Journal of Fundamental and Applied Sciences

Research Article

ISSN 1112-9867

Special Issue

Available online at http://www.jfas.info

INTRODUCING PROTOTYPE FOR CLASSIFYING AND QUANTIFYING EMOTIONS IN SOCIAL NETWORK SITES

M. N. F. Jamaluddin[1,*], S. Z. Z. Abidin[2], N. Omar[3], S. S. M. Fauzi[1] and R. A. J. M. Gining[1]

[1]Faculty of Computer and Mathematical Sciences, UniversitiTeknologi MARA, Perlis, Malaysia

[2]AdvancedAnalytics Engineering Center (AAEC)

[3]Faculty of Computer and Mathematical Sciences, UniversitiTeknologi MARA, Selangor, Malaysia

Published online: 30 May 2018

ABSTRACT

Usage of social network sites for dissemination of information has become common nowadays. Easy publication of opinions has led to various kind of emotions expressed in online discussions. In this paper, we introduce a computerized prototype which is able to classify type of emotion and quantify its degree. The prototype developed using JAVA programming language, features automatic comments extraction from social network sites and produces statistical reports on the emotions expressed. It is implemented with Latent Semantic Analysis (LSA) technique which is known to have capability of extraction and comparison of underlying semantic structure between passages. This initial work shows that the prototype can be further enhanced for better accuracy. Understanding public emotions and sentiments may alert the authority and government for ensuring safety among citizens.

Keywords: emotion classification; emotion quantification; latent semantic analysis; information retrieval; textual modality.

Author Correspondence, e-mail: nabilfikri@perlis.uitm.edu.my

doi: http://dx.doi.org/10.4314/jfas.v10i2s.73

1. INTRODUCTION

Social Network Sites (SNS) nowadays attract vast number of users around the world. One of the features in SNS is data sharing where users are able to share their interest, activities, blogging and all kind of multimedia content in various type such as text, images, audios and videos [1-2]. User from various age ranges spent most of their time on SNS[3].Popular features in SNS allows user to update personal information, interact with other users and browse others profiles had made it becoming very attractive[1]. As of June 2017, more than half of Malaysians (61%) had a Facebook profiles [4].

Excessive growth of SNS also creates opportunity for illegal kind of activities such as spamming, profile replication, spread of hatred, fake news and others. Some of these activities might affects the emotion of other SNS users. Manipulated information for the profits and attentions of certain individuals or organizations might triggered unpleasant condition. This is due to the dissemination of information to the public is very simple and cost effective. Taking blog for example, amount of information available on the internet is increasing very quickly due to the process of publication to the internet becoming very easy [5].

Blogs appears in 1995, it emerged as personal diary or journal published on personal website or diary hosting websites[6]. It is estimated that the existence of blog nowadays reached up to 100 million with addition of 50,000 blogs created every 24 hours [7]. Their popularity differs from time to time dependent on current crisis occurs or due to certain events which usually may ranges from information to entertainment [7].On each post, visitors can leave comments and the author can answer the comments [6]. Various kind of emotion might be expressed by users or visitors based on different kind of topics discussed by the blog's author [8]. Fraud news orinformation spread results in problems for controlling the disseminations of information among citizens. Therefore, a mechanism is required for supporting responsible agency to control or at least know informationthat is circulated among citizens.

In computational linguistic, automatic emotion recognition in texts are becoming increasingly crucial from applicative perspective[9]. This includes automatic identification and extraction of opinions, emotions and sentiment in text [10].Motivation for the development of these computerized application of information analysis is for the usage of government, commercial and political domains forobtaining feedback on attitudes and feelings in news and online

forums [11].

This paper introduces a prototype for recognizing emotion in textual modality from comments posted on the blogspot.com and youtube.com sites. The underlying semantic structure of extracted comments are analyzed using Latent Semantic Analysis (LSA). The technique is used to classify emotion expressed by user and to quantify emotions in order to determine its degree.

2. METHODOLOGY

This section discusses methodology for development of prototype which begins with corpus development, Latent Semantic Analysis, Singular Value Decomposition, query vector, similarity calculations and online comments extraction. The development of corpus is based on selected case study of an incident happened to Aminulrasyhid Amzah, a Malaysian teenager boy who was accidentally shot dead by police officer after trying to escape by car[12-13]. The incident shocked the whole nation and mostly debated in social media.

