# of Words**	**Class S Avg.**	**Class D Avg.**	**Sig.**	**Diff (%)**
1 Social processes	Mate, talk, they, child	455	2,058.9	2,175.3	0.0	-5.7
1.1 Family	Daughter, husband	64	92.7	88.5	0.0	4.6
1.2 Friends	Buddy, friend, neighbor	37	52.9	55.7	0.0	-5.4
1.3 Humans	Adult, baby, boy	61	243.2	247.2	0.0	-1.6
2 Affective proc.	Happy, cried, abandon	915	1,801.3	1,932.4	0.0	-7.3
2.1 Positive emotion	Love, nice, sweet	406	1,165.7	1,220.8	0.0	-4.7
2.2 ***Negative emo.***	***Hurt, ugly, nasty***	499	***626.5***	***701.2***	***0.0***	***-11.9***
2.2.1 Anxiety	Worried, fearful	91	65.0	68.5	0.0	-5.3
2.2.2 ***Anger***	***Hate, kill, annoyed***	184	***323.1***	***364.9***	***0.0***	***-12.9***
2.2.3 ***Sadness***	***Crying, grief, sad***	101	***102.1***	***117.0***	***0.0***	***-14.7***
3 Cognitive processes	cause, know, ought	730	2,525.9	2,700.7	0.0	-6,9
3.1 Insight	think, know, consider	195	323.1	341.0	0.0	-5.5
3.2 Causation	because, effect, hence	108	259.8	282.5	0.0	-8.7
3.3 ***Discrepancy***	***should, would, could***	76	***350.6***	***395.2***	***0.0***	***-12.7***
3.4 Tentative	maybe, perhaps, guess	155	394.0	424.5	0.0	-7.7
3.5 Certainty	always, never	83	313.8	314.6	0.7	-0.3
3.6 Inhibition	block, constrain, stop	111	99.3	101.3	0.0	-2.0
3.7 Inclusive	And, with, include	18	525.6	561.6	0.0	-6.8
3.8 Exclusive	But, without, exclude	17	473.9	517.8	0.0	-9.3
4 Perceptual processes	Observing, heard	273	481.7	508.1	0.0	-5.5
4.1 See	View, saw, seen	72	190.4	203.6	0.0	-6.9
4.2 Hear	Listen, hearing	51	127.5	131.4	0.0	-3.1
4.3 Feel	Feels, touch	75	133.3	143.1	0.0	-7.4
5 Biological processes	Eat, blood, pain	567	667.5	731.6	0.0	-9.6
5.1 Body	Cheek, hands, spit	180	255.5	271.5	0.0	-6.3
5.2 Health	Clinic, flu, pill	236	118.3	124.8	0.0	-5.5
5.3 ***Sexual***	***Horny, love, incest***	96	***242.6***	***285.4***	***0.0***	***-17.6***
5.4 Ingestion	Dish, eat, pizza	111	104.0	105.0	0.3	-1.0
6 Relativity	Area, bend, exit, stop	638	2,344.1	2,388.7	0.0	-1.9
6.1 Motion	Arrive, car, go	168	368.3	398.0	0.0	-8.1
6.2 Space	Down, in, thin	220	873.1	869.4	0.5	0.4
6.3 Time	End, until, season	239	1044.7	1064.8	0.0	-2.0

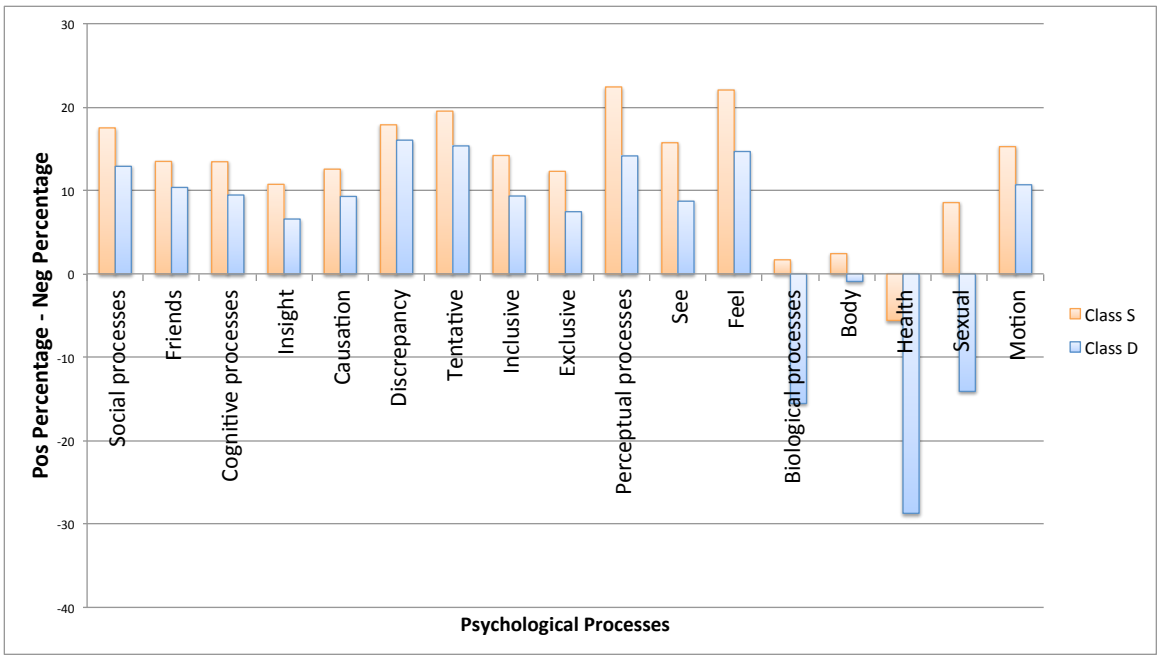


Figure 9.6. Comparison of sentiment for LIWC psychological processes

findings. For example, for Sexual categories, Class S users have more positive sentiment than negative, while Class D users have more negative sentiment than positive. For Health category, Class D users have much more negative sentiment. For most of the other categories, both classes of users have positive sentiment overall. Class S users have more positive sentiment than Class D users for all other categories.

9.3.5 LIWC Personal Concerns

Another component in LIWC is Personal Concerns. Table 9.7 shows the comparison of LIWC Personal Concerns categories for the Class S and Class D users. We didn't show the positive emotion and negative emotion categories like anger, sadness because they naturally have sentiment. All the categories are significantly different for the two classes except Home category. Money, Religion, and Death have more

Table 9.7. Comparison of LIWC personal concerns for Class S and Class D users

Categories having more than 10% differences are highlighted.

Categories	**Example**	**# of Words**	**Class S Avg.**	**Class D Avg.**	**Sig.**	**Diff (%)**
1 Work	Job, majors, xerox	327	226.7	215.2	0.0	5.1
2 Achievement	Earn, hero, win	186	266.6	252.2	0.0	5.4
3 Leisure	Cook, chat, movie	229	295.0	287.3	0.0	2.6
4 Home	Apartment, family	93	80.3	81.1	0.2	-1.0
5 ***Money***	***Audit, cash, owe***	173	***92.4***	***81.9***	***0.0***	***11.3***
6 ***Religion***	***Altar, church***	159	***107.3***	***89.0***	***0.0***	***17.0***
7 ***Death***	***Bury, coffin, kill***	62	***38.0***	***43.9***	***0.0***	***-15.6***

than 10% difference. Class S users have a lot more Money, Religion related words, while Class D users have a lot more Death related words.

We also did the sentiment analysis for each category of Personal Concerns for two classes of users. Figure 9.7 shows the sentiment for those categories for Class S and Class D users. We did not show Death category because it has naturally negative sentiment. We found Class S users have more positive sentiment tweets with religion words, while the average sentiment for Class D users is slightly negative. For all other categories, Class S users have more positive sentiment.

In conclusion, we explored the life satisfaction at the user level in Twitter. We present preliminary analysis of users who changed their mind and we found the number of this kind of users is small, and most of them only change their opinions once. The average change interval for them is about 80 days. In terms of differences between Class S and Class D users, we found the number of LS tweets posted for different classes of users are similar. The numbers of followers and followings are also similar. We did find some characteristics are significantly different for the two classes

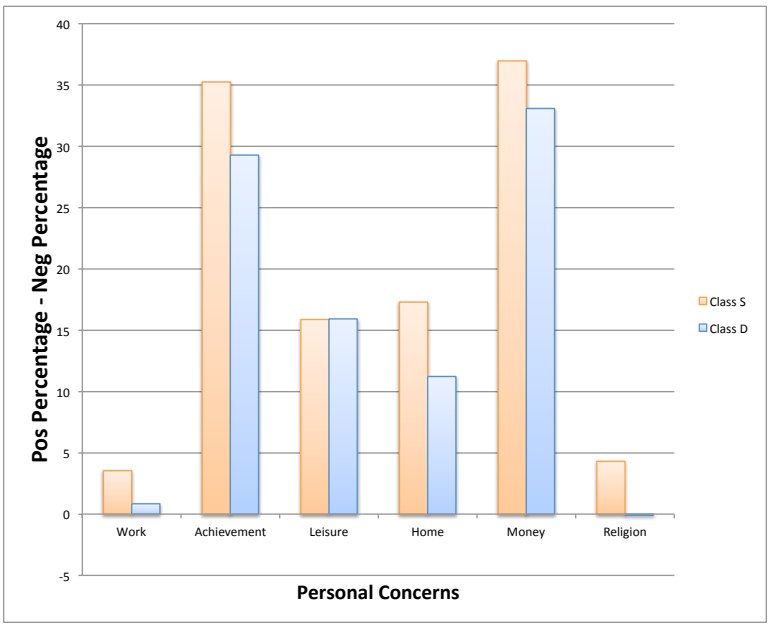


Figure 9.7. Comparison of sentiment for LIWC personal concerns

of users, such as number of URLs and number of hashtags. Using the LIWC lexicon, we are able to find that Class D users have more Swear words, Anger, Sadness, and Sexual words in their tweets. Class S users have more Money and Religion words in their tweets. Using sentiment analysis and the LIWC lexicon, we found users with different life satisfaction have different sentiment for topics like Health, Sexual, Religion, and Death. In the next chapter, we would like to further analyze the users who changed their life satisfaction.

CHAPTER 10

TEMPORAL ANALYSIS OF FACTORS ASSOCIATED WITH LIFE SATISFACTION EXPRESSIONS

In this chapter we are inspired by the work of Lewinsohn et al. [39]. In 1999, they analyzed the relationship between life satisfaction and psychosocial variables (PV). They conducted two surveys at different time points to gather life satisfaction, depression, and other psychosocial variables including social support, pleasant activities, cognition, stress, personality, and health for each participant. The length between two surveys is about eight months. Then four groups of users were selected: users with high/low LS ratings in both surveys, users with high ratings in the first and low ratings in the second, or the other way. Correlations of life satisfaction with the psychosocial variables were analyzed. For example, they found that people with low life satisfaction were more likely to have depression later.

We would like to do similar analysis using our Twitter data. We randomly selected four groups of users from our LSUser2YearsSelect dataset. We call the new dataset LSUser4Groups. LSUser2YearsSelect was used to analyze users who changed their life satisfaction (Section 9.2). So we also have all the tweets crawled for the users in LSUser4Groups dataset. In detail, we selected a similar number of users from the Group S, D, S⇒D, D⇒S by the following criteria: 1) users should have more than 10 tweets; 2) the change interval should be more than 30 days for S⇒D and D⇒S groups; 3) user accounts were created more than 300 days ago before they posted LS tweets for Group S and D. Finally, we have 1,278 users for each group

in the LSUser4Groups dataset. Note that the analysis in this chapter is temporal. This is a key difference from the analysis in the previous chapter even though there is overlap in the psychosocial variables analyzed.

To explore factors associated with life satisfaction, Lewinsohn et al. [39] explored Depression, Social Support and Social Interaction, Pleasant Activities, Cognitions (e.g. irrational beliefs, expectancies of positive and negative outcomes), Stress, Personalities, Health, and Demographic variables like Gender, Age, Marital status, etc. We are not able to infer users' demographic variables, personalities, and cognitions from text. Therefore, we keep Social Support (Friends and Family in LIWC), Pleasant Activities (Leisure in LIWC), and Health. Also we add more factors from LIWC categories like Anger, Anxiety, Death, Sadness, Home, Money, Religion, and Work. We manually created a lexicon for Depression (all words starting with "depress"), because depression is an important factor known to be associated with life satisfaction.

Furthermore, as in chapter 9 we included sentiment analysis for some categories. Such as Money (Pos) and Money (Neg) have the same lexicon, but the sentiment can be totally different in tweets. For some categories like Anger and Anxiety which obviously are negative we didn't use sentiment analysis. We again used the "pattern.en" module[1] for Python to identify "Positive," "Negative," and "Neutral" sentiment from tweets. We skip neutral tweets in our analysis. We show our final set of 19 PV categories in Table 10.1. Notice that the first five categories are naturally

[1]http://www.clips.ua.ac.be/pages/pattern-en

Table 10.1. Psychosocial variable categories

+: Wildcard matching (E.g. "depress*" can match "depress," "depressed," etc.)
SA: Sentiment Analysis

PV Category	**Lexicon Example**	**# Lexicon**	**Use SA?**
Anger	Hate, kill, annoyed	185+	No
Anxiety	Worried, fearful, nervous	91+	No
Death	Bury, coffin, kill	63+	No
Depression	Depress, depression	2+	No
Sadness	Crying, grief, sad	101+	No
Health (Pos)	Health, healthy	234+	Yes
Health (Neg)	Clinic, flu, pill	234+	Yes
Home (Pos)	Apartment, kitchen, family	93+	Yes
Home (Neg)	Apartment, kitchen, family	93+	Yes
Leisure (Pos)	Cook, chat, movie	228+	Yes
Leisure (Neg)	Cook, chat, movie	228+	Yes
Money (Pos)	Audit, cash, owe	173+	Yes
Money (Neg)	Audit, cash, owe	173+	Yes
Religion (Pos)	Altar, church, mosque	160+	Yes
Religion (Neg)	Altar, church, mosque	160+	Yes
Social Support (Pos)	Daughter, husband, friend	102+	Yes
Social Support (Neg)	Daughter, husband, friend	102+	Yes
Work (Pos)	Job, majors, xerox	326+	Yes
Work (Neg)	Job, majors, xerox	326+	Yes

negative.

