Analysis and Visualization of University Twitter Feeds Sentiment

Arlene Caballero[1(✉)], Jasmin D. Niguidula[2], and Jonathan M. Caballero[2]

[1] Lyceum of the Philippines University Intramuros Manila, Manila, Philippines
arlene.caballero@lpu.edu.ph
[2] Technological Institute of the Philippines, 1328 Arlegui Street, Quiapo, Manila, Philippines
jasniguidula@yahoo.com, jonathanmcaballero@gmail.com

Abstract. The exponential growth of online social network as communication channel brought revolutionary changes in our daily lives. For the organizations, Twitter can be used for many reasons. It can be used as a channel of communication for expressing thoughts, emotions, experiences, perspectives, and opinions in a variety of topics and social interest. This study focused on the information theoretic review of sentiment analysis and visualization in Twitter. This paper examined the tweeter feeds from a select institution using a web-based sentiment analysis tool for analyzing tweeter sentiment and tweet visualization. This further investigates the clustering techniques and information theory applied to visualize and analyze the sentiments in the tweeter feeds using a query as the target of sentiments performed on over 1,500 tweeter feeds from a select institution users. In this study, the individual tweets from the users were converted into images and presented in forms of charts, graphs and diagrams to discover the nature of activity of the users. In view of this, an approach to data mining technique – Shannon information theory has been examined to analyze and review how the estimated sentiment in the corpus of data extracted from the tweeter feeds were processed and calculated. The tf-idf calculated for each query term in tweeter feeds were converted into images using information theoretic approach. With this, the nature of activity and opinions of the users in a select institution were discovered. This study also described the tweeter sentiments in an emotional scatter diagram mapped with pleasure and stimulation using the Russel Model of Affect.

Keywords: Social networking · Opinion mining · Twitter profile
Information theory · Term frequency-inverse document frequency

1 Introduction

Twitter is an online social networks which became one of the most popular micro-blogging services available in the web. As of midyear 2011, over 200 million tweets has been posted in Twitter per day and this has been the subject of attention of researchers of various organizations worldwide [1, 2]. The exponential growth of online social network as communication channel brought revolutionary changes in our

J. J. Jung et al. (Eds.): BDTA 2017, LNICST 248, pp. 132–145, 2018.
https://doi.org/10.1007/978-3-319-98752-1_15

daily lives [3]. For the organizations, Twitter can be used for many reasons. It can be used as a channel of communication for expressing thoughts, emotions, experiences, perspectives, and opinions in a variety of topics and social interest [3, 4]. Consequently, it serves as a valuable source of public opinion and communication [4].

Opinion Mining also called sentiment analysis, is the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [5]. It aims to pay attention and process the data being posted by the users in a social media [2].

Sentiment Visualization converts twitter sentiments into images which allow the viewer to see the values and the relationships of data as they form [6]. In this study, the data from a select institution represented by individual tweets from the users were converted into images and presented in forms of charts, graphs and diagrams to discover the nature of activity of the users. In view of this, an approach to data mining technique – Shannon information theory has been examined to analyze and review how the estimated sentiment in the corpus of data extracted from the tweeter feeds were processed and calculated.

This paper utilized and described the tweeter feeds from a select institution using *SentimentViz* - a web-based sentiment analysis tool for analyzing tweeter sentiment and tweet visualization [7]. This further investigates the clustering techniques and information theory applied to visualize and analyze the sentiments in the tweeter feeds using a query as the target of sentiments performed on over 1,500 tweeter feeds from a select institution users.

2 Related Works

This section reviews Shannon information theory and how it relates and intersects to other field of studies such as mathematical or probability theory and algorithmic complexity. This section relates the computational theory applied in sentiment visualization and calculating the estimated sentiment of the twitter feeds.

2.1 Shannon Information Theory

In the early development of human communication, people used pictures, scripts, codes, and symbols to represent objects and to communicate their ideas, actions, names, or by association. In Shannon "theory of communication" which in the latter called "theory of information" explains that the transmission of symbols, script, or message involves sending of information through electronic signals. This communication model consists of transmitter, channel, and a receiver. In the information source, message is being transmitted in a form of signal then, the received signal is sent to the destination [8].

This is further refined in the standard communication model as the source or encoder which translates the message into codes in the form of bits. A code is a set of symbols or a language that is used to transmit message or thought on one or more channel to get response in a receiver or decoder [8].