2.1. Corpus Development

The development is carried out by manually collect and annotate user comments, which are obtained directly from comments section on social media sites. Sites selected for this case study are blogspot.com and youtube.com. Two corpuses are developed proposedly for classification and quantification.

For classification purpose, user comments are manually observed and understood to obtain emotion expressed, then annotated and classified into seven documents which encompasses six basic emotions (angry, disgust, sad, fear, surprise and joy) and a document for ‘normal’ which does not convey any emotion. Short forms in Malay words are replaced with actual words, to reduce redundancyduring indexing. Stop words are also removed and followed by stemming process.

For quantification purposes, user comments are analyzed and understood to select instances with different emotion’s intensity. The criteria considered include usage of “strong words”, Malay language intensifiers, capitalized wording, emoticons, repetitious exclamation and question marks. The instances collected are manually annotated according to the intensity of emotion into three documents which are low, medium and high.

2.2. Latent Semantic Analysis

Latent Semantic Analysis is the core technique for this prototype where it is implemented for classification and quantification. LSA includes construction of Term Document Matrix (TDM) purposely for counting number of occurrences of term in documents. Development begins with building a list of unique wordsthat appear in corpus. Each word is analyzed by prototype that includes replacing the short forms, removing stop words and stemming. Finally, based on the number of documents,the prototype generatesTDM by counting number of word appears in each document. Each value in TDM will be given a weight using Term Frequency - Inverse Document Frequency (TF-IDF) algorithm. The purpose is to evaluate importance of a word to the documents. The TF-IDF of term w_{ij} can be defined as:

$$w_{ij} = tf_{ij} \times \log \frac{N}{df_i}$$

where tf_{ij} term frequency (term counts) or number of times a term i occurs in a document, df_i document frequency or number of documents containing term i and N number of documents in corpus.

2.3. Singular Value Decomposition

After TDM matrix is generated and weighted by TF-IDF, Singular Value Decomposition (SVD) is applied. SVD results the matrix decomposed into three other matrices which are U, S and V^T where it can be written as $X = USV^T$. One matrix describes the original row entity, the other describes original column entity and the third is in a form of diagonal values, such that when all three matrices are multiplied, original matrix can be obtained[14]. To implement this, a Java Matrix Package (JAMA) is used for constructing and manipulating real and dense matrix [15]. To calculate SVD matrix,following simple code are used:

```
Matrix X = new Matrix(weighted_tdm);
SingularValueDecompositionsvd = X.svd();
```

to obtain the decomposed matrix:

```
Matrix U = svd.getU();
Matrix S = svd.getS();
Matrix V = svd.getV();
```

From matrix decomposition U, S and V^T, dimension of the solution can be reduced by

simply deleting the coefficient in the diagonal matrix[14]. This means a reduced matrix or an approximate matrix can be written as$\approx X$ with dimension of $k.k$representsdimension of original decomposed matrix. In literature, this procedure is known as dimensionality reduction.

To approximate matrix $\approx X$, the reduced dimension of matrix U, S and $V$are calculated. The matrix U_k and V_k represents coordinate used to represents terms and documents[16]. To obtain approximate matrix of X, following equation are used:

$$X \approx U_k \cdot S_k \cdot V_k$$

2.4. Query Vector

Query vector is used for comparingits similarity to terms and documents. The query vector can be calculated to compare vector similarity in vector space. The similarity of query vector can be calculated against terms and documents. This is because the matrix U and V are representing vector of terms and documents. The query calculation is as follows:

$$q = q^T \cdot U_k \cdot S_k^{-1}$$

Based on the TDM for classification discussed, the query matrix can be generated by analyzing each word in query and comparing to existing terms in TDM. The value of query matrix will increase by one if any of the query term matches the existing term. Calculation of query vector is similar for classification and quantification processes.

2.5. Similarity Calculation

Similarity calculation is used to find relations between query words with terms and documents in the corpus. Thisis done by comparing the query vector with the vector from the matrix generated by SVD namely matrix U for terms and matrix V for documents. Cosine similarity relations are used for the calculations.