To analyze the relationship of different factors and life satisfaction from tweets, we first define "PV tweets" as tweets that belong to at least one of our 19 PV categories. Next we define "Day 0." Since a Group S (or Group D) users can tweet multiple Class S (or Class D tweets), we select the date when they posted their first LS tweets on "Day 0". For users who changed their life satisfaction (Group S⇒D and D⇒S), "Day 0" is the date when they posted the LS tweet which contradicts their previous LS expression. i.e., for Group S⇒D users, "Day 0" is the date when they

Table 10.2. Number of days before and after "Day 0"
for different percentages of active users

Group	**Before "Day 0"**			**At "Day 0"**	**After "Day 0"**		
	20%	**50%**	**75%**	**100%**	**75%**	**50%**	**20%**
S	-270	-80	-20	1,278	90	150	420
D	-185	-60	-20	1,278	70	190	340
S⇒D	-375	-175	-115	1,278	50	115	285
D⇒S	-375	-200	-115	1,278	50	115	285

posted their first Class D tweets (they posted Class S tweets before); and for Group D⇒S users, "Day 0" is the date when they posted their first Class S tweets (they posted Class D tweets before).

Then we would like to answer the following questions for each PV category. How many PV tweets did users post before "Day 0"? How many PV tweets did users post after "Day 0"? Are there differences between number of PV tweets posted before and after "Day 0"? If there are differences, how to explain them?

It is important to consider the active time span in our analysis as this may differ across users. Thus for each user, we found their first and last tweets. Then we define their active period as the time between their first and last tweets. We show the percentage of active users on different days before and after "Day 0" in Figure 10.1. Table 10.2 highlights some data points, for example, 50% of the Class S users are active at 80 days before "Day 0" and at 150 days after "Day 0."

We first show the percentage of positive, negative, and total PV tweets in the set of tweets for Groups S & D users and Groups S⇒D & D⇒S in Figure 10.2a and 10.2b respectively. The Y axis shows the percentage of PV tweets accumulated from

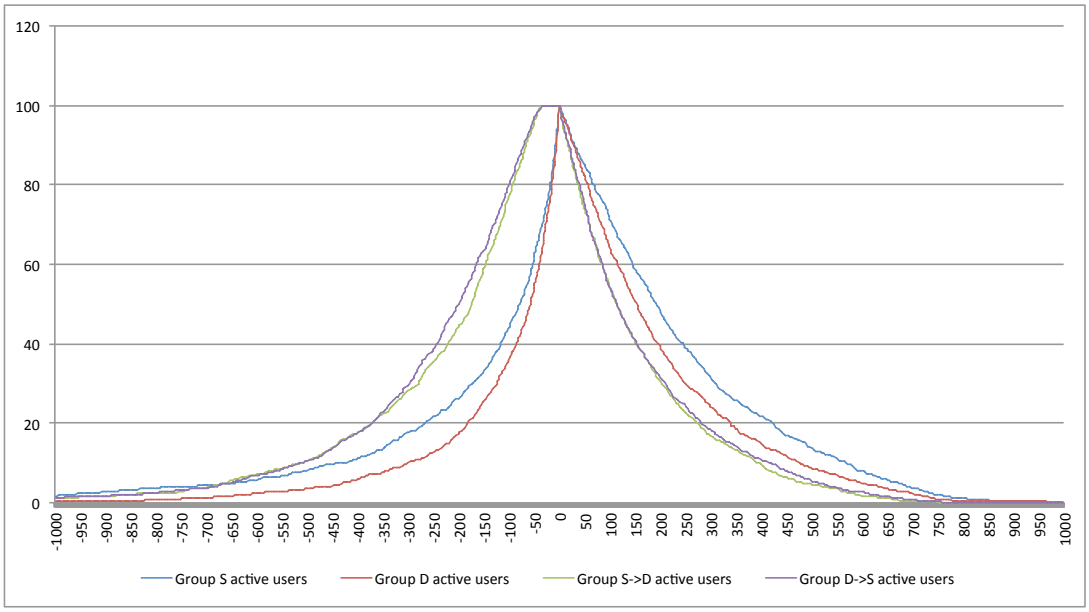
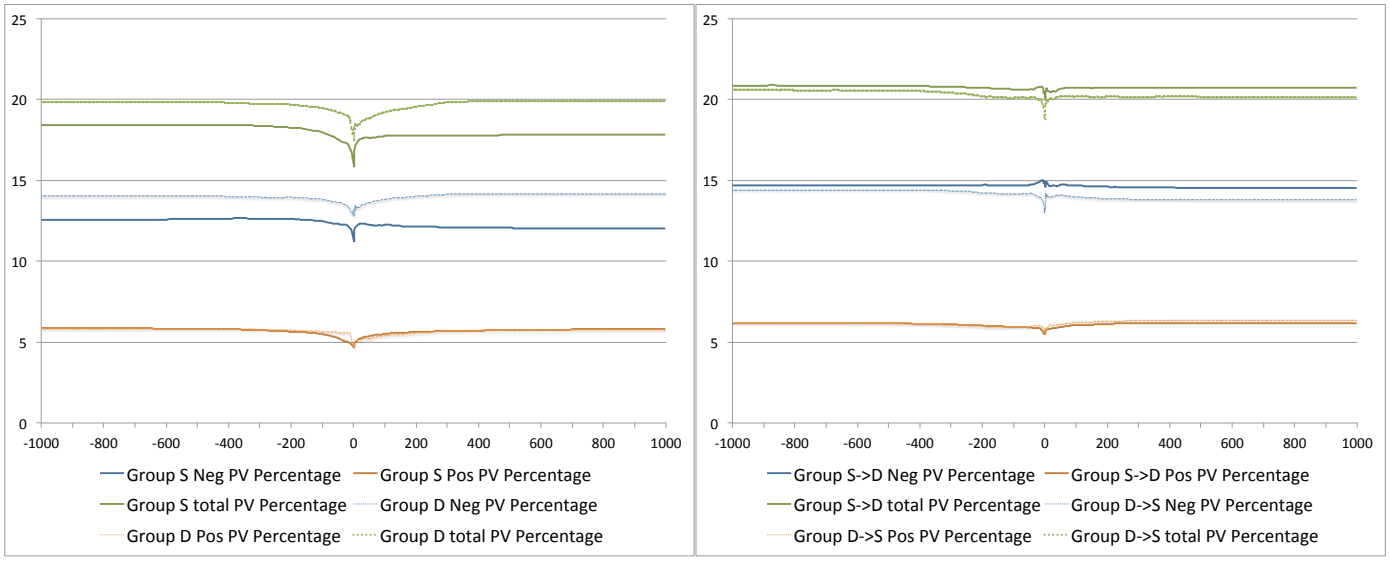


Figure 10.1. Percentage of active users at different days

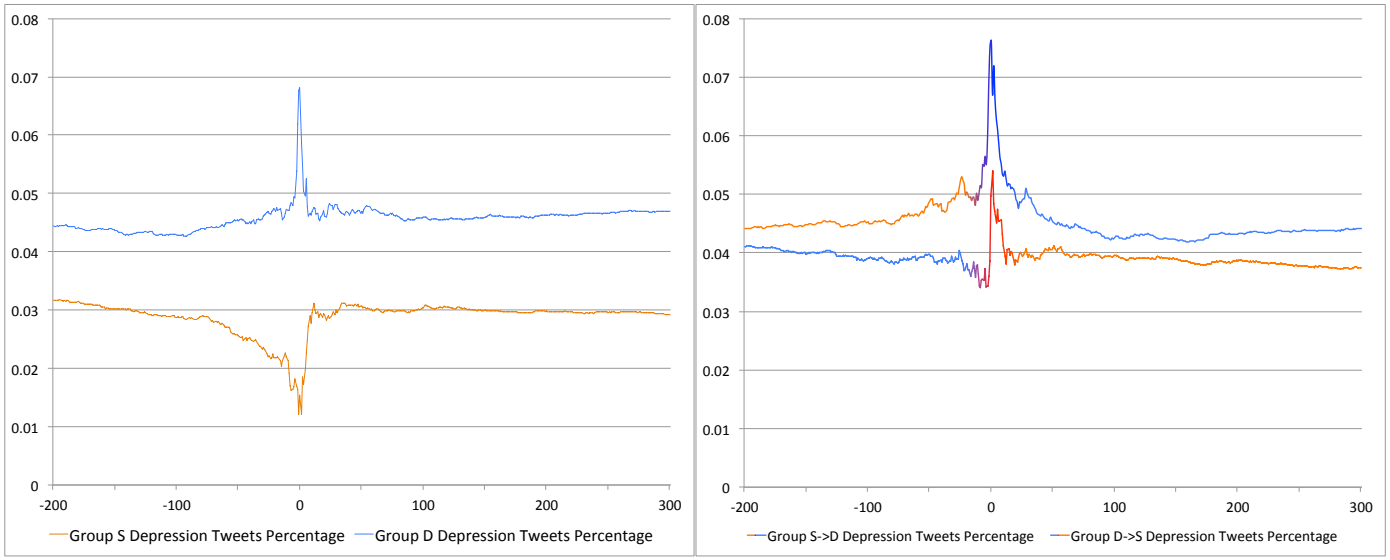
Day 0. For example, the number at Day 100 is the percentage of all PV tweets by that class of users from Day 0 to Day 100 out of all of their tweets from Day 0 to Day 100. The number at Day -100 is the percentage of all PV tweets from Day -100 to Day 0 out of all of their tweets by that class of users from Day -100 to Day 0. First it is interesting to see that around 20% of all tweets for each group are PV tweets. We also see there are more negative PV tweets. The reason could be that we have more negative PV categories. We also find that there is always a dip around "Day 0," which means that users post fewer PV tweets around Day 0. The dips are not an effect of the cumulation process. We checked this by accumulating percentages with other days as the focus (e.g. day -100). We still see the dips around "Day 0." In addition, significance tests show that Group S users have significantly fewer negative PV tweets than Group D users throughout the timeline. This result makes

intuitive sense. Also Group S⇒D users have significantly more negative PV tweets than Group D⇒S users throughout the timeline. Next, we show the figures for all the PV categories, then we analysis the results in the following three sections. Notice we adjusted the range of Y-axis to make them the same within each category.



