Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 159 (2019) 30–37

Procedia Computer Science

www.elsevier.com/locate/procedia

23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

String-based Multinomial Naïve Bayes for Emotion Detection among Facebook Diabetes Community

Vimala Balakrishnan[a,*] and Wandeep Kaur[b]

[a,b] *Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, 50603, Malaysia*

Abstract

This paper determined the emotions among the online diabetes community using the string-based Multinomial Naïve Bayes algorithm. Facebook posts from official diabetes support groups were crawled, for a total of 15 000 pre-processed posts. Of these, 800 were manually annotated by human experts. The posts were first classified according to Plutchik's wheel of emotions, comprising of eight dominant emotions: anger, sadness, fear, joy, surprised, trust, anticipation and disgust using the NRC Emotion Lexicon (Emolex). The emotion classifications were then refined using string-based Multinomial Naïve Bayes algorithm, with results indicating a 6.3% improvement (i.e. 82% vs. 75.7% for average F-score) when compared to the Emolex-approach, and other machine learning algorithms, namely, Naïve Bayes and Multinomial Naïve Bayes. The higher accuracy in emotion classification reflects the feasibility of our approach. Further analysis also revealed emotions such as joy, fear and sadness to be of the highest frequencies for the diabetes community.

Keywords: Emotion detection; diabetes; Facebook; string vector; Multinomial Naïve Bayes

* Corresponding author. Tel.: +6-03-79676377;
E-mail address: vimala.balakrishnan@um.edu.my

1877-0509
10.1016/j.procs.2019.09.157

1. Introduction

Emotion Detection (ED) is an opinion mining task concerning the computational study of natural language expressions in recognizing various emotions from text, such as joy, trust, fear, surprise, and sadness, among others [15;17]. It has proven to have wide applications ranging from building nuanced virtual assistants that cater for the emotions of their users to detecting the emotions of social media users in order to understand their mental and/or physical health.

ED is often performed using two main techniques including Machine Learning (ML), and lexicon-based approach. The lexicon-based approach involves the use of dictionaries that have mapped words to specific emotions, such as SentiWordNet. On the other hand, the ML approach extracts features using a criterion or a combination of criteria, however current feature selection methods pose concerns like huge dimensionality (i.e. high number of features being extracted) and sparse distribution (i.e. features with a limited coverage in a given corpus). Both issues result in inaccurate text classification [13]. The current study proposes to improve the Multinomial Naïve Bayes algorithm (MNB) using a string vector. A similarity matrix is first built from the corpus, and MNB is later modified into string-based version for classification purpose.

The corpus selected for this study is based on Facebook posts focusing on the diabetes community. Diabetes is one of the largest global health concerns with almost 451 million people (age 18-99 years) diagnosed with the disease worldwide, and this number is expected to increase to 693 million by 2040 [16]. Social networks such as Facebook have become an excellent resource for the community as it helps to build a bridge to connect different people who have a similar condition and similar experiences. The information shared by the community is wrapped in their own sentiments and emotions, which is the driving force of ED.

The remainder of the paper is structured as follows: Section 2 presents the background on ED, and ML algorithms. This is followed by the research methodology, which encompasses the corpus, ED approach and the evaluation of the proposed approach. Results and discussion are presented in Section 4.

2. Background

Recent studies have defined emotion as a quintuple (e, a, m, f, t) where (e) represents target of the entity, (a) refers to target aspect of (e) that is accountable for the emotion, followed by (m) which is the type of emotion and its intensity, (f) is the entity that feels the emotion and (t) represents time the emotion was expressed [4;14]. For example, in the text below:

"I am so thrilled with the football team captain today"

(e) is represented by the football team, (a) is the team captain, the emotion (m) is joy, and the use of the word thrilled shows its intensity is greater than happy. The person who experienced the emotion (f) is the author of the statement and (t) is depicted as today.

ED detects and recognizes types of feelings through the expression of texts to determine how happy people are pertaining to different factors such as environmental, health, economic or social. With the advent of social media, people turn to platforms such as Facebook and Twitter to express their views, opinions, suggestions and also frustrations. This provides a colossal of opinion that can be mined, and their emotions detected. A detection system that can extract factual information, sentiments and emotions from text creates a wealth of opportunities for organizations and individuals in advising or predicting customer purchases, for example. Although patient emotion has been shown to be important in healthcare [2;3], studies targeting users' online communication in healthcare is none existent.

As stated in Section 1, text classification documents are encoded in numerical vectors which cause problems such as huge dimensionality and sparse distribution. Numerous techniques have been adopted to address these issues, such as, Al-Anzi et al. [1] who used a singular value decomposition method to extract features based on latent semantic indexing (LSI), with results showing LSI to be a better textual representation technique as it maintains semantic information between words. Authors in [5] presented a number of dimensionality reduction techniques such as root-based stemming, light stemming and singular value decomposition (SVD) to handle large feature sets for text classification, whereas FOREX market exchange from news headlines was predicted using a multi-layer dimension reduction algorithm based on semantics and sentiment [10].

In another study, the researchers used string vector with k-nearest neighbour (KNN) algorithm with results showing an improved accuracy of text classification by 5% compared to using the traditional method of numerical vectors [6]. Further, the researchers extended their work by modifying the Agglomerate Hierarchical Clustering (AHC) algorithm, with an improved classification, and thus showing that converting numerical vectors to string vectors improves dimensionality reduction [7].