Recall the matrix decomposition SVD, $X = USV^T$. Reducing the matrix, we got an approximate matrix, $X \approx U_k \cdot S_k \cdot V_k$. The matrix U_k represents terms vectors while matrix V_k represents document vectors. The query vector q from generated query matrix can be calculated using $q = q^T \cdot U_k \cdot S_k^{-1}$. The similarity between query vector and document vector are calculated using $sim(q = q^T \cdot U_k \cdot S_k^{-1},\ d \cdot U_k \cdot S_k^{-1})$ and similarity between query vector and term vector are calculated using $sim(q = q^T \cdot U_k \cdot S_k^{-1},\ t \cdot U_k \cdot S_k^{-1})$.Based on the value for classification above, cosine similarity relation is as follows:

$$sim(q,d) = \cos A = \left(\frac{q \cdot d}{\|q\|\|d\|}\right)$$

$$\theta = \cos^{-1}\left(\frac{q \cdot d}{\|q\|\|d\|}\right)$$

2.6. Online Comments Extraction

Providing URL to the prototype, it can download and extract comments, commenter username, post title (for blogspot.com) or video title (for youtube.com) from social media sites. To accomplish this task, JSoup version 1.6.1 are utilized. XPATH based technology are used toextract data. The HTML source of a target site are analyzed. Fig.1 shows the code snippet from youtube.com page displaying the location of comment (marked as 1) and username (marked as 2).

Fig.1. Username and comment in youtube.com HTML file

3. RESULTS AND DISCUSSION

This section discusses the output from the development of the prototype and the results obtained from classification and quantification.

3.1. Prototype Overview

Main interface of the prototype is divided into three tabs, which are Initial Processing, Simple Query and Comments Extractor. In Initial Processing tab (Fig.2), Calculate Matrix button performs various text processing beginning with reading all documents in selected corpus and followed by short forms replacement, stop words removal and stemming. Then it generates the TDM, weights the TDM and calculate SVD.