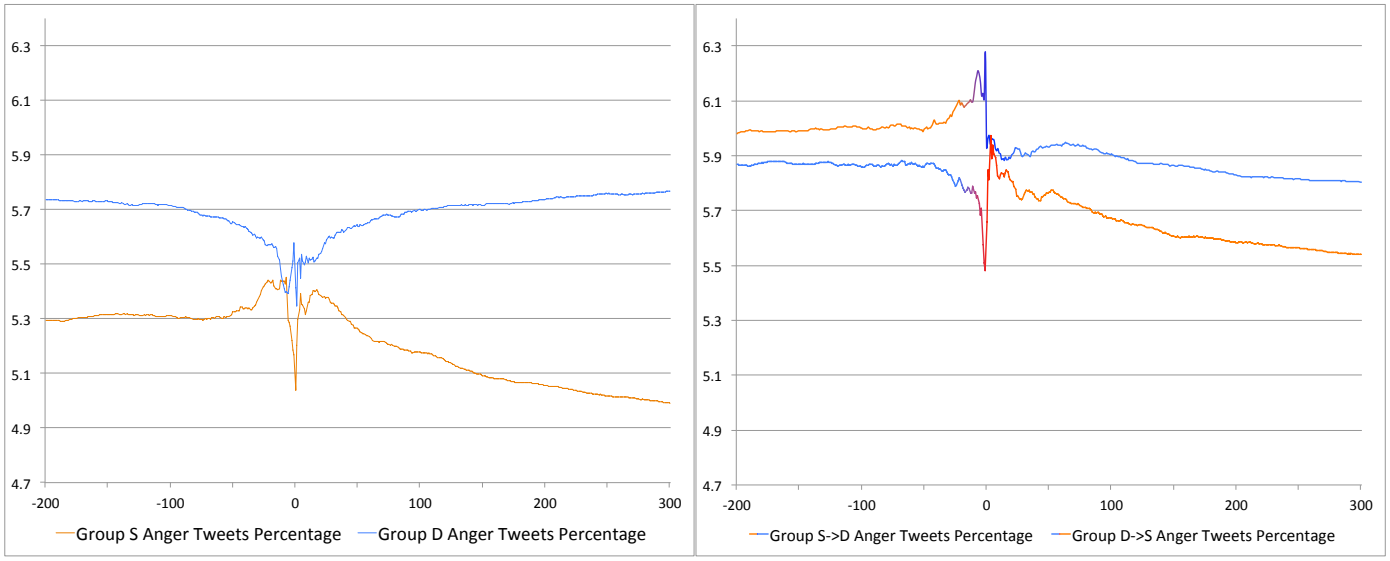
(a) Group S and D (b) Group S⇒D and D⇒S

Figure 10.2. Percentage of postive, negative, and total PV tweets



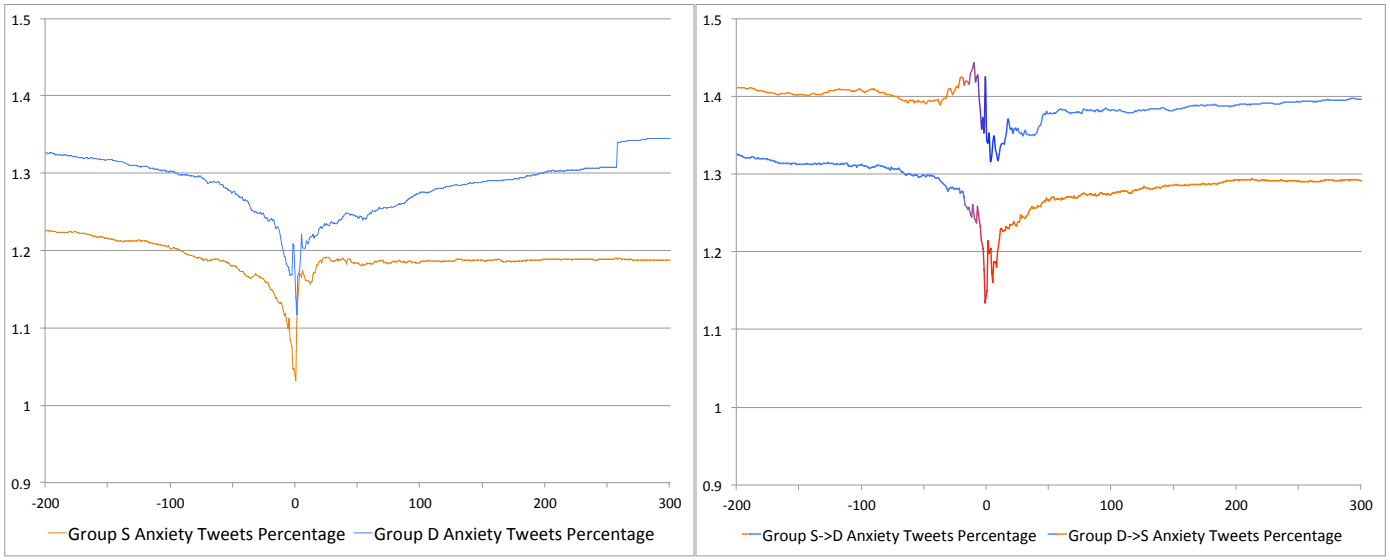
(a) Group S and D (b) Group S⇒D and D⇒S

Figure 10.3. Percentage of depression tweets



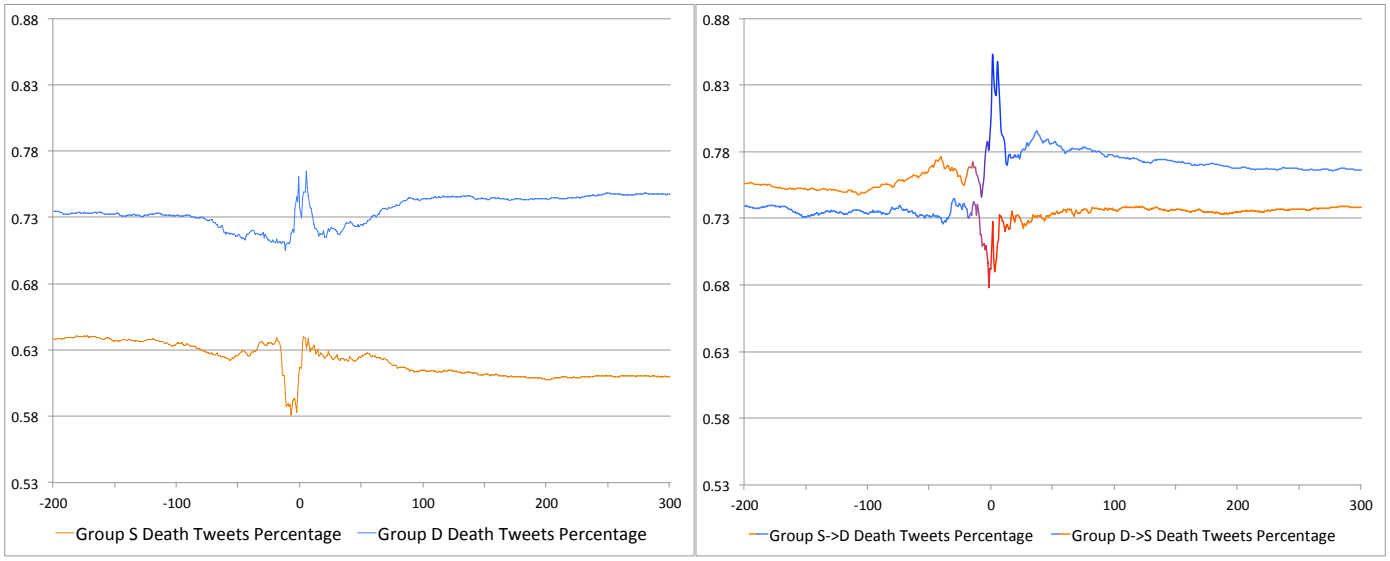
(a) Group S and D (b) Group S⇒D and D⇒S

Figure 10.4. Percentage of anger tweets



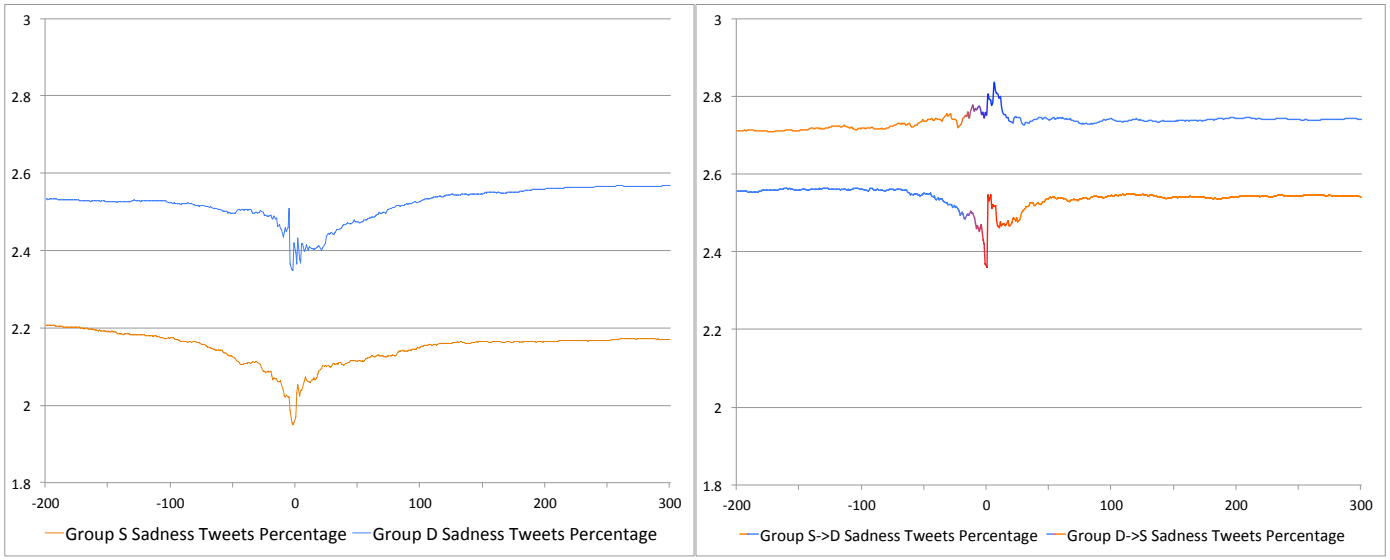
(a) Group S and D (b) Group S⇒D and D⇒S

Figure 10.5. Percentage of anxiety tweets



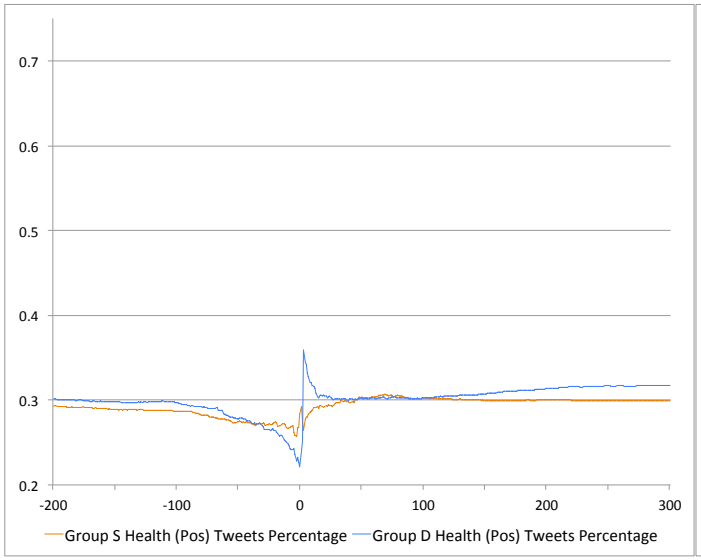
(a) Group S and D (b) Group S⇒D and D⇒S

Figure 10.6. Percentage of death tweets

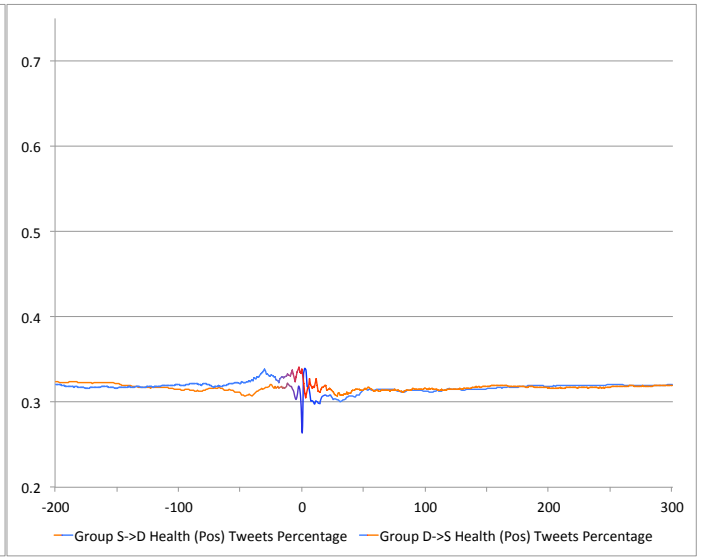


(a) Group S and D (b) Group S⇒D and D⇒S

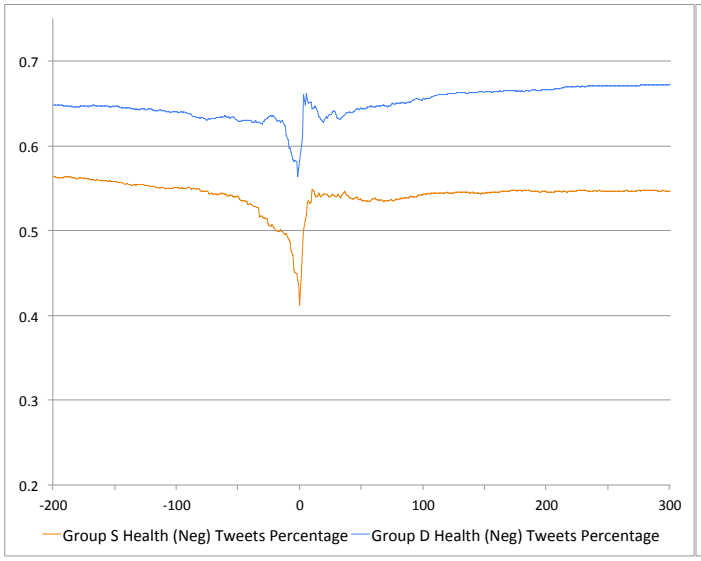
Figure 10.7. Percentage of sadness tweets