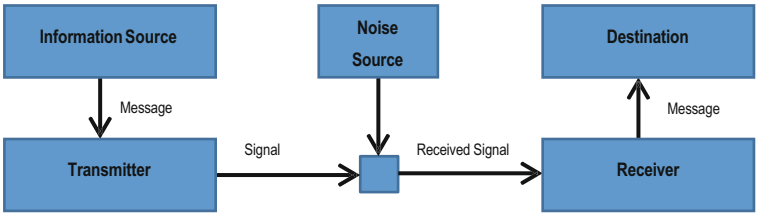

Fig. 1. Shannon model of information theory [8]

Figure 1 show how the message is being transmitted to the receiver through a channel. Whenever a message is transmitted, the noise source which refers to a number of variables makes the message to be distorted or changed. Hence, makes the recipient appear in a different state. These changes refers to entropy which measures the amount of uncertainty of an unknown or random quantity [8]. This communication problem is exactly what Shannon entropy theory tries to measure using the formulation:

$$H(X) = -\sum p(x)\, log_2\, p(x)$$

Over the years, Shannon theory has grown and evolved into modern communication from the attempts of several mathematicians and communication engineers to define and establish how information source is being communicated and how it can be measured [9]. In the same context, messages can also be transmitted through a form of social media such as Twitter and then convey these messages in a form of data visualization through digital and numeric calculation.

Figure 2 show that relationship of information theory to other fields [10]. In the area of computer science and probability theory, Shannon's quantification of entropy initiate the ideas for a relationship with term frequency-inverse document frequency (tf-idf) applied in this study to weight terms in the corpus of data extracted from the tweeter feeds.

2.2 Term Frequency-Inverse Document Frequency (tf-idf)

One of the most commonly used measures in information retrieval is the "term frequency-inverse document frequency (td-idf). This information weighing schemes is used to measure the probability-weighted amount of information in a given document [11].

In the conventional information theory, idf can be interpreted as 'the amount of information' given as the log of the inverse probability [11, 12]. By definition, tf-idf is a measure that multiples the two quantities tf and idf. With this, term frequency provides estimation of the occurrences probability of a term when it is normalized by the total frequency in the document, or the document collection, depending on the scope of the calculation.

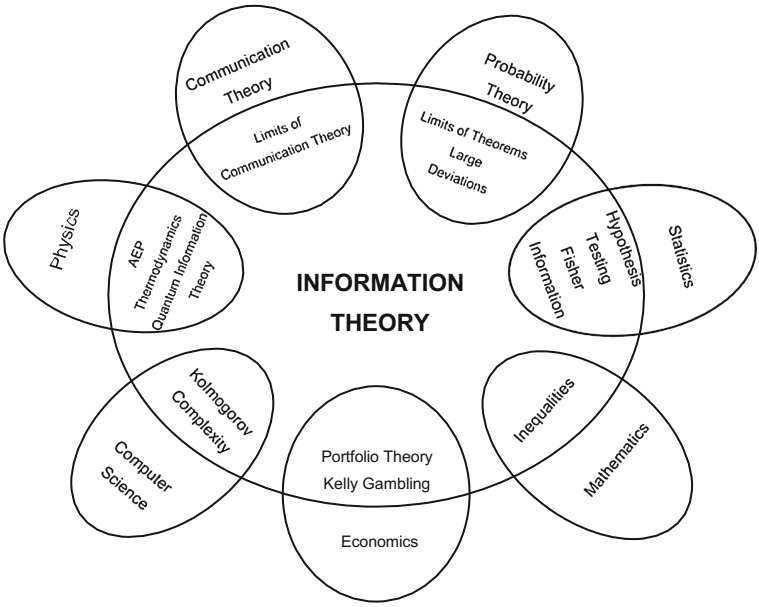

Fig. 2. Relationship of information theory to other fields

Based on the basic formula of information theory [8, 10], a document is assumed to be a given unordered set of terms. Let $D = \{d_j, \ldots, d_n\}$ be a set of documents and $W = \{w_i, \ldots, w_M\}$ be as set of distinct terms contained in D. In this study, the documents D is represented by a corpus of data extracted from the tweeter feeds while W is the query term. The parameters N are the total numbers of documents while M are the number of terms. In adapting the theory, selection of term w_i from W and selection of document d_j from D are also considered [11].

To illustrate the probability distribution of terms in a set of documents, the amount of information is calculated as illustrated in the figure below.