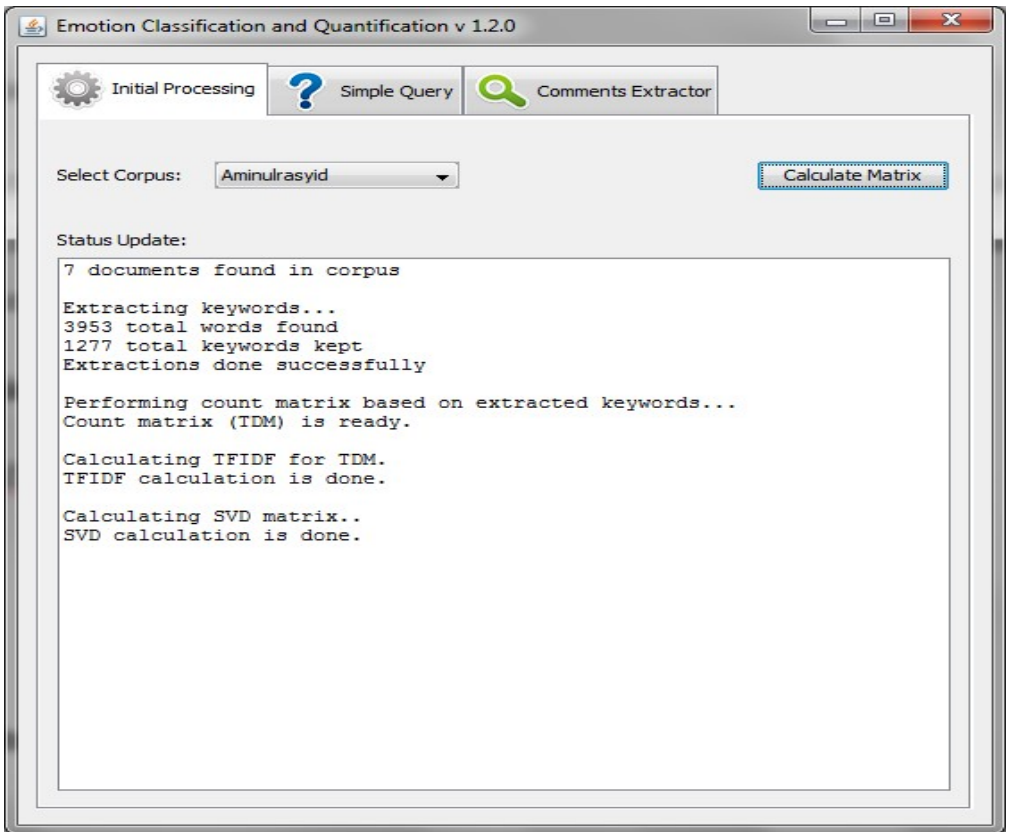

Fig.2. Initial Processing interface

Simple Query tab (Fig.3) allows user to test the calculated SVD. Another four sub tabs in this interface begins with Query Calculation tab used to evaluate words, phrases or sentence for determining emotions and its degree. Value of k that is used in dimensionality reduction can be modified. The prototype start to classify and quantify a given query string, which involves calculating similarity of documents and terms for classifications and documents for quantifications. The calculated similarity for emotions, terms and degree is displayed on the subsequent tabs.

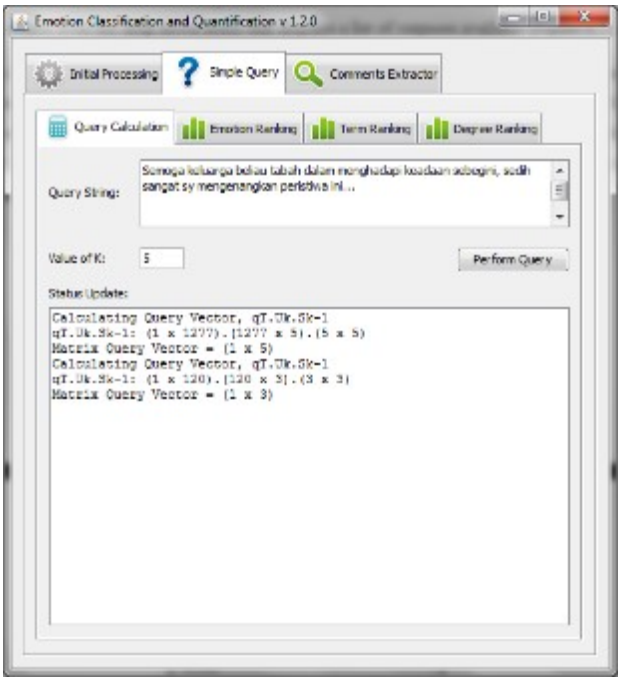

Fig.3. Simple Query interface

Comments Extractor tab consist of three sub tabs which are Social Media Site, Classification Result and Quantification Result. The Social Media Site tab (Fig.4) provide features for extracting comments from SNS. Prototype must be provided with direct URL to blog post of blogspot.com or video of youtube.com. Value of k can be modified to adjust dimensionality reduction.

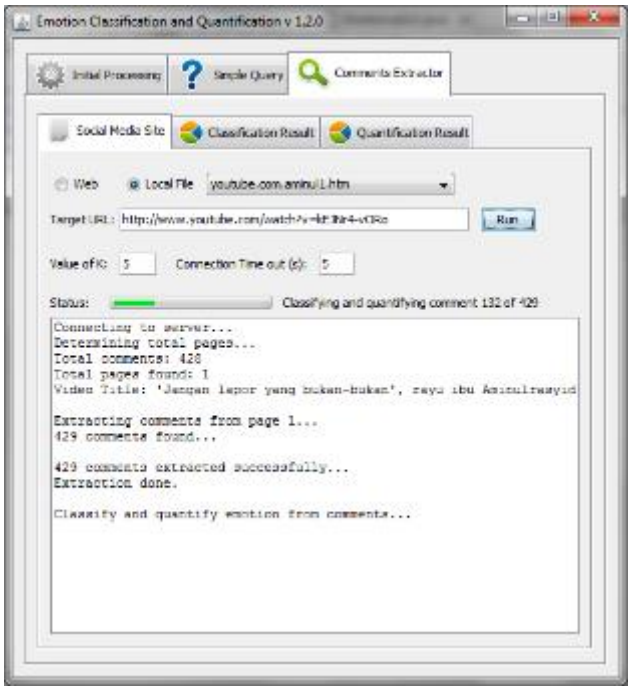

Fig.4. Social Media Site tab

Run button begins by downloading comments, extract information such as commenter's username, page and all necessary details. Each comment will be analyzed by LSA and results will be displayed in subsequent tabs. Classification Result tab (Fig.5) displays counts for each emotion classified and its distribution including percentage. Quantifications Result tab (Fig.6) displays counts for each quantified emotion and its percentage. Detail results including ranks of words and documents for classification and quantification can be accessed from button Detail Results.