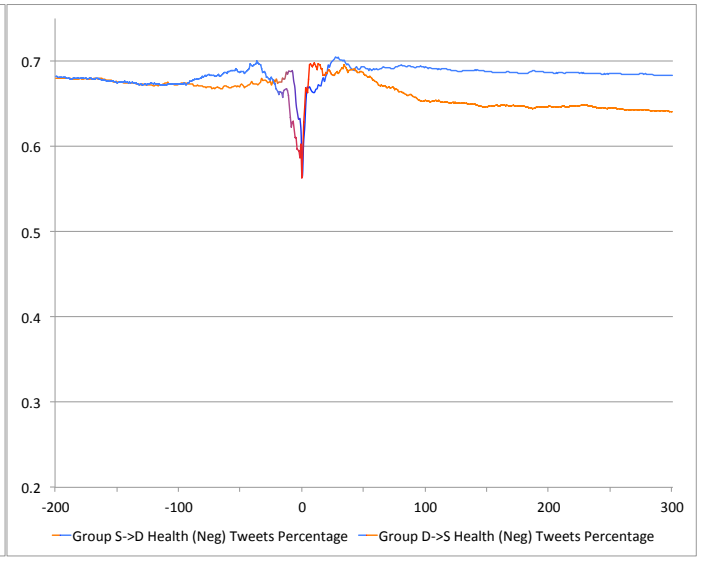
(a) Postive: Group S and D



(b) Postive: Group S⇒D and D⇒S



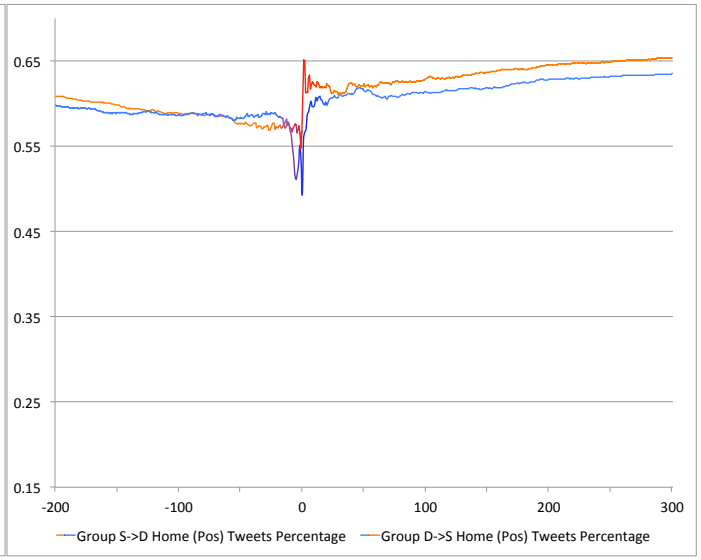
(c) Negative: Group S and D



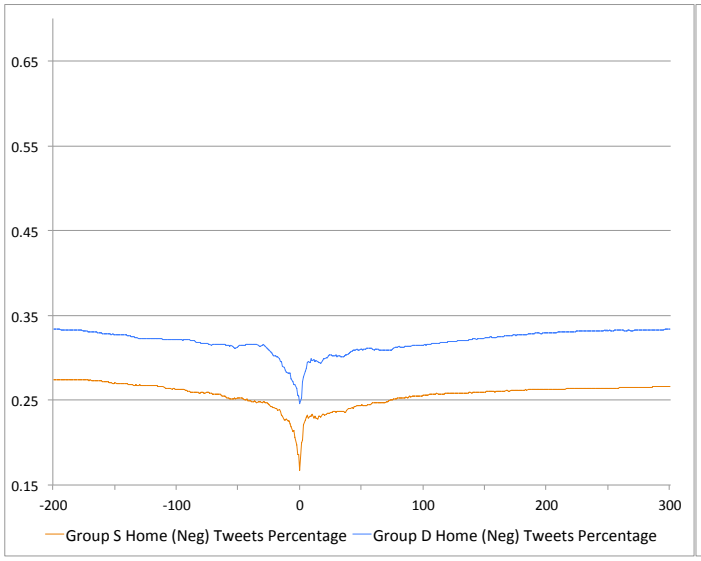
(d) Negative: Group S⇒D and D⇒S

Figure 10.8. Percentage of health tweets

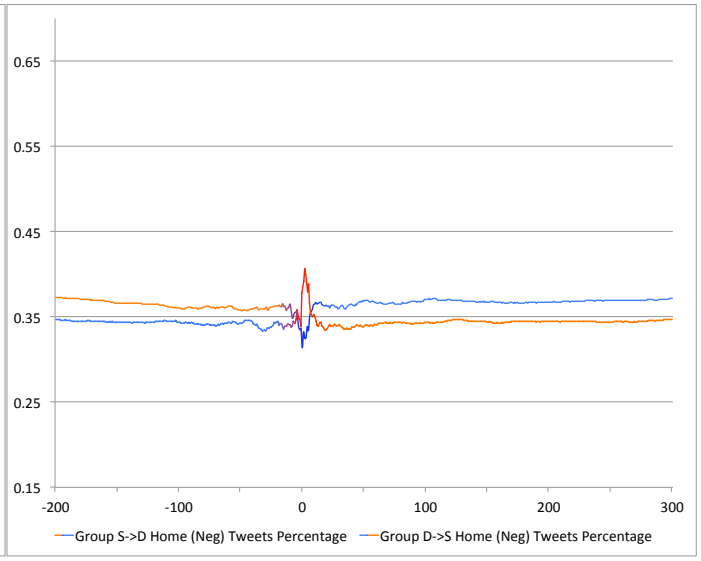
(a) Positive: Group S and D



(b) Positive: Group S⇒D and D⇒S

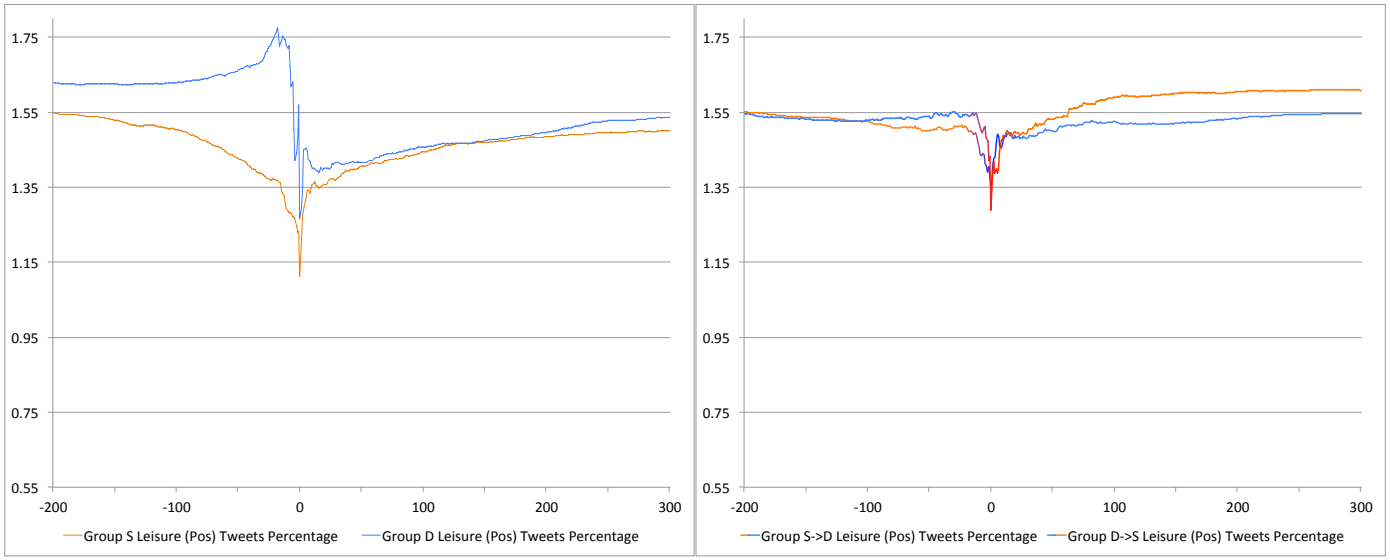


(c) Negative: Group S and D



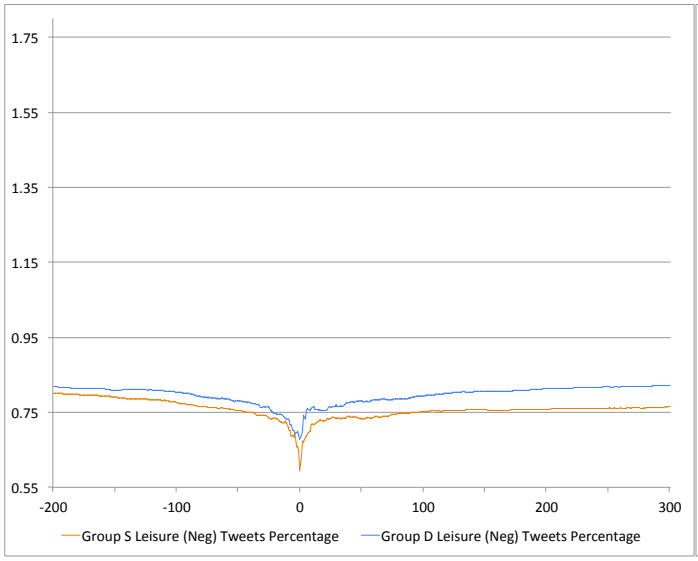
(d) Negative: Group S⇒D and D⇒S

Figure 10.9. Percentage of home tweets

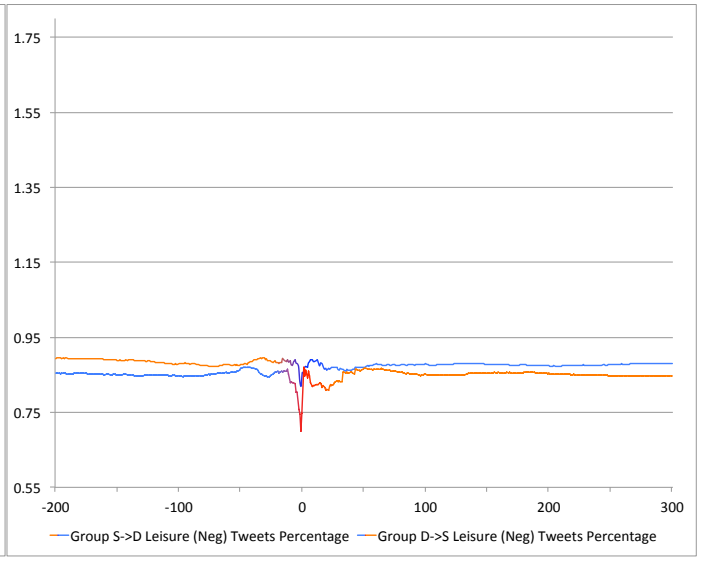


(a) Positive: Group S and D

(b) Positive: Group S⇒D and D⇒S

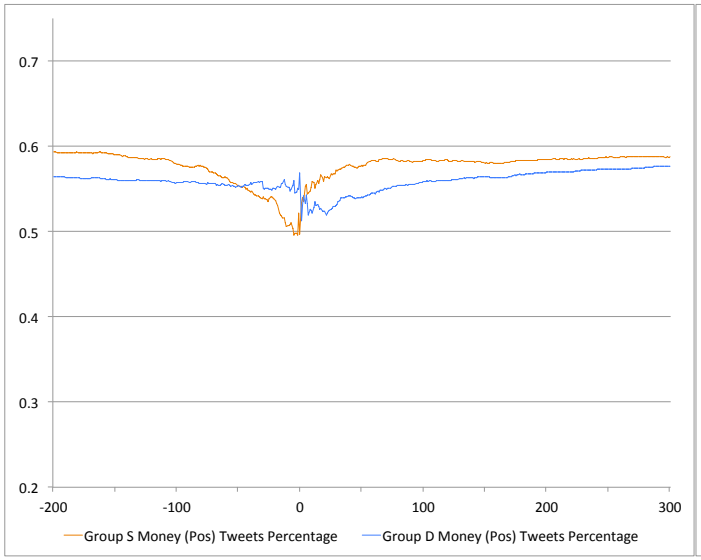


(c) Negative: Group S and D

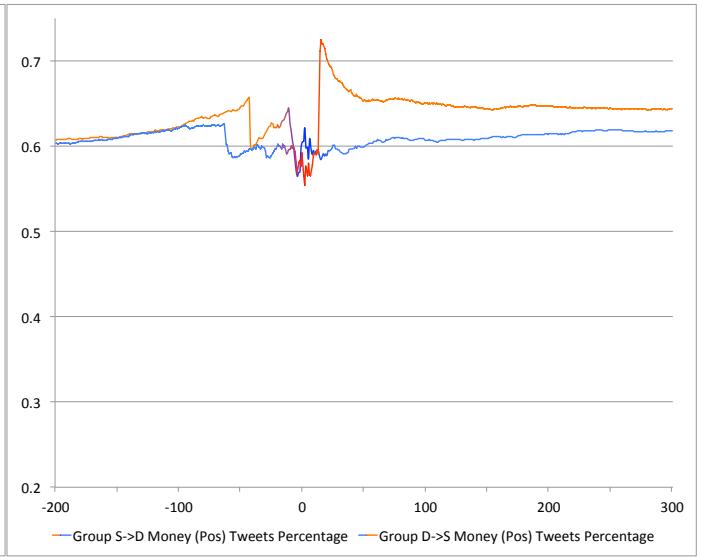


(d) Negative: Group S⇒D and D⇒S

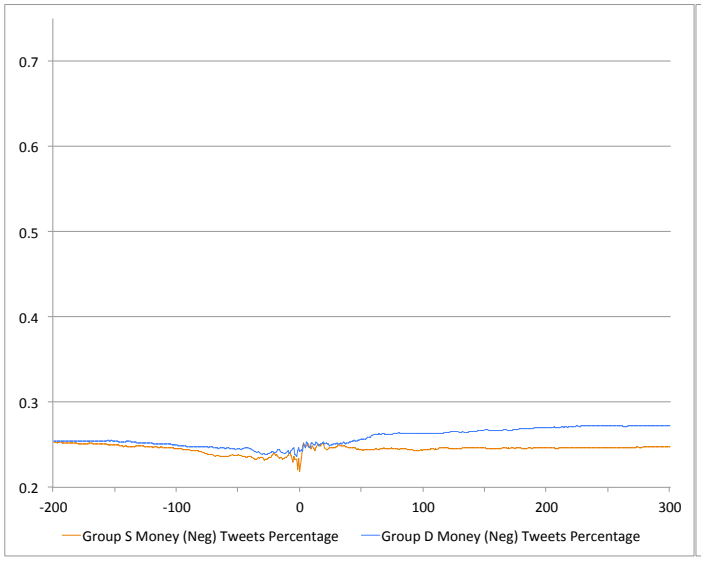
Figure 10.10. Percentage of leisure tweets



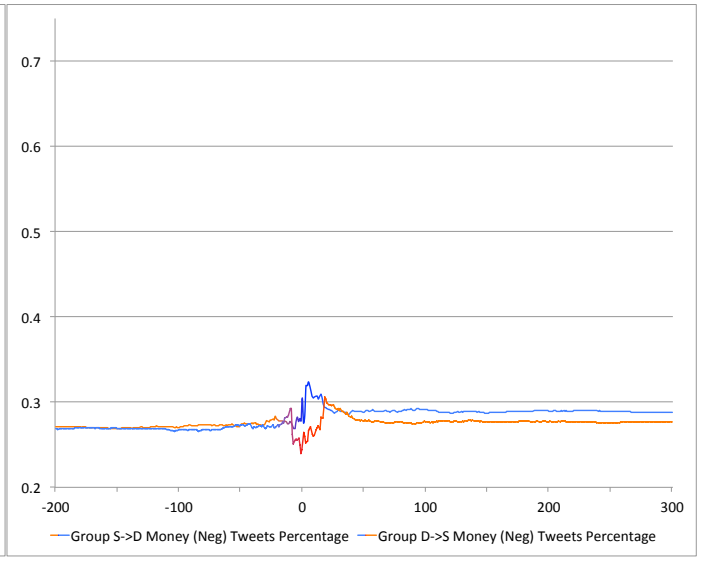
(a) Positive: Group S and D



(b) Positive: Group S⇒D and D⇒S

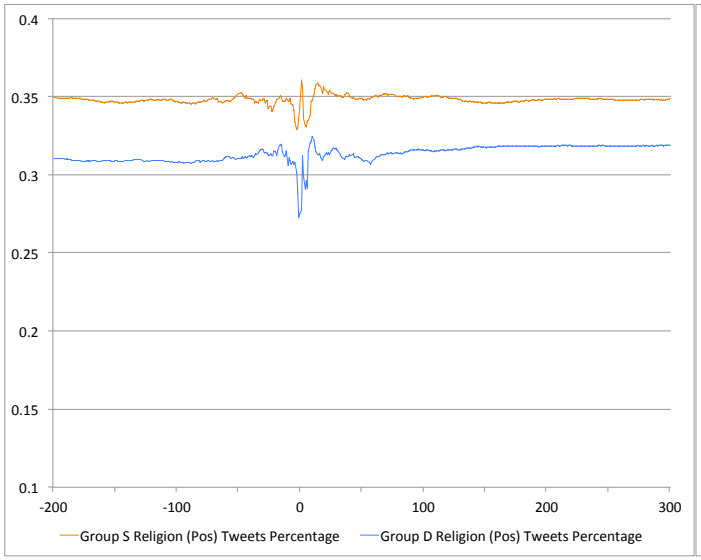


(c) Negative: Group S and D

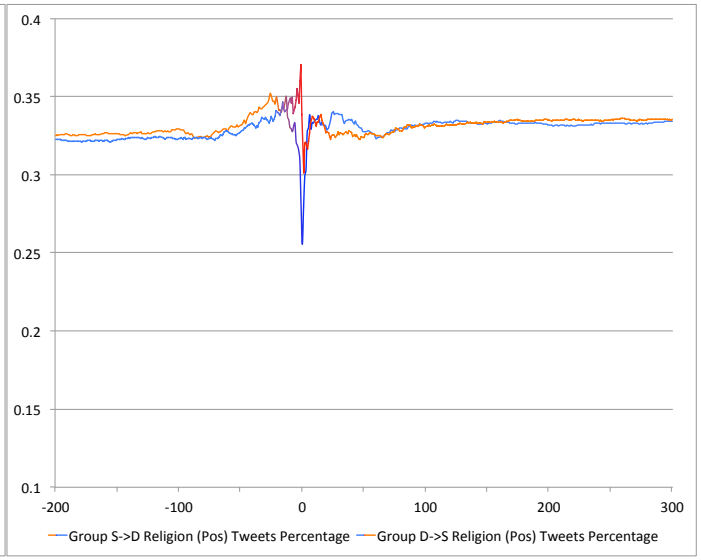


(d) Negative: Group S⇒D and D⇒S

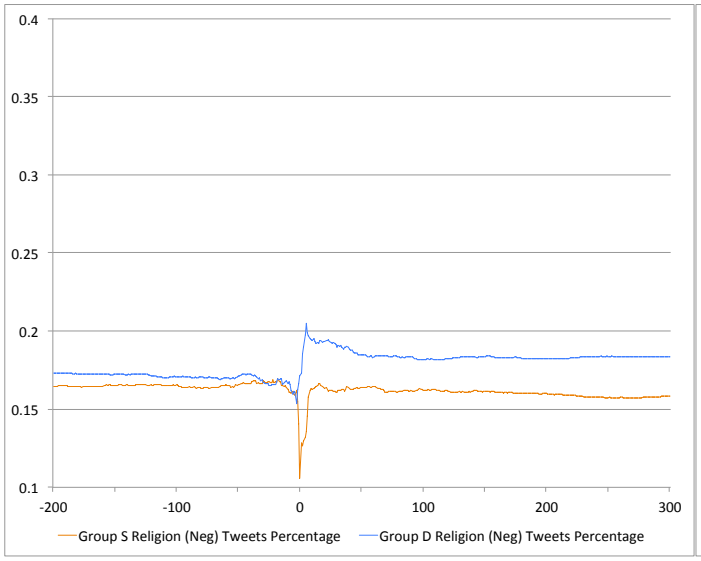
Figure 10.11. Percentage of money tweets



