A Machine Learning Approach to Correlate Emotional Intelligence and Happiness Based on Twitter Data

Sistla Sai Shravani
Third-year Undergraduate student
Department of E&ECE, IIT Kharagpur
shravanisistla1998@iitkgp.ac.in

Niraj Kumar Jha
Third-year Undergraduate student
Department of E&ECE, IIT Kharagpur
nirajj100@iitkgp.ac.in

Rajlaksmi Guha
Assistant Professor, Centre For Educational Technology, IIT Kharagpur
rajg@cet.iitkgp.ernet.in

In this study, we have examined the relation between emotional intelligence and happiness. We have identified the traits of high emotionally intelligent people and low emotionally intelligent people and the corresponding words used on twitter to portray those traits. We have scraped twitter and extracted 1000 tweets for each word at three instances of time namely March 2018, 2013 and 2010. Two classifiers namely Support Vector Machine and Naïve Bayes were trained with large data sets to perform sentiment analysis. Each of them classifies a sentence either as positive(happy) or negative(sad). The different sets of scraped tweets corresponding to each word have been used as test data to the above models. Thus, a correlation between emotional intelligence and happiness over time was established. The underlying assumption of the above study is that the individual is expressing his/her true emotion on twitter.

Emotional Intelligence, Happiness, Twitter Scraping, Classifiers, Sentiment Analysis

1. INTRODUCTION

Emotional intelligence(EI) can be defined as the ability to monitor one's own and other people's emotions, to discriminate between different emotions and label them appropriately, and to use emotional information to guide thinking and behaviour [Coleman, 2008]. Emotions may interact with thinking and allow people to be better decision makers. A person who is more responsive emotionally to crucial issues will attend to the more crucial aspects of his or her life [Mayer, 2008].

Happiness, which is defined in terms of the average level of satisfaction over a specific time period, the frequency and degree of positive affect manifestations, and the relative absence of negative affect [Argyle, 1989]. The fact that some individuals are consistently happier than others suggests dispositional causes underlying the pursuit and experience of happiness. Indeed, personality traits are arguably the most robust predictors of happiness.

The CAT (Conceptual act theory) suggests that language plays a role in emotion because language supports the conceptual knowledge used to make meaning of sensations from the body and world in a given context [Barrett, 2006b]. A person's emotional intelligence is well reflected through language. In the 2000s, sites like Twitter and Facebook brought social networking to the mainstream, reaching millions of members in a fairly short period of time. People tend to share every tiny detail of their life on social media.

Machine Learning is the study of algorithms that can learn from and make predictions on data [Kohavi,1998]. We have scraped tweets which depict various behavioural traits of people having high EI and low EI. Sentiment analysis has been done to establish whether those tweets were used in a happy or a sad context. We made use of this characteristic to establish the correlation between emotional intelligence and happiness using the data obtained from social media platform.

2. LITERATURE REVIEW

There exist strong associations between happiness and personality traits [DeNeve, 1989]. In recent years, dispositional explanations of happiness have also emphasized the importance of emotional intelligence. Unlike cognitive ability, EI is the most reliably assessed via self-report inventories, rather than objective performance tests. Sentiment analysis has been used in the context of e-learning [Ortigosa, 2014], by companies to assess the customers' opinion about their products. In this work, we have made an attempt to relate EI and happiness

using sentiment analysis on social media data, which has never been tried earlier.

2.1 Why Twitter?

Millions of messages appear daily in twitter. Authors of those messages write about their life, share opinions on variety of topics and discuss current issues. Twitter's audience varies from regular users to celebrities, company representatives, politicians, and even country presidents. Therefore, it is possible to collect text posts of users from different social and interests. Twitter contains an enormous number of text posts and it grows every day. The collected corpus can be arbitrarily large and can serve as the perfect sample set for sentiment analysis and thus opinion mining.

3. DATA COLLECTION

The traits of people having high EI [Power, 2017]

Table 1: *EI TRAITS*

EI LEVEL	TRAIT	QUALITIES
High	*Change agents*	*Flexible, versatile, adjustable*
High	*Empathetic*	*Rapport, understanding, gentle*
High	*Self-aware*	*Conscious, alert, mindful*
High	*Balanced*	*Sensible, mature, rational*
High	*Gracious*	*Elegant, Polite, sociable*
High	*Curious*	*Inquistive, wide eyed*
High	*Innovative*	*Inventive, original, Experiment*
High	*Focus*	*Mindful*
High	*Optimistic*	*Positive, confident,*
High	*Assertive*	*Bold, firm, decisive*
Low	*Impatient*	*Eager, jumpy, anxious*
Low	*Frustrated*	*Annoyed, angry, crybaby*
Low	*Overreacting*	*Panic, chaotic, losing it*
Low	*Argumentative*	*Battling, Falling-out, Debate*
Low	*Oblivious*	*Absent, dreamy, inattentive*
Low	*Pessimistic*	*Negative, depressed, gloomy*
Low	*Victimized*	*Con, deceived*
Low	*Oversensitive*	*Moody, tense*

and low EI [Cherry, 2017] have been gathered from various resources and described below in Table 1.