Figure 3 illustrates the calculation of expected mutual information when the user submits a query term to process and how the document is being selected using the term. In the estimate, the probability of query terms is represented by $P(w_i)$ while the probability distribution on the selection of documents is represented by $P(d_j \mid w_i)$. In the process, the co-occurrences of documents and terms are calculated [11].

In this study, the text clustering of Twitter feeds were classified using the theory of term frequency - inverse document frequency (TF-IDF). This measure is applied to evaluate the importance of word to a document in a given data set. The number of times a word appear in each tweet is proportional in the increase of importance and is offset by the frequency of word in a data set [7]. The goal of tf-idf technique for text analysis is to organize a document using a query term. This technique makes each tweet to be classified into sentiments explored in this study.

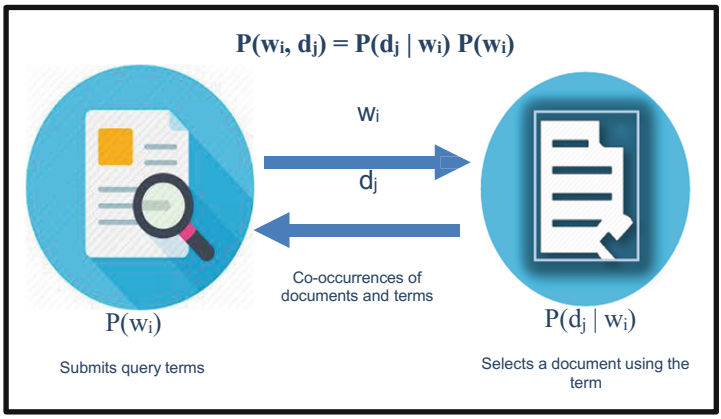

Fig. 3. Calculation of expected mutual information

3 Methodology

3.1 Sentiment Analysis and Visualization

Sentiment analysis is extracting and understanding human emotions from any given text, message or data. In general, sentiment analysis aims to classify documents into polarity of tweets such as positive, negative, or neutral. However, the tf-idf classification technique used in this study aims to classify documents into categories based on query terms. Then, map the terms by emotions using the Russel Model of Affect. The model used in this study proposed the use of valence (or pleasure) and arousal (or stimulation) represented in 2 dimensional plane to build emotional interpersonal circle of affect [7].

Figure 4 depicts the Russel Model of Affect which uses 2D plane emotional dimensions to position feelings or emotions. Along the vertical axis represents arousal (or stimulation) with intense or active on the upper quadrants and mild or passive on the lower quadrants with different levels of pleasures in between. On the horizontal axis, highly unpleasant and highly pleasant were mapped with different levels of pleasure in between. This model suggested using *valence* (or pleasure) and *arousal* (or stimulation) to build emotional interpersonal circle of affect [7].

3.2 Text Clustering Classification

For text clustering, the term frequency - inverse document frequency (tf-idf) was applied to weight the importance of word in a document. Initially, the normalized term frequency (tf) is computed. This refers to the number of times a query term appears in a given data set. Then, the inverse document frequency (idf) is computed which refers to

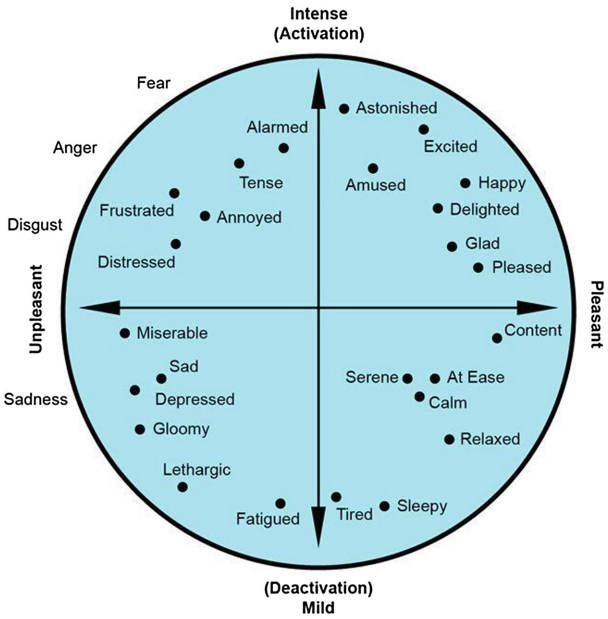

Fig. 4. The Russel Model of Affect

the logarithm of the number of documents where the query term appears. The following are the formula used to calculate tf-idf:

