Estimating happiness of most populous cities by mining user-generated content in Twitter

L.M. Vachan[1], D. Karthik ragunaathan[2]

[1]*Bachelor of Engineering, Department of Electrical and Electronics Engineering, PSG College of Technology, Coimbatore, India*
[2]*Bachelor of Engineering, Department of Electrical and Electronics Engineering, PSG College of Technology, Coimbatore, India*

---***---

Abstract - *In online networking, individuals around the globe produce limitless measures of content substance. Online networking users broadcast their state of mind, suppositions, and sentiments on twitter as tweets. In this paper we evaluate the happiness of the general population living in the top 40 most populous cities around the globe. We gather and analyze a huge corpus of geo referenced messages or tweets, posted by online networking users through twitter information provided by Twitter API. We analyze the text content information using Natural Language Processing (NLP) methods and classify the sentiment of individual tweets using various lexicon based methods associated with Word Sense Disambiguation (WSD) algorithm for computing the happiness coefficients of individual cities and form a happiness index based on it.*

Key Words: **Corpus, Natural Language Processing, Twitter API, Lexicon, Word Sense Disambiguation Algorithm.**

1. INTRODUCTION

In 2011, the UN General Assembly passed a determination welcoming part nations to gauge the happiness of their citizens and to utilize this to direct public policies [1]. Tailing this occasion interest in the fundamental satisfaction or prosperity of a man, gathering of individuals, or country has been growing rapidly, receiving a lot of consideration in the psychological literature [2].Importance of knowing happiness of their citizens for the local administrations has been becoming quickly on the grounds that these days happiness is viewed as an essential component for deciding prosperity of the citizens. Recent work at this measure was undertaken by Ed Diner et al to conceptualize the idea of "subjective prosperity"[3].

The elements that can affect happiness of a man could be diverse because of various human viewpoints. We can't just finish up a man living in Japan is more satisfied than a man living in India as a result of single reason that Gross Domestic Product (GDP) of Japan is more greater than that of India. Happiness is corresponded with numerous components like wealth, well-being conditions, and environment, psychological wellness, which incorporates anxiety, misery, and emotional problems [4].

Along these lines there must be numerous elements considered for measuring happiness. Despite the fact that in present strategies for measuring happiness different elements are considered, dependability of these techniques is low. For estimations of happiness in day by day encounters a present standard is the Experience testing strategy (ESM) in which the members are met at consistent interims to record their everyday feelings and emotions [5]. But the major disadvantage in this technique lies in the way that it is costly and hard to actualize in large samples [5,6]. The unwavering quality in these techniques are low since amid meetings of the as members do not tend to tell that they are dismal despite the fact that they are in awful state of mind or having a tragic living.

In this paper our goal is to measure emotional content of words used in large scale user generated text in twitter and by generating an overall score for the text. This overall score signifies how much the individual is sad or happy.

2. Approach

2.1 Sentiment analysis

Sentiment analysis is defined as the process which automates the mining of emotions, opinions, attitudes from text (tweets) using Natural Language Processing (NLP) methods. It involves classifying emotions contained in twitter text into positive, negative and neutral.

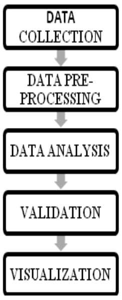

Fig. 1: Process involved in estimating happiness

3. Data collection and data set

As mentioned in the earlier section the data gathered for this survey is from Twitter Inc. In total we have gathered around two lakh geo-tagged tweets for 40 cities around the world. We have pulled approximately5000 tweets for every city over a period of one month (from June 1, 2016 to July 2, 2016). We limited our data set to only first person tweets like tweets containing I, my, mine, me, etc. Moreover, all tweets are labeled as positive, negative or neutral and certain a valence is determined through lexicon based approaches. These two lakh tweets came from 1, 89,546 distinctive users spread uniformly across 40 cities considered for this survey

4. Data pre-processing

The tweets contain lot of opinions; emotions expressed by users and also have lots of redundancy and inconsistency so that of pre- processing has to be done in twitter text. The following pre-processing steps are involved in cleaning the text data. The original twitter text obtained from Twitter Inc does contain lot of insignificant data.

Before analysis of twitter text there has be a lot of cleaning and parsing to be done. Because text data is human generated, it can be understood by humans and not considerably by machines. Hence data is structured in a way that machines can classify the text. Steps involved in this level are removing url's (like www.cvb.com), punctuations and numbers (.,@,#,$,%,etc.), conversion of upper case elements into lower case elements, removing stop words(very common words like the, of ,is ,was etc) and stemming (the process of replacing derived words with their linguistic root words). All words in the tweets are stemmed using Porter stemming (Porter, 1996) [8].