3.1 Scraping Twitter Data

Since we have used twitter to collect the database, informal words have been taken into consideration which are analogous to the above-mentioned qualities. For example, people with high EI are hopeful about their future so 'brightfuture' is one such words used to portray this quality. The most recent 1000 tweets containing the phrase '#brightfuture' which were posted beginning from March 2018, were scraped. Also, the most recent 1000 tweets beginning from March 2013 and most recent 1000 tweets beginning from March 2010 have been extracted. So, for each word, we have obtained three sets of 1000 tweets each. The same process has been repeated for all the words listed in Table 2. The tweets are downloaded using TweePy library. The method doesn't exclude the possibility of retweets but the entire data set contained original tweets in general (99.9%). We removed non-Unicode characters, and eliminated tweets that contain hyperlinks and also tweets that are shorter than 6 words in length as longer tweets are more likely to get classified when sentiment analysis is performed [Joshi, 2015].

Table 2: *WORDS USED FOR SCRAPING*

HIGH EI WORDS	LOW EI WORDS
#brightfuture	*#angry*
#teamup	*#irritated*
#change	*#clueless*
#sorry	*#darkfuture*
#excited	*#deceived*
#thankful	*#sorrynosorry*
#hope	*#depressed*
#thoughtful	*#restless*
#happy	*#fallingapart*
#innovation	*#hopeless*
#Ican	*#fml*
#Iwill	*#novemberrains*

4. SENTIMENT ANALYSIS

It is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in-order to determine whether the writer's attitude towards a particular topic is positive, negative, or neutral.

4.1 Model Training

Conventionally, supervised lexicalized Natural Language Processing(NLP) approaches take a word and convert it to a symbolic ID, which is then transformed into a feature vector where each

dimension of the vector represents a feature. One of such representations is word embeddings. Word2vec is a group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space.

The corpus in this case was the common crawl dataset [Mikolov, 2017] and the word vectors were obtained using fast text algorithm of Facebook AI research (FAIR) which gave us around 2 million word vectors. The data set containing the product/movie reviews of imdb, amazon and yelp which was present initially in the form of words, was converted into word vectors using the above mentioned common crawl dataset. The final vectors served as the training set. Two classifiers – SVM, Naïve Bayes models were trained with the final feature set being the word vectors and the output being 0 or 1.

The SVM classifier has showed a training accuracy of 87.9% and validation accuracy of 82.33%. The Naïve Bayes has a training accuracy of 83.2% and validation accuracy of 81.05%. The models scale the output between 0 and 1, if the output value is greater than 0.5, it is classified to be positive and if it is less than 0.5, it is classified to be negative. Positive symbolizes happy whereas negative is used to represent sad sentences.

4.2 Model Testing

The above trained models have been used to test for around 72000 scraped tweets. All the tweets containing words related to trait EI have been classified as positive(happy) or negative(sad). For a specific word, if the number of positive tweets is way more than the number of negative tweets, that trait EI word has been classified to be used in a happy context (on a general note) whereas if it is the other way round, that trait EI word refers to sad context.

5. RESULTS

For each word in consideration, we have shown the number of tweets classified as positive by different models in Table 3. For example, if we consider the word '#excited', it has been classified as positive 878 times and negative for 122 times when tested using SVM. The person is excited since he/she is not merely using 'excited' in a sentence but describing his state by using '#excited' which concludes that an excited person (High EI) is generally happy. If we consider the hashtag '#clueless', it has been used 99 times in a positive context and 901 times in a negative context (when classified using SVM). The person is considered to have tweeted in a clueless situation by using the hashtag which concludes that a clueless person (Low EI) is sad/disturbed most of the times.

5.1 Discussions

All the words except change, sorry, ican, iwill were observed to follow a regular trend. All the high EI words excluding the above mentioned ones were found to be majorly used in happy context whereas all the low EI words are concluded to be used in a sad context. '***' in Table 3 indicates a conflict in classification by SVM and Naïve Bayes. 'Change' was classified as sad using SVM in 2018 and 2010 but as happy using Naïve Bayes. In 2013 tweets, 'change' was classified as sad using both the models.