1. *tf (t) = (number of times a query term appears in a document) / (total number of query terms in the document).*
2. *idf (t) = log_e(total number of documents / number of documents with query term in it).*
3. *tf-idf = (term frequency * inverse document frequency)*

In computing the tf, all query terms from the select institution tweet feeds are assumed to be equally important. However, there are terms such as "the", "is", and "of" which may appear often but not as important as the query term. Therefore, an inverse document term frequency is computed to balance down the frequent terms and measure the non-frequent terms. Then, the tf-idf is computed by multiplying the result of term frequency (tf) and the inverse document frequency (idf).

The query term in each tweet feed calculated to have the highest tf-idf result will be selected for the estimate of sentiment. After weighing each term, the standard cosine similarity is applied to compute the pairwise similarity between all pairs of tweets from the corpus of data. The cosine similarity is used to compute for the Dot Product of two vectors which refers to the position of one point in a plot relative to another.

3.3 Estimating Sentiments

In estimating sentiment, an alternative method was used such as the use of dictionaries to determine the sentiments on words along a given tweet or message. The dictionary used *in SentimentViz* provided measures of valence and arousal for approximately 10,680 English words. As a result of previous research, words included in the dictionary were selected as candidates to express emotion [7].

To compute for the estimated sentiment, ratings for the common words found in the dictionary are combined with a mean rating and a standard deviation of the ratings for each dimension rating. For each word *wi* in the tweet that exists in the tweeter visualizer tool dictionary, word's mean valence $\mu_{v,i}$ arousal $\mu_{a,i}$ standard deviation of valence $\sigma_{v,i}$ and arousal $\sigma_{a,i}$ are saved.

Table 1 shows that in the tweet posted on September 1, 2016, there were more than two (2) words found in the dictionary. The words *"high", "school", "open",* and *"low"* has an overall valence of 5.55 while the overall mean arousal is 5.30. If a tweet contains less than $n = 2$ words found in the dictionary, it is ignored because it has an insufficient number of ratings to estimate its sentiment. The statistical average of the n means and standard deviations is computed to obtain the tweet's overall mean valence M_v and arousal M_a [7].

Table 1. Tweet valence and arousal

Date/Time *Sep 1, 2016 4:01am*	'Enrollment for LPU Manila Senior ***high School*** is NOW ***OPEN***! For a ***low*** downpayment of P5;000'				
Keywords	**Mean Valence**	**Standard Deviation**	**Mean Arousal**	**Standard Deviation**	**Frequency**
high	μ: 6.64	σ: 1.21	μ: 4.75	2.91	*fq* = 50
school	μ: 6.26	σ: 1.88	μ: 5.74	σ: 2.46	*fq* = 50
open	μ: 6.1	σ: 1.36	μ: 5.92	σ: 2.55	*fq* = 50
low	μ: 3.66	σ: 1.12	μ: 4.54	σ: 3.19	*fq* = 50
Valence			***v = 5.55***		
Arousal			***a* = 5.30**		

3.4 Sentiment Visualization

In visualization of sentiments, each tweet's estimated sentiment is represented by a circle mapped by emotions. An unpleasant tweet is plotted in blue circles while pleasant tweets are mapped in green circles. The stimulation or arousal is represented as brighter circles which indicate that the brighter the circle, the more active are the tweets. The confidence in the sentiment estimate is represented by the size and transparency. The larger the sizes of the circle, the more confident are the estimates.

Another measure of confidence of the tweet's emotion is the transparency. The more opaque or less transparent tweets, the more confident are the estimates.

Figure 5 illustrates how a single tweet with overall mean valence of *5.55* and overall mean arousal of *5.30* is being plotted in an emotional map in horizontal and vertical axes of pleasure and arousal. It further shows that the tweet lies on the upper right quadrant in the Russel Model which depicts that the tweet is generally pleasant.