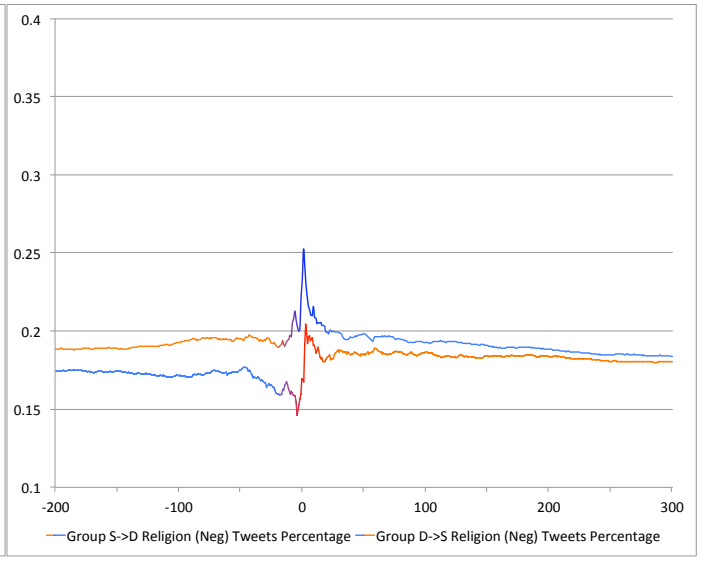
(a) Positive: Group S and D



(b) Positive: Group S⇒D and D⇒S

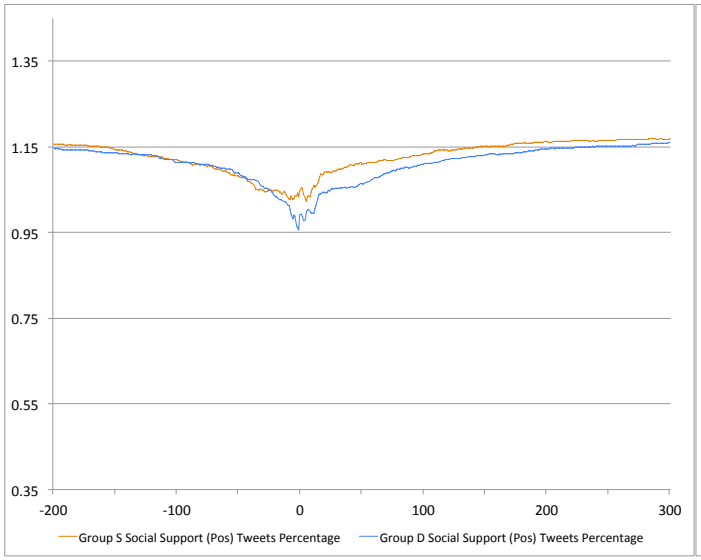


(c) Negative: Group S and D

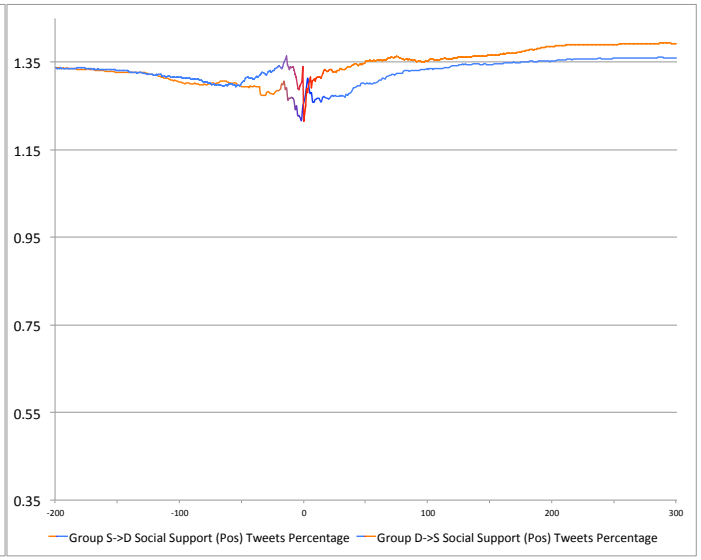


(d) Negative: Group S⇒D and D⇒S

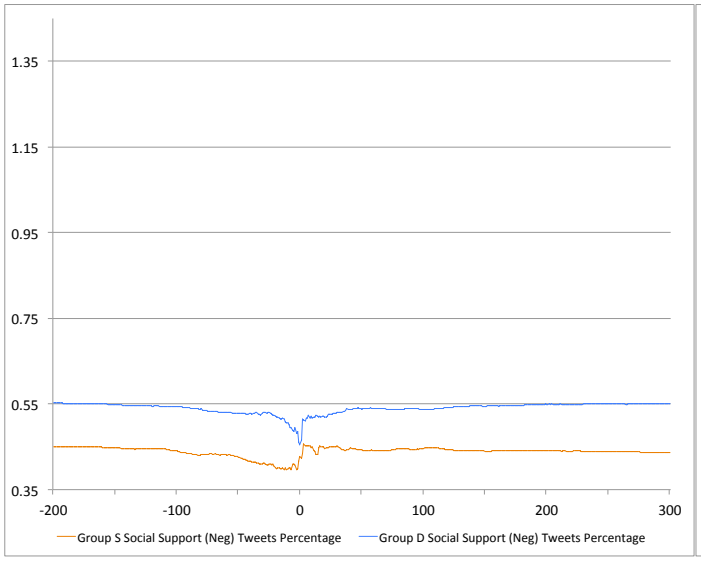
Figure 10.12. Percentage of religion tweets



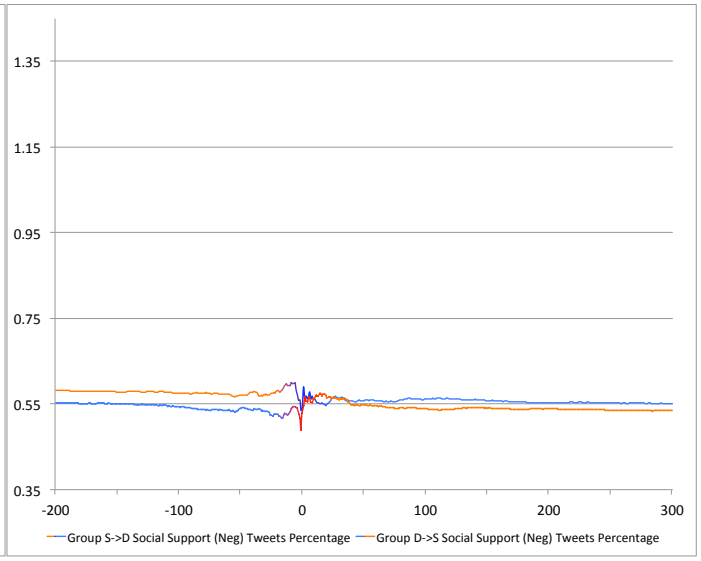
(a) Positive: Group S and D



(b) Positive: Group S⇒D and D⇒S

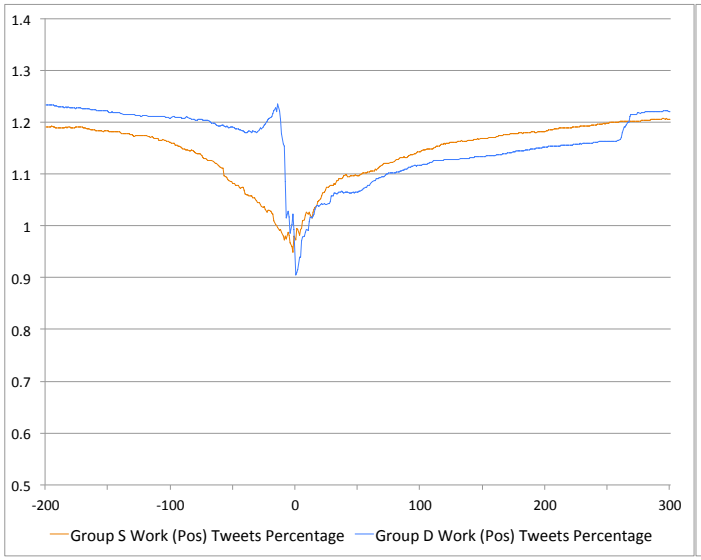


(c) Negative: Group S and D

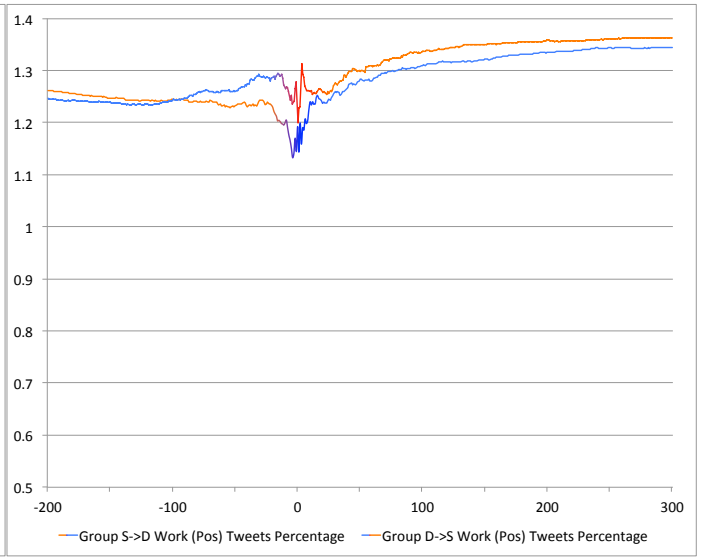


(d) Negative: Group S⇒D and D⇒S

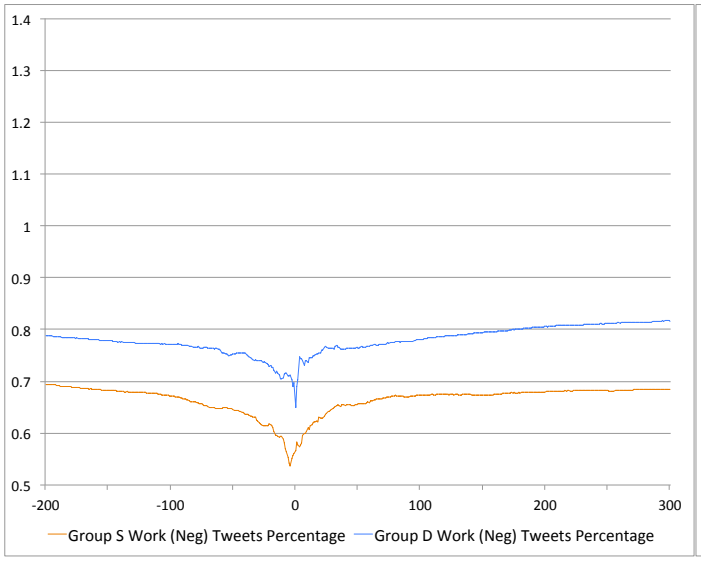
Figure 10.13. Percentage of social support tweets



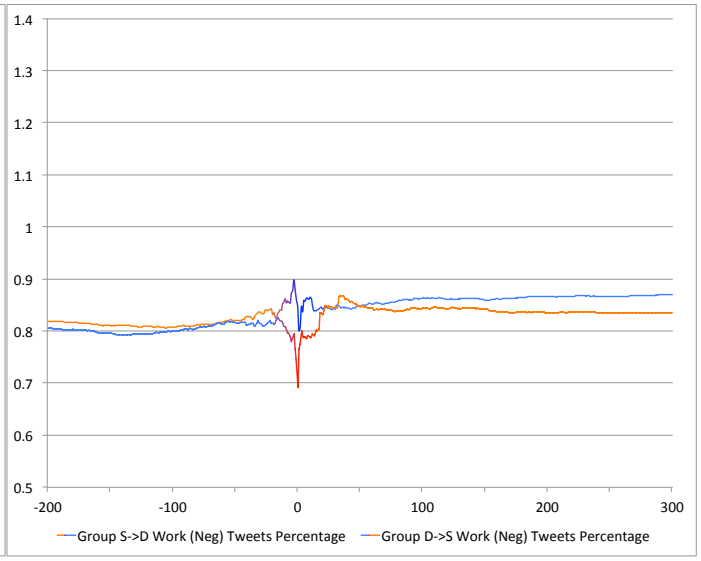
(a) Positive: Group S and D



(b) Positive: Group S⇒D and D⇒S



(c) Negative: Group S and D



(d) Negative: Group S⇒D and D⇒S

Figure 10.14. Percentage of work tweets

In the following three sections, we will analysis the results for all the categories in the following ways. We first show the factors associated with posting the LS tweets (using Group S & D users). Then we show the factors associated with the change of users' life satisfaction (using Group S⇒D & D⇒S users). Finally, we compare the results from the four groups.

10.1 Group S vs. Group D

For all the figures comparing Group S and Group D, the orange lines represent Group S users and the blue lines represent Group D users. We summarize the patterns for Group S and D.

- Pattern 1: We found Anger, Anxiety, Sadness, Health (Neg), Home (Neg), Leisure (Neg), Social Support (Neg), and Work (Neg) have a similar pattern (Figure 10.4a, 10.5a, 10.7a, 10.8c, 10.9c, 10.10c, 10.13c, and 10.14c respectively) with a dip around "Day 0" for both groups. Importantly, Group D has significantly more PV tweets in these negative categories than Group S. (There is slight overlap around "Day 0" for Anxiety and Anger.) These are interesting findings which follow our intuition. We also found Group D has more Health (Pos) and Leisure (Pos) tweets than Group S throughout timeline. (Figure 10.8a, 10.10a)

- Pattern 2: Death and Depression follow another pattern (Figure 10.6a, 10.3a). Although Group D has significantly more death and depression tweets than Group S throughout the timeline, Group S has a dip around "Day 0," while

Group D has a peak. The results are consistent with previous researchers [30] who also found the negative correlation between depression and life satisfaction.

- Pattern 3: For Money (Neg), Religion (Neg) (Figure 10.11c, 10.12c respectively), Group D has significantly more PV tweets after "Day 0." Before "Day 0," even thought the two groups are still significantly different, the gap is much smaller.

- Pattern 4: For positive categories, we found for Home (Pos) and Work (Pos) (Figure 10.9a, 10.14a), Group S has significantly fewer PV tweets before "Day 0," but has significantly more PV tweets after "Day 0."