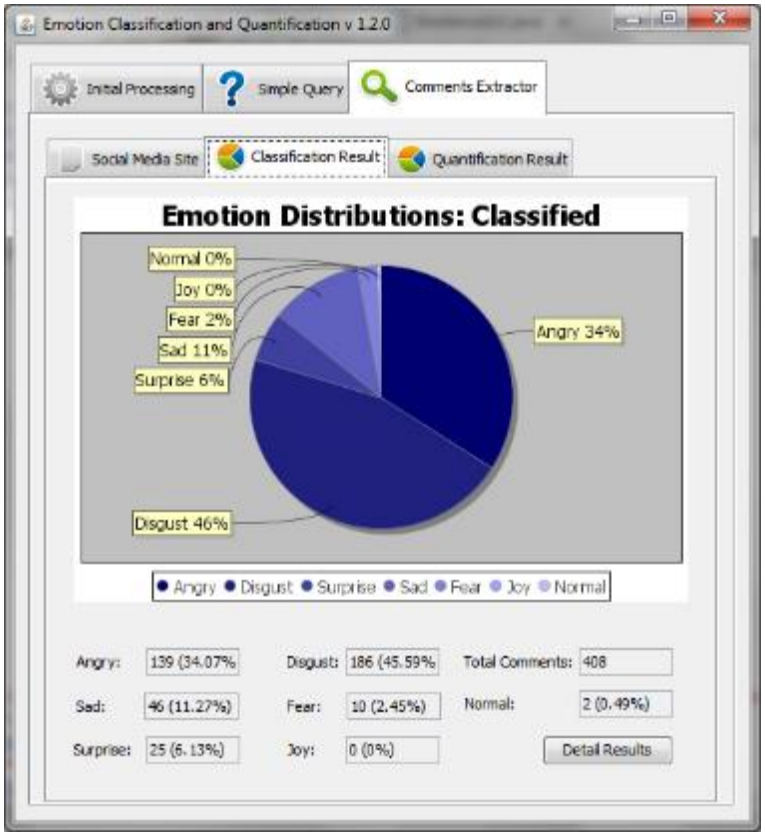

Fig.5. Classification result tab

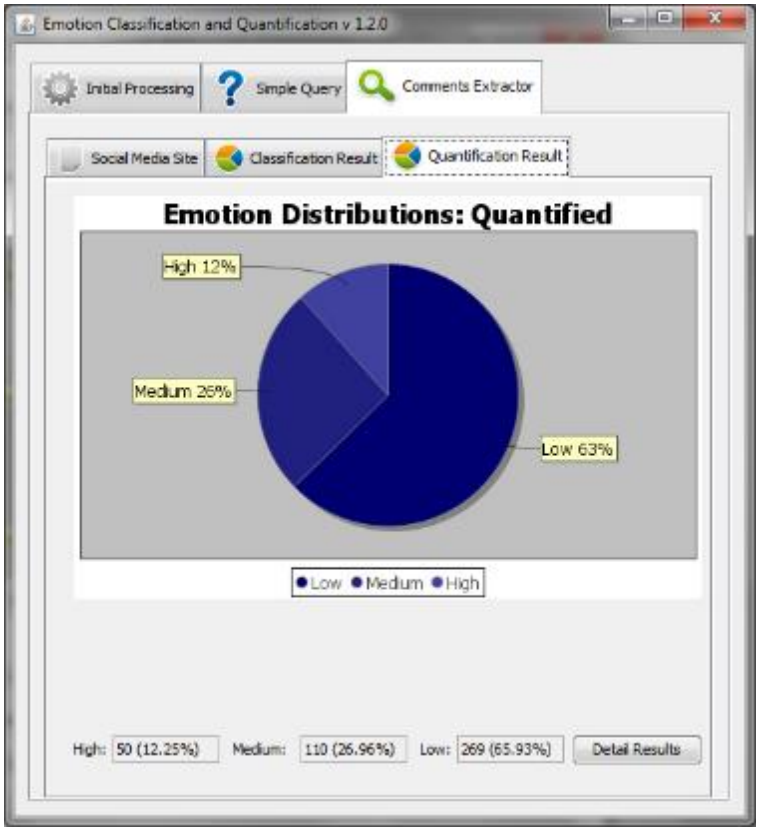

Fig.6. Quantification result tab

3.2. Results and Analysis

In this project, two types of corpuses are built purposely for classification and quantification. Three versions of corpus are prepared for classifications, where each version is an extended version with differences in size which are numbers of words and unique words they have. This can be depicted in Table 1. Accuracy testing is carried out to evaluate accuracy of classification and quantifications between annotation made by human and automatic annotation by prototype. Manual annotation is done by selecting the best sample of comments portraying seven types of emotions and three degrees of emotions and stored in text file. For this purpose, total of 173 samples are collected and annotated manually.

Table 1. Total words and unique words in each version of corpus

Corpus	Total Words	Unique Words
Version 1	3,953	1,277
Version 2	4,788	1,458
Version 3	6,434	1,751

3.2.1. Classification Accuracy

From the collected samples of 173 user comments, classification testing is carried out to determine the capability of the prototype. The corpus version one is used for indexing and preparing the matrix for Latent Semantic Analysis (LSA). Different values of k are tested to obtain best classification accuracy. Result for testing is presented in Fig.7.