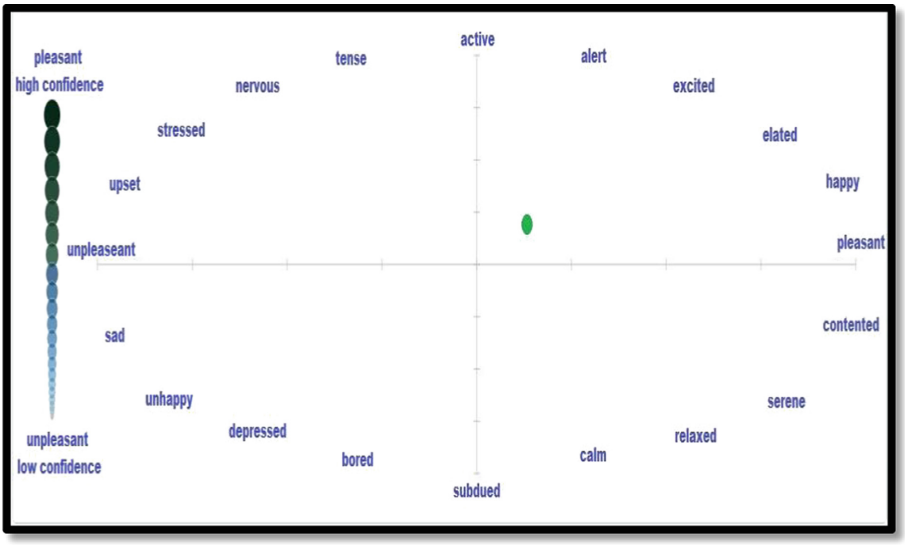

Fig. 5. Tweeter visualization

4 Results and Discussions

4.1 Select Institution Profile

This section describes the profile of the select institution in terms of frequency of posts and number of user interactions. The corpus of data examined in this study were gathered from the tweet feeds of a select institution from December 1, 2011 to June 29, 2016. The tweets were first analyzed by the average frequency per day, user mentions, and hashtags; percentage of retweets, and replies; and percentage of tweets retweeted, and retweets favorited.

As shown in Table 2, the analysis was performed over 1,557 tweets as corpus of data subject for estimated sentiment. On the average, there was 0.93 mean frequency of tweets posted each day and about 655 user mentions with 0.42 average number of mentions per tweet, 1087 hashtags with 0.70 average number of hashtags, 83 retweets or 5% retweets and 361 replies or 23% replies in the total of analyzed tweets. Tweets retweeted has a frequency of 967 or 62.1% while tweets favorited has frequency of 1174 or 75.4%. This shows that the users has more interaction with each other as the numbers calculated increases.

Table 2. Tweets analytics of the select institution

Total of 1557 Tweets From December 1, 2011 to June 29, 2016			
Tweet activity	Frequency	Mean average	Percentage
Tweets per day	0.93		
User Mentions	655	0.42	
Hashtags	1087	0.70	
Retweets	83		5%
Replies	361		23%
Tweets retweeted	967		62.1%
Tweets favorited	1174		75.4%

4.2 Tweets Calculation Using a Query 'Enrollment'

The following tables were the results from the query term 'Enrollment' or 'Enrolment" as the target of sentiments. The table shows the highlighted keywords found in the dictionary and the corresponding measures of valence and arousal.

Table 3 depicts that the tweet feed dated *June 8, 2016, 9:51 pm* highlighted keywords such as "*situation", "best", "option", "enroll", "will", "be", and "enrollment"* which has an overall valence of 6.24 and an overall mean arousal of 3.86.

Table 3. Tweets on enrollment dated June 8, 2016, 9:51 pm

Date/Time *Jun 8, 2016* *9:51pm*	@lxmntrsh_ if they are having ***situation***; the ***best option*** is to ***enroll*** in LPU and they ***will be*** assisted by ***enrollment*** advisers				
keywords	**Mean Valence**	**Standard Deviation**	**Mean Arousal**	**Standard Deviation**	**Frequency**
situation	μ: 5.0	σ: 1.31	μ: 4.08	σ: 1.92	*fq* = 50
best	μ: 7.18	σ: 1.69	μ: 4.6	σ: 2.67	*fq* = 50
option	μ: 6.49	σ: 1.31	μ: 4.74	σ: 2.23	*fq* = 5
enrol	μ: 6.19	σ: 1.94	μ: 3.76	σ: 2.63	*fq* = 23
will	μ: 6.83	σ: 2.04	μ: 2.76	σ: 2.05	*fq* = 19
be	μ: 6.18	σ: 1.44	μ: 3.43	σ: 2.31	*fq* = 21
enrollment	μ: 6.19	σ: 1.94	μ: 3.76	σ: 2.63	*fq* = 23
Valence			***v* = 6.24**		
Arousal			***a* = 3.86**		

Table 4 shows that the tweet feed dated *Oct 11, 2016 1:44am* highlighted keywords such as "*Enroll", "students", "deficiencies", "Please", "See", and "notice"* which has an overall valence of 5.35 and an overall mean arousal of 4.79.