5. Base line data analysis

5.1 Lexicon

Senti word net [7] lexical resource was used for our sentiment analysis application. It provides annotations based up on three sentiments (positive, negative, and neutral). For proficient classification by this approach we associated this lexicon with Word Sense Disambiguation (WSD) algorithm.

5.2 Methodology

In lexicon based approaches used in sentiment analysis, the valance (sentiment) of a text depends upon the polarity of words contained in the text. Let us consider a tweet T which has N terms $P_1, P_2,...,P_N$. In this proposed method, the valance of tweet T is obtained by considering the valence of all terms $P_1, P_2,...,P_N$ in T. The valence of each term in the tweet T is obtained by matching each term in the text T with the terms in the lexicon using string matching algorithm.

5.3 Description of the approach

Consider a tweet T which consists of N terms $P_1, P_2,...,P_N$. The sentiment score Q of the tweet T is defined as follows

$$Q \triangleq \sum_{i=1}^{N} \text{valence (Pn)} \qquad (1)$$

Normalization is done to weigh tweets of varying lengths. Normalization of sentiment score Q of each tweet T is done with the number of terms N in the tweet T

$$S \triangleq \frac{\sum_{i=1}^{N} \text{valance(Pn)}}{N} \qquad (2)$$

Here, overall score S represents the sentiment of tweet T.

The above process is repeated for all tweets pulled from the cities and happiness coefficients of individual cities are calculated. The happiness coefficient of a city is the average of all sentiments S collected in the city, where sentiment S is computed for every tweet T pulled from the city. The Happiness coefficient of a city c is

$$H(c) \triangleq \frac{\sum_{i=1}^{K} S_i}{K} \qquad (3)$$

Where S_i is the Sentiment of a tweet T_i pulled from the city c and K is the total number of tweets from c.

6. RESULTS AND DISCUSSION

Rather than classifying the cities into happy and unhappy cities based on clustering methods, we developed an index using the obtained results from this analysis. Following **Error! Reference source not found.**shows number of cities with a given happiness coefficient. The mode of the observed results lies in the range 0.2 to 0.3.

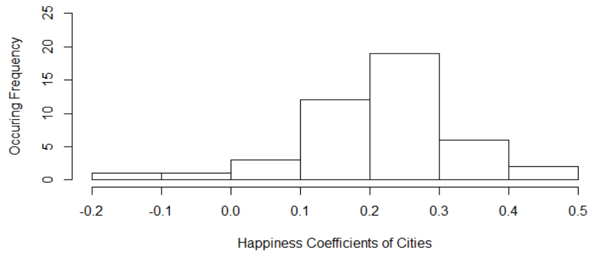

Fig.2 Happiness coefficients and frequency of occurrence

Fig.1 illustrates the distribution of Happiness coefficient for 40 most populous cities around the globe. The range of Happiness coefficient in the following distribution is 0.607. The minimum value is -0.137 which corresponds to Istanbul and the maximum is 0.4700 which relates to Bern

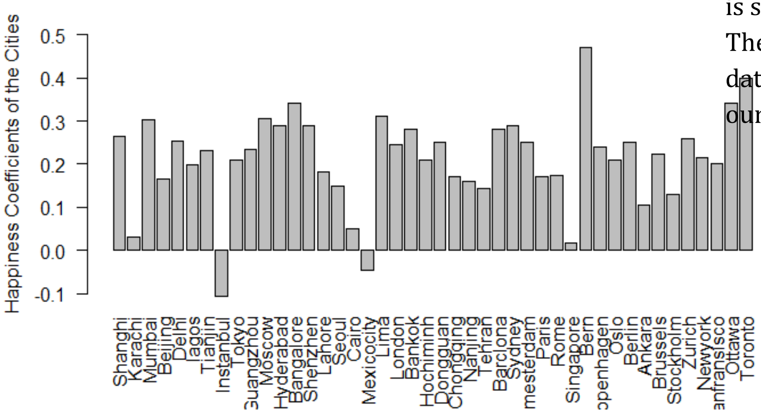

Fig.1: Distribution of Happiness coefficients of Top 40 Most Populous cities

Fig.2 shows the Happiness coefficients of the top 40 most populous cities around the globe. The illustration shown below is developed through ggmaps package [9] in R Studio. The coefficients are represented by shades of blue with light blue representing happier cities.