- Pattern 5: For Religion (Pos) (Figure 10.13a, 10.12a), Group S has significantly more PV tweets throughout the timeline. This is also suggested in the literature from Chaeyoon and Robert [40]. Money (Pos) and Social Support (Pos) also follow this pattern, even they have slightly overlap.

We summarized the comparing of Group S and D for the patterns before and after "Day 0" in Table 10.3. We also compared the PV tweets posted before and after "Day 0" for Group S and D separately in Table 10.4. We see that there is less than 5% difference for most of the categories. For Group S, we found that they have more Anger, Depression tweets before "Day 0." The results are consistent with our intuition.The only exception is Home (Pos) for Group S. For Group D, we found that they have more Depression, Health (Pos), Money (Neg), Religion (Neg) after "Day 0," and have fewer Leisure (Pos) tweets after "Day 0." The results also follow our

Table 10.3. Comparison of PV categories for Groups S and D

The patterns refer to Section 10.1
">": S has significantly more PV tweets than D.
"<": S has significantly fewer PV tweets than D.

PV Category	**Before "Day 0" S vs. D**	**After "Day 0" S vs. D**	**Pattern**
Anger	<	<	Pattern 1
Anxiety	<	<	Pattern 1
Death	<	<	Pattern 2
Depression	<	<	Pattern 2
Sadness	<	<	Pattern 1
Health (Pos)	<	<	Pattern 1
Health (Neg)	<	<	Pattern 1
Home (Pos)	<	>	Pattern 4
Home (Neg)	<	<	Pattern 1
Leisure (Pos)	<	<	Pattern 1
Leisure (Neg)	<	<	Pattern 1
Money (Pos)	>	>	Pattern 5
Money (Neg)	<	<	Pattern 3
Religion (Pos)	>	>	Pattern 5
Religion (Neg)	<	<	Pattern 3
Social Support (Pos)	>	>	Pattern 5
Social Support (Neg)	<	<	Pattern 1
Work (Pos)	<	>	Pattern 4
Work (Neg)	<	<	Pattern 1

intuition.

10.2 Group S⇒D vs. Group D⇒S

In the last section, we analyzed the relationship of LS and PV categories for Group S and Group D users. More, we compare Groups S⇒D & D⇒S users in this section. To represent the change of LS for Group S⇒D users, the line starts with orange, ends with blue. For Group D⇒S users, the line starts with blue, and ends with orange.

From the figures, we find that the comparisons between the two groups, S⇒D

Table 10.4. Comparison of PV categories before and after "Day 0"

">": Total percentage before is more than 5% larger than after "Day 0."
"<": Total percentage after is more than 5% larger than before "Day 0."
"=": Total percentage before and after "Day 0" is less than 5% difference.

PV Category	**Group S Before vs. After**	**Group D Before vs. After**
Anger	>	=
Anxiety	=	=
Death	=	=
Depression	>	<
Sadness	=	=
Health (Pos)	=	<
Health (Neg)	=	=
Home (Pos)	<	=
Home (Neg)	=	=
Leisure (Pos)	=	>
Leisure (Neg)	=	=
Money (Pos)	=	=
Money (Neg)	=	<
Religion (Pos)	=	=
Religion (Neg)	=	<
Social Support (Pos)	=	=
Social Support (Neg)	=	=
Work (Pos)	=	=
Work (Neg)	=	=

and D⇒S fall into 4 patterns.

- Pattern 1 is the most uniform. In this pattern, we find that S⇒D has significantly more PV tweets than D⇒S for most of the timeline. In details, S⇒D has significantly more Depression and Religion (Neg) tweets (Figure 10.3b and 10.12d) throughout the timeline with peaks for both two groups. For Anxiety, Leisure (Neg), and Work (Neg) tweets (Figure 10.5b, 10.10d, and 10.14d), both groups have dips around "Day 0" (Leisure (Neg), and Work (Neg) has overlap around "Day 0"). For Anger, Death and Sadness (Figure 10.4b, 10.6b and 10.7b), S⇒D peaks around "Day 0", while D⇒S dips. For Home (Neg) (Figure 10.9d), S⇒D dips around "Day 0", while D⇒S peaks. In addition, for Social Support (Neg) (Figure 10.13d), both groups have overlaps around "Day 0".

- Pattern 2: similar in the beginning but significantly different after "Day 0." The two groups are similar in the beginning, but S⇒D has significantly more Health (Neg) and Money (Neg) tweets (Figure 10.8d and 10.11d) and significantly less Leisure (Pos), Money (Pos), and Social Support (Pos) tweets (Figure 10.10b, 10.11b, and 10.13b) than D⇒S (after "Day 0").

- Pattern 3: difference before and around "Day 0," similar after. S⇒D has significantly more Religion (Pos) (Figure 10.12b) than D⇒S before "Day 0," then the two groups have similar number of Religion (Pos) to the end. For Health (Pos) (Figure 10.8b), two groups also have similar number of tweets

after "Day 0."

- Pattern 4: difference before and after, group switched. S⇒D has significantly more Home (Pos) and Work (Pos) tweets (Figure 10.9b and 10.14b) than D⇒S at beginning, but after "Day 0," D⇒S has significantly more of those tweets.

We summarize the patterns in Table 10.5. We also compared the PV tweets posted before and after "Day 0" for Group S⇒D and D⇒S in Table 10.6. For negative categories, we found Group S⇒D has fewer Money (Neg) and Work (Neg), and more Social Support (Neg) before "Day 0." Group D⇒S has more Anger, Depression, Health (Neg) tweets before "Day 0." For positive categories, we found Group S⇒D has fewer Work (Pos) before "Day 0." Group D⇒S has fewer Home (Pos), Money (Pos), and Work (Pos) tweets before "Day 0." Most of the results make intuitive sense.

10.3 Comparison Of Four Groups

In the previous two sections, we compared Group S vs. Group D, and Group S⇒D vs. Group D⇒S. In this section, we would like to see what can we find when comparing the 4 groups. Therefore, we summarized the results for comparing the four groups.

- Comparing summary trends (Figure 10.2a and 10.2b)

  We mentioned before, it is interesting that there are 18 to 21% of tweets are about psychosocial variables (PV). The four groups are similar in the prevalence of positive PVs (about 6%). In negative aspects Group S posts fewer tweets

Table 10.5. Comparison of PV categories for Groups S⇒D and D⇒S

The patterns refer to Section 10.2
">": S⇒D has significantly more PV tweets than D⇒S.
"<": S⇒D has significantly fewer PV tweets than D⇒S.
"=": Groups has similar number of PV tweet in the figure
(even it maybe significant in statistic test).

PV Category	**Before "Day 0" S⇒D vs. D⇒S**	**After "Day 0" S⇒D vs. D⇒S**	**Pattern**
Anger	>	>	Pattern 1
Anxiety	>	>	Pattern 1
Death	>	>	Pattern 1
Depression	>	>	Pattern 1
Sadness	>	>	Pattern 1
Health (Pos)	>	<	Pattern 3
Health (Neg)	=	>	Pattern 2
Home (Pos)	>	<	Pattern 4
Home (Neg)	>	>	Pattern 1
Leisure (Pos)	>	>	Pattern 2
Leisure (Neg)	>	>	Pattern 1
Money (Pos)	=	<	Pattern 2
Money (Neg)	=	>	Pattern 2
Religion (Pos)	>	=	Pattern 3
Religion (Neg)	>	>	Pattern 1
Social Support (Pos)	=	<	Pattern 2
Social Support (Neg)	>	>	Pattern 1
Work (Pos)	>	<	Pattern 4
Work (Neg)	>	>	Pattern 1

Table 10.6. Comparison of PV categories before and after "Day 0"

">": Total percentage before is more than 5% larger than after "Day 0."
"<": Total percentage after is more than 5% larger than before "Day 0."
"=": Total percentage before and after "Day 0" is less than 5% difference.

PV Category	**Group S⇒D Before vs. After**	**Group D⇒S Before vs. After**
Anger	=	>
Anxiety	=	=
Death	=	=
Depression	=	>
Sadness	=	=
Health (Pos)	=	=
Health (Neg)	=	>
Home (Pos)	=	<
Home (Neg)	=	=
Leisure (Pos)	=	=
Leisure (Neg)	=	=
Money (Pos)	=	<
Money (Neg)	<	=
Religion (Pos)	=	=
Religion (Neg)	=	=
Social Support (Pos)	=	=
Social Support (Neg)	>	=
Work (Pos)	<	<
Work (Neg)	<	=

(about 12.5%) compared to the other three groups (about 14% to 15%), which appear similar to each other in this regard.

- Anger, anxiety, death, depression and sadness

  In these 5 PV categories that are naturally negative, Group S has significantly fewer tweets compared to the other three groups. For Group D we do not find consistent difference with the groups that changed their life satisfaction. The two exceptions are anger and sadness where Group D has fewer posts than S⇒D.

- Negative aspects determined with sentiment analysis (Health (Neg), Home (Neg), Leisure (Neg), Money (Neg), Religion (Neg), Social Support (Neg), and Work (Neg))

  We found Group S also has significantly fewer tweets than all other 3 groups. For Group D we do not find consistent difference with users who changed their life satisfaction. The exceptions are for Home (Neg) and Leisure (Neg) where Group D has fewer tweets.

- Positive aspects determined with sentiment analysis (Health (Pos), Home (Pos), Leisure (Pos), Money (Pos), Religion (Pos), Social Support (Pos), and Work (Pos))

  In Home (pos), Social Support (pos), and Work (pos) , both Group S and D each have fewer tweets than the groups that change status. In Money (Pos) , only Group D has fewer tweets than both groups that change status. For Health

(Pos) and Leisure (Pos) the trends are mixed. The one graph which stands out in this mix is that Group S has far more tweets than any of the other 3 groups for Religion (Pos).

- Positive vs. Negative aspects determined with sentiment analysis (Health, Home, Leisure, Money, Religion, Social Support, and Work)

  We found that users in the four groups always have more positive tweets than negative tweets, except for Health category.

In summary, Group S stands out in that it has fewer tweets than the other three groups in the clearly negative aspects (anger, anxiety, death, depression, sadness) and also in all negative aspects determined through sentiment analysis (home, social support etc.). This is a clearly interesting result because it is consistent with intuition. Group S also stands out in having more religion (pos) tweets than the other 3 groups. Group D looks more similar to the groups that change status with a few unremarkable exceptions.

Comparing Group S and S⇒D, we found S⇒D has significantly more negative PV tweets on every category. This is an important observation that may be useful to identify individuals at risk of becoming dissatisfied with their lives, even thought they express satisfaction at some point. Identifying such individuals opens up possibilities for positive interventions. This is a direction for future research.

In conclusion, we explored the relationship between 19 psychosocial variable categories and life satisfaction on social media. From the analyses of Group S and Group D users, we summarized five patterns. Group D users have more PV tweets for

most of the negative PV categories throughout timeline. They posted more death and depression tweets around the time they posted Class D tweets. Group S users have more positive money, social support, especially religion tweets throughout timeline.

For the analyses of Group S⇒D and Group D⇒S users, we also summarized five patterns. In general, S⇒D has more tweets for most of the negative PV categories. For positive health, home, and work categories, we found S⇒D has more those tweets before change status, and fewer those tweets after change status.

We also compared the four groups, and found that in general Group S has significantly fewer tweets in all the negative PV categories than other three groups. Users posted more positive tweets than negative for the PV categories with sentiment analysis expect health. We found it is possible to detect the Class S users who may at risk of becoming dissatisfied.

CHAPTER 11

ANALYSIS OF LS USERS BY LOCATION

Our dataset allows us to average the geographical distribution of LS users. This may be useful because as governments are increasingly focusing on the well-being of their population as an indication of success in policies of strategies. In this chapter we present the geographical distribution of LS users using LSUsers1Year dataset. Table 11.1 summaries the data.

Tweets can have accurate geo tags via the GPS of mobile devices. However, the tweets with geo tags in our dataset are pretty rare, with only 0.001% of them having geo tags. Since users can write their locations in their profiles in unstructured sentences, we decided to use the location fields in the users' profiles. However, the challenge is users could write "Mars," "Space," etc. Finally, we employed the Bing API[1] to map the free text into a location. The Bing API is good at recognizing

[1]http://www.bing.com/dev/en-us/dev-center

Table 11.1. Summary description of LSUsers1Year

Time Span	2012-10 to 2013-10
# total FP tweets	About 1.58 billion
# Class S tweets	1,096,862
# Class D tweets	1,034,437
# unique Class S users	976,834
# unique Class D users	894,976
# LS tweets with geo tag	15,800 (0.001%)
# LS users with location from Bing API	677,328 (About 40%)

Table 11.2. Summary of LS user with inferred location

Time Span	2012-10 to 2013-10
# Class S users with inferred location	368,237
# Class D users with inferred location	315,176
# Cites with at least 1 LS user	13,961
# Cites with at least 100 LS users	653
# Countries with at least 1 LS user	241
# Countries with at least 100 LS user	95

location from a sentence of free-text.

Previous work by Hecht et al. [31] suggested that the commercial map API like Yahoo! Geocoder may not return the correct geo info when parsing free text in 2011, e.g., "BieberTown" could be mapped to Missouri in the US. We randomly selected 100 free text location fields from our corpus and sent them to Bing API in 2013. Then we manually evaluated how many of them have been correctly mapped. We define "correctly mapped" as follows: If the location field is valid (e.g. "HOUSTON *TEXAS*"), it should be mapped to "Houston—TX—United States." If the location field is not valid (e.g."With Tygaaaa(:"), Bing API should not map it to anywhere. After the evaluation, we found that 83 of the free texts were correctly mapped. Therefore, the precision of Bing API in our experiment is 83%. Therefore, we decided to use Bing API to process our data. We first show the results for US locations, including US cities and states, because a large portion of LS tweets are posted in the US. Then we show the results for world locations, including big cities and countries. Table 11.2 shows the summary of the LS users with inferred location.

Table 11.3. Top 10 U.S. cities with life satisfaction

Rank By Population	City	Total LS Users	Class S User Ratio
20	Memphis	767	66.36 %
25	Nashville	866	64.78 %
40	Atlanta	3,416	62.76 %
42	Raleigh	654	62.69 %
69	Greensboro	517	62.67 %
51	New Orleans	1,107	62.42 %
17	Charlotte	1,069	61.83 %
12	Jacksonville	967	61.43 %
14	San Francisco	975	60.51 %
23	Denver	859	60.30 %

11.1 US Locations

We selected the top 100 US cities by population from Wikipedia[2]. If a city has less than 500 LS users in our dataset, we skipped it. In our final US city list, we have 52 cities. We ranked them by the percentage of their Class S users over all of their LS users. Table 11.3 and Table 11.4 show the top and bottom 10 US cities with Class S user ratio respectively. Memphis leads in the top 100 US cities, while Pittsburgh is at the bottom of the list. The biggest 3 cities, New York, Los Angeles, and Chicago, are ranked at 47/52, 32/52, and 41/52 respectively.