Table 4. Tweets on enrollment dated Oct 11, 2016 1:44 am

Date/Time *Oct 11, 2016 1:44am*	#*Enrolment* Schedule is now up at http://t.co/zcX5MI0ZCF; For *students* w/ *deficiencies*; *please see notice* there also.@LPUPirates @TagaLPU				
keywords	Mean Valence	Standard Deviation	Mean Arousal	Standard Deviation	Frequency
Enroll	μ: 6.19	σ: 1.94	μ: 3.76	σ: 2.63	*fq* = 23
students	μ: 6.28	σ: 1.83	μ: 5.12	σ: 2.46	*fq* = 50
deficiencies	μ: 2.74	σ: 1.24	μ: 4.2	σ: 2.53	*fq* = 19
please	μ: 6.36	σ: 1.68	μ: 5.44	σ: 2.88	*fq* = 50
see	μ: 6.06	σ: 1.06	μ: 6.1	σ: 2.19	*fq* = 50
notice	μ: 5.16	σ: 1.5	μ: 3.93	σ: 2.56	*fq* = 50
Valence			***v* = 5.35**		
Arousal			***a* = 4.79**		

Table 5. Tweets on enrollment dated Oct 16, 2016 1:26 am

Date/Time *Oct 16, 2016 1:26am*	Do you have any concern about # Enrollment? Just send a PM to http://t.co/T9cG1fdIJJ and we'd be glad to address them. #OnlineHelpDesk				
keywords	Mean Valence	Standard Deviation	Mean Arousal	Standard Deviation	Frequency
concern	μ: 4.04	σ: 1.62	μ: 5.07	σ: 2.74	*fq* = 50
Enrollment	μ: 6.19	σ: 1.94	μ: 3.76	σ: 2.63	*fq* = 23
send	μ: 5.38	σ: 1.35	μ: 5.63	σ: 2.36	*fq* = 50
be	μ: 6.18	σ: 1.44	μ: 3.43	σ: 2.31	*fq* = 21
glad	μ: 7.48	σ: 1.52	μ: 6.49	σ: 2.77	*fq* = 50
address	μ: 5.6	σ: 1.05	μ: 5.62	σ: 2.25	*fq* = 50
Valence			***v* = 5.80**		
Arousal			***a* = 4.98**		

Table 5 shows that the tweet feed dated *Oct 16, 2016 1:26am* highlighted keywords such as "*concern", "enrollment", "Send", "Be", "glad", and "address"* which has an overall valence of 5.80 and an overall mean arousal of 4.98.

Table 6 depicts that the tweet feed dated *Oct 16, 2016 9:23am* highlighted keywords such as "*thank", "will", "Be", "address", "concerns", and "enrollment"* which has an overall valence of 5.84 and an overall mean arousal of 4.12.

Table 7 depicts that the tweet feed dated *Oct 17, 2016 1:29am* highlighted keywords such as "*please", "check", "website", "enrollment", "students", and "follow"* which has an overall valence of 6.19 and an overall mean arousal of 4.67.

4.3 Tweeter Sentiment Visualization

The tweet feeds as a result of the query term *"Enrollment"* were estimated and mapped in a scatterplot diagram to visualize the sentiments.

Table 6. Tweets on enrollment dated Oct 16, 2016 9:23 am

Date/Time *Oct 16, 2016 9:23am*	#OnlineHelpDesk ***Thank*** you for your queries; we ***will be*** around to ***address*** your ***concerns*** throughout ***enrollment***; Monday to Friday; 9am - 5pm				
keywords	**Mean Valence**	**Standard Deviation**	**Mean Arousal**	**Standard Deviation**	**Frequency**
thank	μ: 6.89	σ: 2.29	μ: 4.34	σ: 2.31	*fq* = 6
will	μ: 6.83	σ: 2.04	μ: 2.76	σ: 2.05	*fq* = 19
Be	μ: 6.18	σ: 1.44	μ: 3.43	σ: 2.31	*fq* = 21
address	μ: 5.6	σ: 1.05	μ: 5.62	σ: 2.25	*fq* = 50
concerns	μ: 4.06	σ: 1.56	μ: 5.07	σ: 2.74	*fq* = 50
enrollment	μ: 6.19	σ: 1.94	μ: 3.76	σ: 2.63	*fq* = 23
Valence			***v* = 5.84**		
Arousal			***a* = 4.12**		