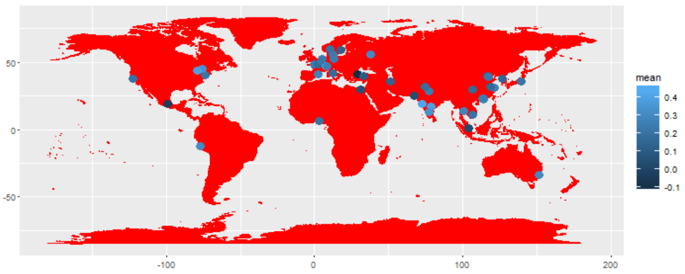

Fig.2 : Map Illustrating Happiness index of the cities developed using ggmaps package in R

7. VALIDATION

For Validation of this process naive bayes machine learning classification technique is utilized. Naive bayes classification technique is based on conditional probability model. The efficiency of this approach is determined by training the naive bayes classifier. In this process the data set (Tweets obtained from Twitter) is split in to training (35%) and test data (65%). The classifier is trained by training set data. The sentiment scores of test data tweets is predicted by this classifier.

7.1 Description of data set

The data set consists of 4205 user generated tweets. It is split in to 1445 training tweets and 2760 test tweets. The following table 1.1shows the description of training data set in which the sentiment of tweets are evaluated by our lexicon based method.

Table 1: Description of training data set

Sentiment of tweets	Number of tweets
Positive	676
Negative	356
Neutral	413

The sentiments of the tweets in test set are predicated by the naive bayes trained classifier. The results obtained in this analysis are satisfactory. The confusion matrix of the classification method illustrates the efficiency of the method.

Table 2 illustrates the confusion matrix of sentiment scores with actual and predicted sentiment values using naive bayes classification method

Table 2: confusion matrix with predictive parameters

		PREDICTED		
		Positive	Negative	Neutral
ACTUAL	positive	632	260	318
	Negative	196	401	199
	Neutral	121	111	522

The efficiency obtained by naive bayes classification technique is 56.34%

8. CONCLUSION

This paper is a survey of sentiment analysis on twitter data using Senti-word net lexicon and gives functional ways to identify and extract happiness of users from their twitter text. A lexicon based methodology is a straightforward and suitable way to deal with Sentiment Analysis of Twitter data. To derive better results, word sense disambiguation algorithmic associated with lexicon based approach. These results are validated with Naïve bayes machine learning algorithm and the efficiency of this approach is determined. Finally, The Happiness Index of 40 cities considered in our survey is computed.

Acknowledgements

We would like to thank Dr. S. Lovelyn Rose, Associate Professor in PSG College of Technology for her immense support. We would like to thank all annotators especially Mr. R. Ragavandra kumar, software development engineer from Intel and Dr.Vishal kesari from Defence Research and Development Organization. This work was supported by data grant from Twitter.

REFERENCES

[1] World Happiness REPORT Edited by John Helliwell, Richard Layard and Jeffrey Sachs

[2] Cacioppo, J. T., Gardner, W. L., & Bernston, G. G. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. Personality and Social Psychology Review, 1, 3-25.

[3] Diener, Diener, & Diener (1995). Factors predicting the subjective well-being of nations. Journal of Personality and Social Psychology, 69, 851-864.

[4] Schyns, Peggy. "Cross national differences in happiness: Economic and cultural factors explored. " Social Indicators Research 43.1-2, pp. 3-26, 1998.

[5] Csikszentmihalyi M, Larson R. 1987. "Validity and reliability of the Experience-Sampling Method." Journal of Nervous and Mental Disease, Sep;175(9):526-36.

[6] Ferring, D., S.-H. Filipp and K. Schmidt: 1996, The "Skala zur Lebensbewertung:" Scale construction and findings on reliability, stability, and validity, Zeitschrift für Differentielle und Diagnostische Psychologie 17, pp. 141–153.

[7] Andrea Esuli Baccianella, Stefano and Fabrizio Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of LREC, volume 10, pages 2200–2204, 2010.

[8] Ahmad, F., Yusoff, M. and Sembok, T.M.T. (1996), "Experiments with a stemming algorithm for Malay words", Journal of the American Society for Information Science, Vol. 47 No. 12, pp. 909-918

[9] Kahle,D and wickham,H. (2013) ggmap:Saptial Visualization with ggplot2. The R *Journal*, 5:144-161.