We used a similar method to rank the 50 US states and Washington, D.C. Table 11.5 and Table 11.6 show the top and bottom 10 US states with Class S user ratio respectively. We see that while differences between cities and states chose in rank are not noticeable, differences between top and bottom ranks appear large.

[2] http://en.wikipedia.org/wiki/List_of_United_States_cities_by_population

Table 11.4. Bottom 10 U.S. cities with life satisfaction

Rank By Population	**City**	**Total LS Users**	**Class S User Ratio**
73	Buffalo	967	53.26 %
34	Fresno	592	53.04 %
72	Lincoln	613	52.53 %
21	Boston	2,766	52.06 %
1	New York	9,955	51.34 %
43	Omaha	549	50.64 %
19	El Paso	533	50.09 %
33	Tucson	779	49.94 %
31	Las Vegas	3,135	49.47 %
61	Pittsburgh	1,786	49.16 %

Table 11.5. Top 10 U.S. states with life satisfaction

Rank By Population	**State**	**Total LS Users**	**Class S User Ratio**
25	Louisiana	6,108	60.51 %
24	South Carolina	5,599	60.37 %
9	Georgia	13,137	59.72 %
17	Tennessee	7,038	59.70 %
33	Arkansas	4,440	59.55 %
23	Alabama	7,370	59.40 %
50	DC	1,977	58.47 %
4	Florida	22,333	56.95 %
18	Missouri	7,909	56.87 %
10	North Carolina	12,731	56.72 %

Table 11.6. Bottom 10 U.S. states with life satisfaction

Rank By Population	**State**	**Total LS Users**	**Class S User Ratio**
46	Delaware	952	50.53 %
42	Maine	2,782	49.96 %
36	Nevada	5,809	49.92 %
30	Connecticut	4,019	48.94 %
43	New Hampshire	3,224	48.67 %
11	New Jersey	9,196	48.59 %
45	Montana	1,304	48.54 %
48	Alaska	2,371	47.95 %
14	Massachusetts	10,031	47.67 %
44	Rhode Island	1,754	45.32 %

Figure 11.1 shows the distribution of LS users across US states as a heat map. The percentage for Class S users ranges from 45.3% to 60.5%. Orange indicates the states have more people who are satisfied with their lives while purple shows more people who are dissatisfied with their lives. In terms of geo location, we see that people in northeastern states appear less satisfied with their lives, such as Massachusetts, Rhode Island, Connecticut, New Hampshire and New Jersey. People living in southeastern states appear more satisfied such as South Carolina, Georgia, Tennessee, Alabama, Louisiana, and Florida. Notice that Iowa ranked in the middle (22nd place). Future research is needed to understand the observation and validate them,

11.2 World Locations

Although we have LS users from 13,961 cities in the world, most cities do not have many users. There are several possible reasons; one is that Twitter users are mostly from several countries like the U.S. Another reason is that our method can only detect life satisfaction in English. Therefore, the detection of LS using our method for non-English speaking countries is not accurate, e.g., Shanghai which is the biggest city by population, only has 31 LS users by our method. Therefore, we only chose the top 50 cities by population from all over the world and only used the cities which have more than 200 LS users. This resulted in 16 cities. We ranked the cities and countries the same way we did in the last section.

We show the results in Table 11.7. The world cities are ranked by English Class S tweets ratio. We mapped the cities in the world map and show the result in

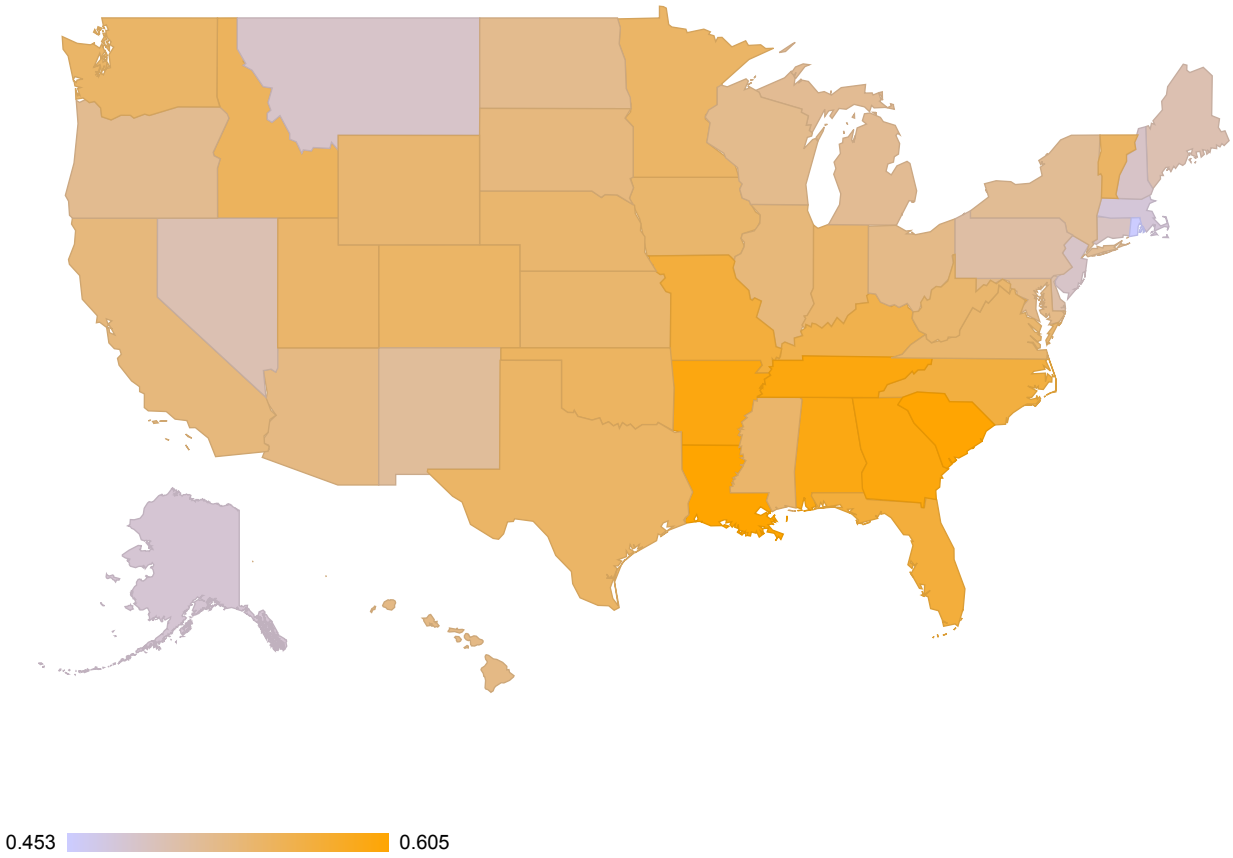


Figure 11.1. Life satisfaction for U.S. states

Table 11.7. Life satisfaction (English only) rank for world cities

Rank By Population	**City**	**Total LS Users**	**Class S User Ratio**
2	Lagos	1,335	78.13 %
14	Jakarta	1,580	66.96 %
5	Mumbai	396	63.13 %
6	Moscow	227	62.56 %
36	Rio de Janeiro	369	61.79 %
10	Delhi	800	56.38 %
18	Cairo	1,089	55.74 %
25	Bangkok	456	55.26 %
24	London	10,227	55.23 %
13	Seoul	531	54.61 %
48	Alexandria	453	52.54 %
42	Riyadh	889	51.86 %
23	New York	9,955	51.34 %
17	Tokyo	329	49.85 %
40	Singapore	648	47.69 %
30	Lima	305	40.33 %

Figure 11.2. Lagos, Jakarta, and Mumbai are the top 3 cities with highest Class S tweets ratio. Tokyo, Singapore, and Lima ranked the lowest. Again this is the life satisfaction rank only for English speakers.

Next, we analyzed the life satisfaction for countries. We only choose 39 countries which have more than 1,000 LS users. Table 11.8 and Table 11.9 show the top and bottom 10 countries with highest Class S tweets ratio respectively.

The heat map for life satisfaction of the 39 countries is shown in Figure 11.3. The percentage of Class S users ranges from 43.9% to 74.3%. The orange color shows the high Class S user percentage, while the purple color indicates the low Class S user percentage. If a country has not enough LS users (less than 1000 LS users), we

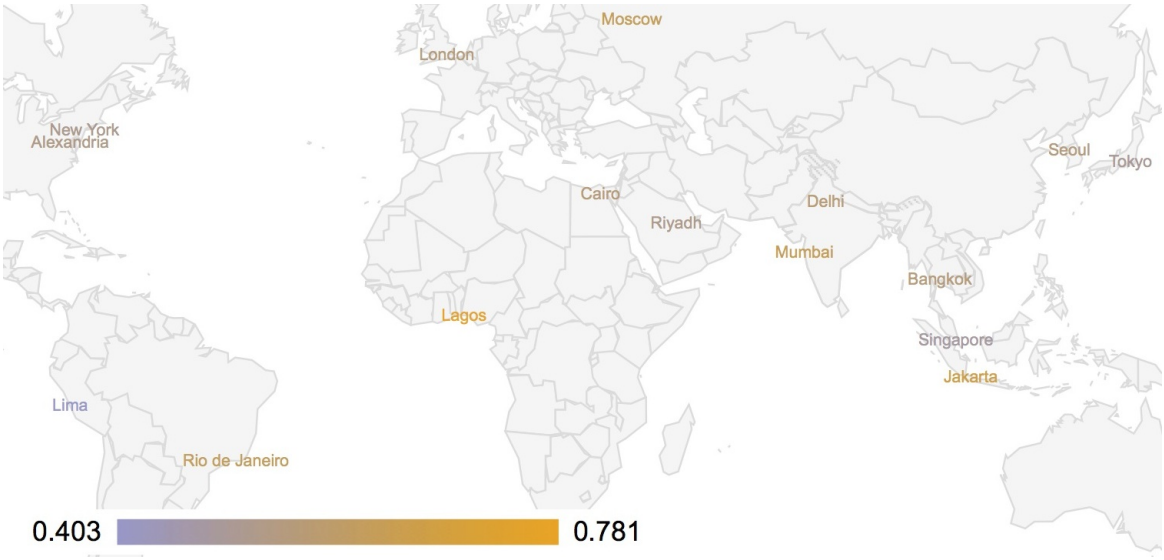


Figure 11.2. Life satisfaction (English only) for world cities

Table 11.8. Top 10 countries with life satisfaction (English only)

Rank By Population	**Country**	**Total LS Users**	**Class S User Ratio**
7	Nigeria	3,856	74.30 %
4	Indonesia	13,608	68.52 %
24	South Africa	7,473	65.70 %
177	The Bahamas	1,070	65.23 %
141	Jamaica	1,476	62.47 %
12	Philippines	11,874	59.63 %
28	Spain	3,057	56.79 %
2	India	6,590	56.18 %
5	Brazil	3,293	56.00 %
10	Japan	1,597	55.35 %

Table 11.9. Bottom 10 countries with life satisfaction (English only)

Rank By Population	**Country**	**Total LS Users**	**Class S User Ratio**
199	Jersey (UK)	1,497	49.90 %
51	Australia	8,687	49.74 %
23	Italy	5,374	49.53 %
31	Argentina	3,113	49.28 %
89	Sweden	1,631	49.05 %
113	Denmark	1,353	48.78 %
84	Portugal	1,036	48.55 %
18	Turkey	2,779	47.39 %
34	Poland	1,123	46.48 %
116	Singapore	4,910	43.91 %

marked them using white. From the figure we see that in southeastern Asia and some parts of Africa people seem more satisfied with their lives. People from the Middle East or eastern Europe appear less satisfied of their lives. Again, we only considered the English LS tweets when we were doing world location analysis.

In conclusion, we explored geographical distribution of LS users in this chapter. We presented distribution across U.S. cities and states, world cities and countries (English speaking only) They are preliminary observation, further research with additional data is needed to validate and explain the observation.

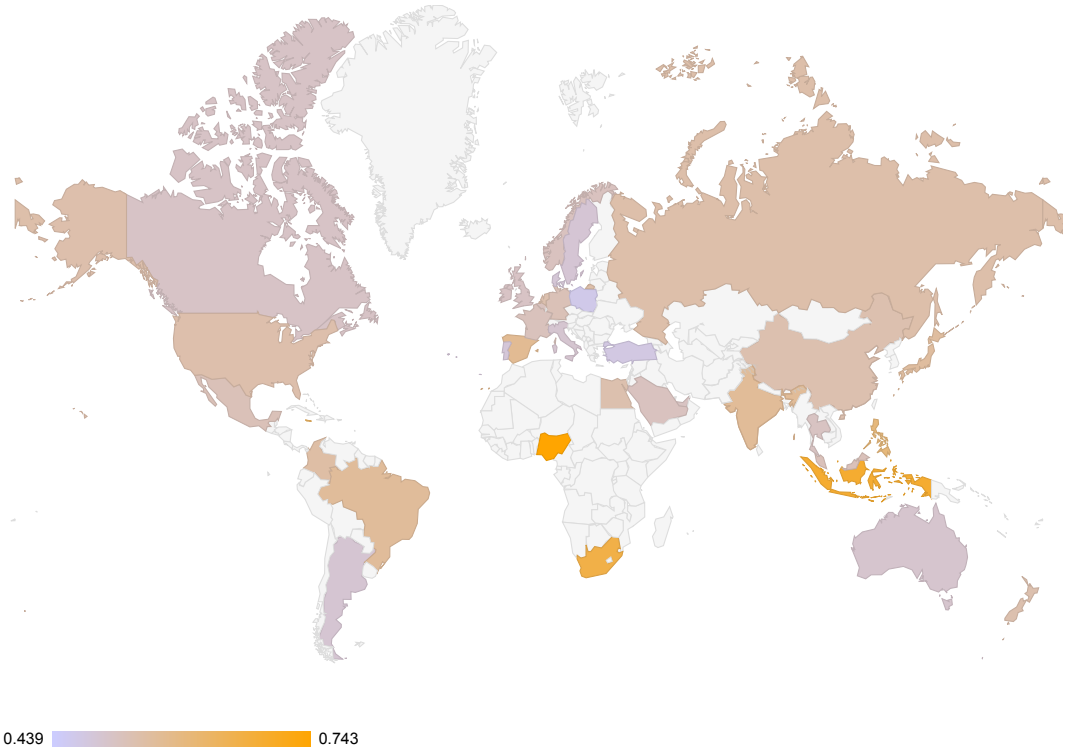


Figure 11.3. Life satisfaction (English only) for countries

CHAPTER 12

CONCLUSION AND FUTURE WORK

This thesis presents the first survey-to-surveillance method for translating a survey of interest into a surveillance strategy on social media. We start with a survey, obtain synonymous expressions for the survey statements, then generalize the expressions to templates using lexicons that are built simultaneously. The templates are then used as queries to retrieve tweets which can be considered as valid responses to the original survey statements. Filters are used to boost the accuracy of retrieval process.