The tweets with 'Ican' were used in an equal sense that is the number of negative tweets and positive tweets were nearly equal when scraped in 2013 and 2010. The tweets with 'Iwill' were more towards being used in a happy context in 2018, whereas they were found to be stagger in 2013 and 2010. Being sorry is a quality of people with high EI but the tweets having #sorry were found to be used in sad contexts.

The reason we have chosen three different time instances is an attempt to analyse how emotional expression has changed over the years. 800 million tweets are posted on an average in a day in 2018, 500 million in 2013 [Krikorian, 2013] and 35 million in 2010 [Weil, 2010]. This is an evidence to the fact that the number of twitter users is increasing drastically over time. We observed that the time required to collect 1000 tweets is not the same for each word and also varies for each word from time to time. The word 'sorrynossorry' has not been found in 1000 instances before 2010. In order to address the above question, we define a quantity 'score' corresponding to each word for each year which is inversely proportional to the normalized frequency of an individual tweeting about his emotions. For a particular word, if 'time' denotes the amount of time in hours between the first tweet and the last(1000^{th}) tweet, and 'num' is the number tweets in a single day in a particular year which is 8, 5 and 0.35 respectively for years 2018, 2013 and 2010. As num is increasing over the years and time is expected to decrease, the product will represent a normalized measure of frequency of tweeting.

For example, if we take '#happy', in 2018, the first tweet was scraped on 20-03-2018 was at 20:20 hours, the last tweet on 20-03-2018 at 18:00 hours, so the time is 2.5 hours, and hence the score is given by 20. The above procedure was repeated to calculate the scores for all the words at all time

Table 3: *RESULTS*

SCRAPED WORD	NUMBER OF POSITIVE TWEETS IN 2018				NUMBER OF POSITIVE TWEETS IN 2013				NUMBER OF POSITIVE TWEETS IN 2010			
	SVM	*NB*	*SCORE*	*CLASS*	*SVM*	*NB*	*SCORE*	*CLASS*	*SVM*	*NB*	*SCORE*	*CLASS*
#brightfuture	*748*	*944*	*4792*	*Happy*	*685*	*861*	*1312*	*Happy*	*762*	*943*	*549*	*Happy*
#change	*366*	*700*	*112*	***	*263*	*481*	*95*	*Sad*	*220*	*554*	*55*	***
#excited	*878*	*842*	*232*	*Happy*	*882*	*722*	*7.5*	*Happy*	*860*	*793*	*104*	*Happy*
#happy	*884*	*795*	*20*	*Happy*	*905*	*737*	*5*	*Happy*	*841*	*785*	*28*	*Happy*
#hope	*706*	*674*	*120*	*Happy*	*680*	*654*	*48*	*Happy*	*600*	*625*	*70*	*Happy*
#Ican	*518*	*739*	*3180*	*Happy*	*544*	*371*	*102.5*	***	*477*	*514*	*720*	***
#Iwill	*749*	*752*	*2960*	*Happy*	*655*	*436*	*468*	***	*529*	*324*	*1587*	***
#Innovation	*593*	*836*	*36*	*Happy*	*543*	*777*	*75*	*Happy*	*532*	*804*	*19*	*Happy*
#Sorry	*96*	*326*	*488*	*Sad*	*79*	*306*	*17.5*	*Sad*	*122*	*296*	*26*	*Sad*
#teamup	*764*	*822*	*10556*	*Happy*	*672*	*711*	*6600*	*Happy*	*615*	*651*	*2189*	*Happy*
#thankful	*708*	*836*	*120*	*Happy*	*836*	*668*	*45*	*Happy*	*803*	*743*	*175*	*Happy*
#thoughtful	*915*	*901*	*622*	*Happy*	*894*	*885*	*1118*	*Happy*	*849*	*955*	*12*	*Happy*
#angry	*106*	*135*	*185*	*Sad*	*97*	*190*	*15*	*Sad*	*130*	*205*	*146*	*Sad*
#clueless	*99*	*177*	*928*	*Sad*	*96*	*123*	*140*	*Sad*	*129*	*178*	*256*	*Sad*
#darkfuture	*342*	*332*	*VERY-LARGE*	*Sad*	*304*	*335*	*VERY-LARGE*	*Sad*	*298*	*323*	*VERY-LARGE*	*Sad*
#deceived	*110*	*214*	*168*	*Sad*	*143*	*202*	*153*	*Sad*	*130*	*245*	*77*	*Sad*
#depressed	*79*	*124*	*10*	*Sad*	*68*	*161*	*103*	*Sad*	*78*	*190*	*3*	*Sad*
#fml	*62*	*197*	*20*	*Sad*	*61*	*84*	*10*	*Sad*	*73*	*103*	*4*	*Sad*
#hopeless	*100*	*146*	*3788*	*Sad*	*71*	*208*	*322*	*Sad*	*97*	*159*	*1411*	*Sad*
#irritated	*16*	*178*	*5284*	*Sad*	*29*	*162*	*178*	*Sad*	*45*	*205*	*7*	*Sad*
#restless	*198*	*177*	*839*	*Sad*	*242*	*278*	*563*	*Sad*	*268*	*270*	*145*	*Sad*
#sorrynosorry	*341*	*394*	*81760*	*Sad*	*298*	*332*	*10800*	*Sad*	*--*	*--*	*--*	*--*
#fallingapart	*68*	*106*	*88*	*Sad*	*42*	*181*	*29*	*Sad*	*104*	*194*	*19*	*Sad*