Table 7. Tweets on enrollment dated Oct 17, 2016 1:29 am

Date/Time *Oct 17, 2016 1:29am*	***Please check*** LPU ***Website*** (http://t.co/51zfnHyX) for the schedule of 2ns Semester ***enrollment***. ***Students*** must ***follow*** the schedules accordingly.				
keywords	**Mean Valence**	**Standard Deviation**	**Mean Arousal**	**Standard Deviation**	**Frequency**
please	μ: 6.36	σ: 1.68	μ: 5.44	σ: 2.88	*fq* = 50
check	μ: 6.1	σ: 1.53	μ: 6.1	σ: 2.19	*fq* = 50
website	μ: 6.76	σ: 1.64	μ: 3.58	σ: 2.22	*fq* = 22
enrollment	μ: 6.19	σ: 1.94	μ: 3.76	σ: 2.63	*fq* = 23
students	μ: 6.28	σ: 1.83	μ: 5.12	σ: 2.46	*fq* = 50
follow	μ: 5.66	σ: 1.17	μ: 4.1	σ: 2.12	*fq* = 50
Valence			***v* = 6.19**		
Arousal			***a* = 4.67**		

Figure 6 demonstrates the visualization of the tweeter feeds streamed from the corpus of data using the query term *"Enrollment"* as a target of sentiments. It further revealed that the tweets lies on the right side of the scatterplot diagram which indicates that the tweets using the query term *"Enrollment"* are generally pleasant.

Figure 7 demonstrates the general tweeter sentiments of the select institution visualized in an emotional scatter diagram. It depicts that the general estimated sentiment reclined on the positive emotions where majority of the sentiments represented by circles in color green are pleasant [7].

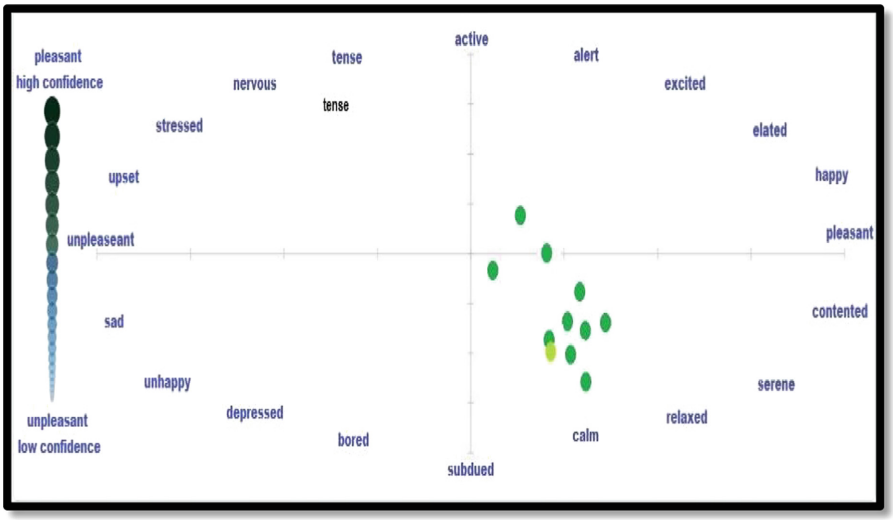

Fig. 6. Tweeter sentiments on query term "Enrollment"