We used life satisfaction (starting with a will respected survey) as a case study to illustrate our method. Our surveillance strategy for life satisfaction on social media outperforms other standard approaches that we designed for the same goal. Specifically, we tested two lexicon based methods: a) labMT, which is used to detect happiness, and b) ANEW, a more general sentiment lexicon. Both have low precision and are not efficient at detecting life satisfaction expressions. We conducted three machine learning experiments, while these have better performance than lexicon based methods. Their F scores are still not satisfactory compared to our template-based method. Therefore, our survey-to-surveillance approach builds the state-of-art surveillance method for tracking life satisfaction expressions on social media. The approach can be adopted to transform other surveys to their corresponding surveillance strategies. For example, surveys on surveillance of depression, personality, etc can be

the basis of surveillance on social media. This is our first contribution.

The thesis also presents a novel way to build a very large gold standard dataset, particularly when number of interesting data instances is far less than the uninteresting ones. Instead of evaluating every instance in the dataset or in a random subset as in the traditional approach, we go through a process of finding the positives; the rest are assumed to be negatives. Our "annotate by finding" method supports more complete evaluation: we can compute recall, since we have a reasonable estimate of the number of positives in the dataset. We built a gold standard dataset for life satisfaction using this method and used this to test our surveillance of life satisfaction on social media. We further validated the method by building three more TREC microblog topic datasets. Compared with TREC's "pooling" method, our method finds more positives. This "annotation by finding" method can be easily adopted to build other gold standard datasets (e.g., for expressions of depression). This is our second contribution.

The third contribution of this thesis is the empirical analysis of observations regarding life satisfaction. At the tweet level, we show the time series, daily and weekly cycle of life satisfaction. We found the overall time series of life satisfaction seems to fluctuate randomly. Therefore, one important inference is that life satisfaction appears to be unaffected by local events, world events, and seasons. This is very different from conclusions in previous happiness studies, which found that "happiness" on Twitter associated with events like Michael Jackson's death and U.S. Independence Day. This difference is consistent with differences in the definitions of

two components of subjective well being. Life satisfaction is the stable and long term assessment of one's life. For the daily and weekly cycle of life satisfaction, we found the LS tweets were posted less at noon and much more in the night. We also found Wednesdays have the least LS tweets, while Sundays have the most LS tweets.

At the user level, we compared differences of characteristics for users who are satisfied or dissatisfied with their lives. The characteristics include Twitter metadata, LIWC categories such as psychological processes, personal concerns, etc. We found the number of followers and followings are similar for the two classes of users. Class S users have more user active days and hashtags, while Class D users have more URLs. For all the categories which have more than 10% difference in LIWC, Class S users have more positive sentiment than Class D users. In particular, Class S users have positive sentiment towards sex, while Class D users have negative sentiment towards it. Also Class S users have more positive sentiment tweets with religion words, while the average sentiment for Class D users is slightly negative.

We analyzed users who posted both Class S and Class D tweets. The number of these users is small, and most changed their opinions only once. The average change interval is about 80 days. It often shows that life satisfaction is a stable variable, as most users don't change their life satisfaction in a short time. To further explore users' life satisfaction, we did temporal analysis of factors associated with life satisfaction. Exploring 19 psychosocial variables, we found depression, anger, anxiety, death, sadness, money, leisure, social support, religion, and home events are associated with changes of users' life satisfaction. We found that in general Group

S has significantly fewer tweets in all the negative PV categories than other three groups. The result can be use to detect the Class S users who may at risk of becoming dissatisfied.

We finally explored the geography distribution of life satisfaction expressions across the U.S. and across places around the world. Bing API was used to detect the location of LS users from their free-text location profile fields. Memphis, Nashville, and Atlanta are the three U.S. cities with the highest Class S tweet ratio. New York, Los Angeles, and Chicago rank at 47/52, 32/52, and 41/52 respectively. We found that people in northeastern states appear less satisfied with their lives compared to people living in southeastern states. The global distribution analysis is limited by the fact that we only explore English only tweets.

In the future, we will focus on the detection of long tweets which will boost the recall performance of our method. We also would like to utilize the result from analysis of retrieval rules and improve the method. In terms of the observation of life satisfaction on Twitter, we could study models to predict which users may change from satisfaction to dissatisfaction with their lives. We could also analyze aspects such as assortativity by studying LS using user network.

With the continuing popularity of social media, obtaining meaningful information is becoming more and more important. Using surveys to query a random sample of individuals on social media is not practical and efficient. We conclude that the development of sophisticated survey-to-surveillance methods is critical for the social media research. Our method can translate a well-respected survey to a state-of-the-

art surveillance strategy on social media. In this manner, the area of computer science and domains such as psychology and social science could be bridged more efficiently.

REFERENCES

[1] Charu C Aggarwal and ChengXiang Zhai. *Mining Text Data*. Springer, 2012.

[2] Michael Argyle. *The Psychology of Happiness*. Routledge, 2013.

[3] Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE, 2010.

[4] Krisztian Balog, Gilad Mishne, and Maarten de Rijke. Why are they excited?: Identifying and explaining spikes in blog mood levels. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, pages 207–210. Association for Computational Linguistics, 2006.

[5] Jen Beaumont and Jennifer Thomas. Measuring national well-being-health. Technical report, UK Office for National Statistics, 2012.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[7] Catherine A Bliss, Isabel M Kloumann, Kameron Decker Harris, Christopher M Danforth, and Peter Sheridan Dodds. Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science*, 3(5):388–397, 2012.

[8] Johan Bollen, Bruno Gonçalves, Guangchen Ruan, and Huina Mao. Happiness is assortative in online social networks. *Artificial Life*, 17(3):237–251, 2011.

[9] Margaret M Bradley and Peter J Lang. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, University of Florida, 1999.

[10] Jilin Chen, Gary Hsieh, Jalal U Mahmud, and Jeffrey Nichols. Understanding individuals' personal values from social media word use. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 405–414. ACM, 2014.

[11] Luigi Curini, Stefano Iacus, and Luciano Canova. Measuring idiosyncratic happiness through the analysis of Twitter: An application to the Italian case, year = 2014. *Social Indicators Research*, pages 1–18.

[12] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116. Association for Computational Linguistics, 2010.

[13] Munmun De Choudhury, Scott Counts, and Michael Gamon. Not all moods are created equal! exploring human emotional states in social media. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[14] Munmun De Choudhury, Michael Gamon, and Scott Counts. Happy, nervous or surprised? Classification of human affective states in social media. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[15] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[16] Ed Diener and Robert Biswas-Diener. *Happiness: Unlocking the Mysteries of Psychological Wealth*. John Wiley & Sons, 2011.

[17] Ed Diener, Marissa Diener, and Carol Diener. Factors predicting the subjective well-being of nations. *Journal of Personality and Social Psychology*, 69(5):851, 1995.

[18] Ed Diener, Robert A Emmons, Randy J Larsen, and Sharon Griffin. The satisfaction with life scale. *Journal of Personality Assessment*, 49(1):71–75, 1985.

[19] Peter Sheridan Dodds and Christopher M Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456, 2010.

[20] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS one*, 6(12):e26752, 2011.

[21] Maeve Duggan, Nicole B. Ellision, Cliff Lampe, Amanda Lenhart, and Mary Madden. Social media update 2014. `http://www.pewinternet.org/2015/01/09/social-media-update-2014`, 2015. [Online; accessed 04-March-2014].

[22] M R Frank, L Mitchell, PS Dodds, and et al. Happiness and the patterns of life: A study of geolocated tweets. *Scientific Reports*, 2013.

[23] Gallup. Gallup healthways well-being index. `http://www.well-beingindex.com`, 2013. [Online; accessed 30-September-2013].

[24] Gallup-Healthways. State of well-being 2011: City, state and congressional district wellbeing reports. `http://www.well-beingindex.com/files/2011CompositeReport.pdf`, 2012. [Online; accessed 01-October-2013].

[25] Alastair J Gill, Robert M French, Darren Gergle, and Jon Oberlander. The language of emotion in short blog texts. In *CSCW*, volume 8, pages 299–302, 2008.

[26] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.

[27] Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 581–586. Association for Computational Linguistics, 2011.

[28] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[29] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[30] Bruce Headey, Jonathan Kelley, and Alex Wearing. Dimensions of mental health: Life satisfaction, positive affect, anxiety and depression. *Social Indicators Research*, 29(1):63–82, 1993.

[31] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 237–246. ACM, 2011.

[32] John Helliwell, Richard Layard, and Jeffrey Sachs. *World Happiness Report*. The Earth Institute, Columbia University, 2013.

[33] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[34] Yu-Heng Hung, Yang-Yen Ou, Ta-Wen Kuan, Chin-Hui Cheng, Jhing-Fa Wang, and Jaw-Shyang Wu. An emotional feedback system based on a regulation process model for happiness improvement. In *Orange Technologies (ICOT), 2014 IEEE International Conference on*, pages 205–208. IEEE, 2014.

[35] Elsa Kim, Sam Gilbert, Michael J Edwards, and Erhardt Graeff. Detecting sadness in 140 characters: Sentiment analysis and mourning Michael Jackson on Twitter. *Web Ecology*, 3, 2009.

[36] Adam D.I. Kramer. An unobtrusive behavioral model of gross national happiness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 287–290. ACM, 2010.

[37] Ethan Kross, Philippe Verduyn, Emre Demiralp, Jiyoung Park, David Seungjae Lee, Natalie Lin, Holly Shablack, John Jonides, and Oscar Ybarra. Facebook use predicts declines in subjective well-being in young adults. *PloS one*, 8(8):e69841, 2013.

[38] Gilly Leshed and Joseph 'Jofish' Kaye. Understanding how bloggers feel: Recognizing affect in blog posts. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1019–1024. ACM, 2006.

[39] Peter M Lewinsohn, J Redner, and J Seeley. The relationship between life satisfaction and psychosocial variables: New perspectives. *Subjective well-being: An interdisciplinary perspective*, pages 141–169, 1991.

[40] Chaeyoon Lim and Robert D Putnam. Religion, social networks, and life satisfaction. *American Sociological Review*, 75(6):914–933, 2010.

[41] Jalal Mahmud. Why do you write this? Prediction of influencers from word use. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[42] Yelena Mejova, Padmini Srinivasan, and Bob Boynton. GOP primary season on Twitter: Popular political sentiment in social media. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 517–526. ACM, 2013.

[43] Rada Mihalcea and Hugo Liu. A corpus-based approach to finding happiness. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 139–144, 2006.

[44] Gilad Mishne and Maarten de Rijke. Capturing global mood levels using blog posts. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 145–152, 2006.

[45] Lewis Mitchell, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth. The Geography of Happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5):e64417, 2013.

[46] Saif M Mohammad and Tony Wenda Yang. Tracking sentiment in mail: How genders differ on emotional axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (ACL-HLT) 2011*, pages 70–79, 2011.

[47] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[48] James W Pennebaker, Cindy K Chung, Molly Ireland, Amy Gonzales, and Roger J Booth. The development and psychometric properties of LIWC2007. *Austin, TX, LIWC. Net*, 2007.

[49] Daniele Quercia. Don't worry, be happy: The geography of happiness on Facebook. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 316–325. ACM, 2013.

[50] Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. Tracking "gross community happiness" from tweets. In *CSCW '12: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM Request Permissions, February 2012.

[51] Lenore Sawyer Radloff. The CES-D scale A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1(3):385–401, 1977.

[52] Ulrich Schimmack, Ed Diener, and Shigehiro Oishi. Life-satisfaction is a momentary judgment and a stable personality characteristic: The use of chronically accessible and stable sources, volume = 70, year = 2002. *Journal of Personality*, (3):345–384.

[53] George W. Snedecor and William G. Cochran. *Statistical Methods*. Iowa State University Press, 1989.

[54] David Stillwell and Michal Kosinski. mypersonality project. http://mypersonality.org/, 2013. [Online; accessed 30-September-2013].

[55] Allegra Stratton. Happiness index to gauge britain's national mood. http://www.theguardian.com/lifeandstyle/2010/nov/14/happiness-index-britain-national-mood, 2010. [Online; accessed 01-October-2013].

[56] Trevor Strohman, Donald Metzler, Howard Turtle, and W Bruce Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, volume 2, pages 2–6. Citeseer, 2005.

[57] Facebook Data Team. Continuing our study of happiness. http://www.facebook.com/notes/facebook-data-team/continuing-our-study-of-happiness/375901788858, 2010. [Online; accessed 01-October-2013].

[58] Rianne Van der Zanden, Keshia Curie, Monique Van Londen, Jeannet Kramer, Gerard Steen, and Pim Cuijpers. Web-based depression treatment: Associations of clients' word use with adherence and outcome. *Journal of Affective Disorders*, 160:10–13, 2014.

[59] Dongsheng Wang, Abdelilah Khiati, Jongsoo Sohn, Bok-Gyu Joo, and In-Jeong Chung. An improved method for measurement of gross national happiness using social network services. In *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, pages 23–30. Springer, 2014.

[60] N Wang, M Kosinski, D J Stillwell, and J Rust. Can well-being be measured using Facebook status updates? Validation of Facebook's Gross National Happiness Index. *Social Indicators Research*, pages 1–9, 2012.

[61] Yi-Chia Wang, Moira Burke, and Robert E Kraut. Gender, topic, and audience response: An analysis of user-generated content on Facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 31–34. ACM, 2013.

[62] David Watson and Lee Anna Clark. *The PANAS-X: Manual for the positive and negative affect schedule-expanded form*. Ames: The University of Iowa, 1994.

[63] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: Finding topic-sensitive influential twitters. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010.

[64] Chao Yang. *Mining Human Emotions, Sentiment, And Happiness In Texts*. The University of Iowa, 2013.

[65] Chao Yang, Padmini Srinivasan, and Philip M Polgreen. Automatic adverse drug events detection using letters to the editor. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1030. American Medical Informatics Association, 2012.

[66] Huahai Yang and Yunyao Li. Identifying user needs from social media. Technical report, IBM Tech Report. goo. gl/2XB7NY, 2013.

[67] Tal Yarkoni. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3):363–373, 2010.

[68] Jichang Zhao, Li Dong, Junjie Wu, and Ke Xu. Moodlens: An emoticon-based sentiment analysis system for chinese tweets. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1528–1531. ACM, 2012.