instances and is shown in Table 3.

High score shows that the time taken to obtain 1000 tweets is more, and low score shows that the time taken to obtain 1000 tweets is less, indicating that large number of people are tweeting that particular word more often. The scores in 2018 were observed to be more than scores in 2013, 2010 which can be attributed to many factors such as the expression of emotion might have changed over the years which means people are resorting to using new words. Also, Global events can influence the frequency of usage of particular word which can give a very skewed value of score.

6. CONCLUSION

We observed that the tweets having high EI words were majorly classified to be used in a happy situation where as tweets having low EI words were classified to be used in a negative context. The implication could be that people with high emotional intelligence tend to be have greater distress tolerance which helps them extract happiness from the surroundings without being overwhelmed by any situation whereas people with low emotional intelligence have less control over situations and easily get unhappy. The above work has been done with assumptions that an individual is expressing one's own self on twitter and also the conclusions written below are not specific with respect to any individual but are drawn on a general note. So the variability of EI of an individual is not a matter of study.

6.1 Future Work

More words which depict traits of people picturing their emotional intelligence can be gathered to obtain better understanding. Artificial neural networks can be employed. Tweets can also be scraped with respect to geographical location. Different conclusions can be drawn with respect to Emotional Intelligence of people residing in different parts of the world and thus relating it to happiness.

7. REFERENCES

Argyle M, Martin M, & Crossland, J. (1989) Happiness as a function of personality and social encounters. In J. P. Forgas & M. Innes(Eds.), Recent advances in social psychology: an international perspective, North Holland.

Barrett L. F. (2006b) Solving the emotion paradox : categorization and the experience of emotion. Pers. Soc. Psychol. Rev., 10, 20–46.

Cherry, Kendra. (2017) Signs of Low Emotional Intelligence. www.verywellmind.com/signs-of-low-emotional-intelligence-2795958.

Coleman, Andrew. (2008) A Dictionary of Psychology (3 ed.). Oxford University Press, England.

DeNeve, K. M., & Cooper, H. (1998) The happy personality: a meta-analysis of 137 personality traits and subjective well-being. Psychological Bulletin, 124, 197–229.

Joshi Aditya, Mishra Abhijit, Balamurali AR, Bhattachacharyya Pushpak, Carman Mark J. (2015) A Computational Approach to Automatic Prediction of drunk-texting. In proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (short papers), Beijing, China.

Kohavi Ron, Foster Provost. (1998) Glossary of Terms. *Machine Learning*, 30, 271–274.

Krikorian, Raffi. (VP, Platform Engineering, Twitter Inc.). "New Tweets per second record, and how!" Twitter Official Blog. August 16, 2013.

Mayer, John D. (2008) Human Abilities: Emotional intelligence. Annual Review of Psychology, Vol. 59, 507-536.

Mikolov T., Grave E., Bojanowski P., Puhrsch C., Joulin A. (2017) Advances in Pre-Training Distributed Word Representations.

Ortigosa A., J.M. Martín, R.M. Carro. (2014) Sentiment analysis in Facebook and its application to e-learning. Comput. Hum. Behav., 31, 527-541.

Power, Rhett. (2017) Qualities of People with High EmotionalIntelligence. www.success.com/article/7-qualities-of-people-with-high-emotional-intelligence.

Turian Joseph, Lev Ratinov, and Yoshua Bengio. (2010) Word representations: a simple and general method for semi-supervised learning. In 48th Proceedings of ACL, 384–394.

Weil, Kevin. (VP of Product for Revenue and former big data engineer, Twitter Inc.). "Measuring Tweets." Twitter Official Blog. February 22, 2010.