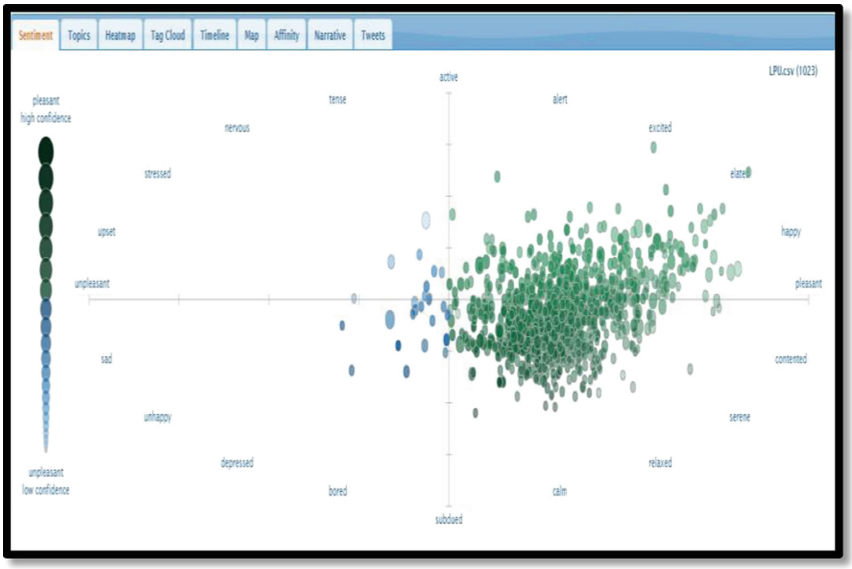

Fig. 7. General sentiments of the select institution

5 Conclusions

Based on the information theoretic review of the study, tweeter feeds can be visualized in a form of charts, graphs and diagrams. Understanding and extracting feelings or emotions from messages or tweeter feeds can also be performed through sentiment analysis. The tf-idf calculated for each query term in tweeter feeds were converted into images using information theoretic approach. With this, the nature of activity and opinions of the users in a select institution were discovered.

In the analysis performed over 1,557 tweets, there was 0.93 mean frequency of tweets posted each day and about 655 user mentions per tweet, 1087 hashtags with 0.70 average number of hashtags, 83 retweets or 5% retweets and 361 replies or 23% replies in the total of analyzed tweets. It showed that the users has more interaction with each other as the numbers calculated increases. As to the query term "Enrollment" used as the target of sentiment, there were 6 tweet feeds analyzed for estimated sentiment. It was visualized that the tweets lie on the right side of the scatterplot diagram which indicates that the tweets as to query term *"Enrollment"* were generally pleasant.

References

1. T. Engineering: Twitter, 30 June 2011. https://blog.twitter.com/2011/200-million-tweets-per-day. Accessed 17 Aug 2016
2. Fornacciari, P., Mordonini, M., Tomaiuolo, M.: A case study for sentiment analysis on Twitter. In: 16th Workshop on From Object to Agents, Naples, Italy (2015)
3. Peng, S., Yang, A., Cao, L., Yu, S., Xie, D.: Social influence modeling using information theory in mobile social networks. Inf. Sci. **379**, 146–159 (2017)
4. Sheela, L.: A review of sentiment analysis in Twitter data using hadoop. Int. J. Database Theory Appl. **9**(1), 77–86 (2016)
5. Bing, L.: Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, Chicago (2012)
6. Healey, C., Hao, L., Hutchinson, S.E.: Visualizations and analysts. In: Cyber Defense and Situational Awareness, p. 329. Springer, New York (2014)
7. Healey, C.: Visualizing Twitter Sentiment, 22 May 2016. https://www.csc.ncsu.edu/faculty/healey/tweet_viz/. Accessed 17 Aug 2016
8. Shannon, C.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423, 623–656 (1948)
9. Cherry, E.C.: A history of the theory of information. In: IEE-Part III: Radio and Communication Engineering (1951)
10. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (2012)
11. Aizawa, A.: An information-theoretic perspective of tf–idf measures. Inf. Process. Manag. **39**(1), 45–65 (2003)
12. Brookes, B.C.: The Shannon model of IR systems. J. Doc. **28**(2), 160–162 (1972)
13. Ebay: A Guide to Buying Voice Changers (2013). http://www.ebay.co.uk/gds/A-Guide-to-Buying-Voice-Changers-/10000000177317588/g.html
14. University of Twente, January 2017. https://www.utwente.nl/cw/theorieenoverzicht/. Accessed 21 Feb 2017

15. Huang, A.: Similarity measures for text document clustering. In: Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand (2008)
16. Robertson, S.: Understanding inverse document frequency: on theoretical arguments for IDF. J. Doc. **60**(5), 503–520 (2004)
17. S. a. J. M. P. ". t. s. o. t.-i. b. c. s. p. A. S. R. 3. (. 7.-1. Tata
18. Bakshy, E., Hofman, J.H., Mason, W.A., Watts, D.J.: Everyone's an influencer: quantifying influence on Twitter. In: Fourth ACM International Conference on Web Search and Data Mining, Hongkong, China (2011)