The classification accuracy is very low on the reduced matrix with value of k is 2, 3 and 4. This is because the matrix is losing a lot of valuable information from the reduced SVD matrix. Most accurate classification is obtained from the value of k = 5 with 52.0%. This is followed by k value of 6 and 7 with 46.2% and 34.7% classification accuracy respectively. The decreasing accuracy conditions from value of k = 5 to 7 is due to the increase in number of coefficient which leads to the increase of noise in SVD matrix and results in poor classifications. The best dimension reduction for this prototype is achieved when the value of k is 5.

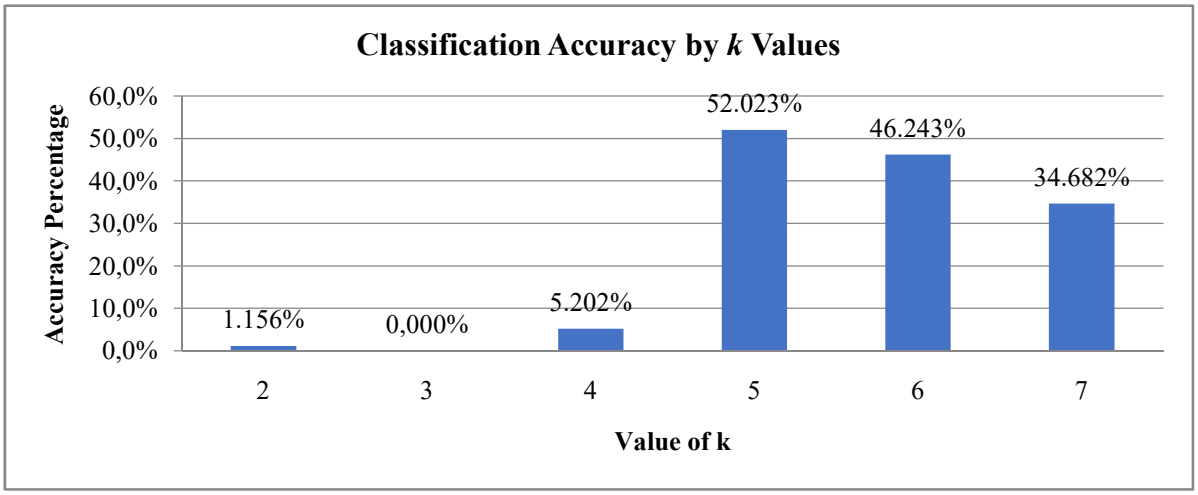

Fig.7. Classification accuracy by the value of matrix reduction, k

3.2.2. Quantification Accuracy

The value of k is set to 3 for quantification in the prototype. Simple observation is carried out to check if the quantification is successful. Testing with the value of $k = 2$ shows poor result and cannot properly recognize degree of emotion. This is due to limited number of documents and implementsmultiple conditional checks. Because ofthislimitation, tests on different value of k cannot be carried out. The quantification accuracy can only be tested on the value of $k = 3$. The quantifications accuracy results in 53.757% or 93 from 173 comments are correctly quantified.

4. CONCLUSION

In this paper, a prototype for classifying and quantifying emotions in social network sites is presented. LSA technique has been successfully implemented for classifying six basic emotions which are angry, disgust, sad, fear, surprise and joy and quantifying three levels of emotion which are low, medium and high. The prototype features comments extractor which capable of extracting comments from popular social network sites. Statistical information is displayed upon the completion of classification and quantification processes, which might give indications on the overall emotions posed by commenters.

Main constraint in developing this project is corpus. Selected case study posed unequal distribution of emotions which leads to difficulty in obtaining samples for emotion joy, surprise and fear. Some selected samples are annotated to single type of emotion even it may have mixed emotions. Manual annotations are done by the judgment and perception of annotators which may be biased and inconsistent.

Considerations for future improvements includes the development of corpus, the distribution of samples should be balanced. Multiple documents can be used to store samples, which might allow bigger range for the value of k to be tested. Artificial Neural Network (ANN) can be implemented at the end of classification and quantification process, where it can determine the emotions and its degree based on calculated similarity degree by the LSA engine. The current method used is by taking the closest emotion document to the query vector. In some cases, the correct emotion document is at the second closest. In this case, the ANN can be used to learn pattern and predict correct emotion by taking values calculated from this prototype.

5. REFERENCES

[1] Rathore S, Sharma P K, Loia V, Jeong Y S, Park J H. Social network security: Issues, challenges, threats, and solutions. Information Sciences, 2017, 421:43-69

[2] Boyd D M, Ellison N B. Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 2007, 13(1):210-230

[3] Ramalingam D, Chinnaiah V. Fake profile detection techniques in large-scale online social networks: A comprehensive review. Computers and Electrical Engineering.

2017,65(3):165–177

[4] Internet World and Statistics.Research, research report on asia marketing internet usage, population statistics and facebook information. 2018, https://www.internetworldstats.com/asia.htm#my

[5] Jiang Z, Lu C. A latent semantic analysis based method of getting the category attribute of words.In International Conference on Electronic Computer Technology, 2009, pp. 141–146

[6] Miura A, Yamashita K. Psychological and social influences on blog writing: An online survey of blog authors in Japan. Journal of Computer-Mediated Communication, 2007, 12(4):1452-1471

[7] Macias W, Hilyard K, Freimuth V. Blog functions as risk and crisis communication during Hurricane Katrina. Journal of Computer-Mediated Communication, 2009, 15(1):1-31

[8] Chen J, Liu Y, Zou M. User emotion for modeling retweeting behaviors. Neural Networks, 2017, 96:11-21

[9] Strapparava C, Mihalcea R. Learning to identify emotions in text. In ACM Symposium on Applied Computing, 2008, pp. 1556-1560

[10] Ghazi D, Inkpen D, Szpakowicz S. Hierarchical approach to emotion recognition and classification in texts. InCanadian Conference on Artificial Intelligence, 2010, pp. 40-50

[11] Wiebe J, Wilson T, Cardie C. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 2005, 39(2-3):165-210

[12] Utusan Online. Remaja mati ditembak polis: Sepuluh saksi utama diambil keterangan. 2010, http://ww1.utusan.com.my/utusan/info.asp?y=2010&dt=0429&pub=Utusan_Malaysia&sec=Jenayah&pg=je_01.htm

[13] Wikipedia. Aminulrasyid Amzah. 2018, https://en.wikipedia.org/wiki/Aminulrasyid_Amzah

[14] Landauer T K, Foltz P W, Laham D. An introduction to latent semantic analysis. Discourse Processes, 1998, 25(2-3):259-284

[15] Hicklin J, Moler C, Webb P, Boisvert R F, Miller B, Pozo R, Remington K.JAMA:A Java matrix package. 2012, http://math.nist.gov/javanumerics/jama/

[16] Deerwester S, Dumais S T, Furnas G W, Landauer T K, Harshman R. Indexing by latent

semantic analysis. Journal of the American Society for Information Science, 1990, 41(6):391-407

How to cite this article:
Jamaluddin M. N. F., Abidin S. Z. Z.,Omar N., Fauzi S. S. M. and Gining R. A. JM.Introducing Prototype for Classifying and Quantifying Emotions in Social Network Sites. J. Fundam. Appl. Sci., 2018, *10(2S), 999-1012.*