Nevertheless, the implementation of string vectors instead of numerical vectors as an effort for dimensionality reduction seems to be limited. The most related work at implementing string vectors look into adopting the neural network models which are not only computationally expensive to implement but also require a very large amount of training data [8]. Therefore, this study looks into improving the emotion classification using string vectors and a traditional machine learning algorithm, that is, Multinomial Naïve Bayes (MNB).

3. Methodology

Fig. 1 depicts the emotion detection pipeline, comprising of all the tasks involved.

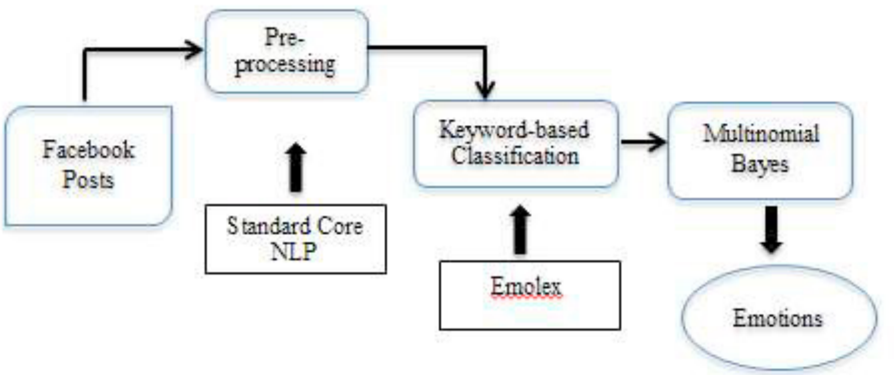

Fig. 1. Emotion Detection Pipeline

3.1. Diabetes corpus

Facebook posts from six diabetes related groups were extracted for a duration of six months (July 2016 - January 2017) using Facebook Graph API. The data were extracted for all three diabetes types, that is, Type 1, 2 and 3. Pre-processing included the removal of spams, emojis, misspelled words, etc. resulting in a total of 82 120 posts. Of these, approximately 15 000 posts were selected for further Part of Speech (POS) tagging, tokenization, stop word removal and stemming, using the Standard Core NLP parser. Out of these, 6 000 posts were randomly annotated by seven annotators comprising of medical practitioners and linguists, with an inter-coder reliability of 89%.

3.2. ED approach

The keyword-based emotion filter proposed identifies posts that contain some form of emotion using keywords extracted from the said posts. The NRC Emotion Lexicon (Emolex) consisting of 14,181 words with eight basic emotions (i.e. anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) associated with each word in the lexicon, was used as the emotion lexicon [9].

Once the emotions have been identified, the post is then passed through the MNB algorithm that converts each post into a string of vectors to map the words around the word identified as an emotion using bag of words concept. MNB is an extension of Naïve Bayes intended to solve text classification problems, and it adopts multinomial distribution with the number of occurrences of a word within a text. The MNB equation is as shown below:

$$P\ (X \mid c) = \log \frac{N_c}{N} + \sum_{i=1}^{n} log \frac{t_i + \alpha}{\sum_{i=1}^{n} t_{i+\alpha}}$$

Where

$P\ (X \mid c)$ = probability of document X in class c

N_c = total documents in class c
N = total documents
t_i = weight term t

$\sum_{i=1}^{n} t$ = total term weight t in class c

α = smoothing parameter

The following example illustrates how the ED differs between the two approaches:

I am so grateful for the night nurses at Mayo Clinic for their service. Without their hard work, support and patience, my son would not have made it through. You not only saved my son but also my life!

Table 1 shows the ED classifications for the above-post, and the keywords associated to determine the classification. Both techniques identify the keywords associated to an emotion, but the refined MNB specifically identifies the emotion with the strongest emotion intensity, which is *trust* in the given example.

Table 1: Emotion classification for Emolex versus MNB-string

Model	Trust	Surprise	Sadness	Joy	Fear	Disgust	Anger	Anticipation
Emolex	1 *(support, patience, saved)*	1 *(saved)*	0	1 *(grateful, life)*	0	0	0	1 *(service)*
MNB-string	1 *(support, patience, saved)*	0	0	0	0	0	0	0

3.3. Evaluation

A total of 800 posts were randomly selected and used for the purpose of evaluation. The study used two metrics, namely, F-measure, which is the weighted average of precision and recall. Precision refers to the number of false positives (i.e. how many were correctly classified as per human annotations), whereas recall measures the number of false negatives (i.e. how many emotions were correctly found by the classifier from the pool of annotations). F-measure therefore, is a measure that places equal emphasis on both optimization objectives. The second metric used was the overall accuracy. A higher score for both the metrics indicates the effectiveness of the ED classification.

The proposed ED model was compared with several other approaches and algorithms. To be specific, the proposed string-based MNB was compared against Emolex (i.e. lexicon based approach), and two other ML algorithms, namely Naïve Bayes and Multinomial Naïve Bayes. The two latter ML algorithms were selected as they represent probabilistic classification models as well.

4. Results and Discussion

Table 2 shows the accuracy and F-measure scores for all the compared models. In general, it can be observed that the effectiveness of the ED classification improved when the string vectors were used, regardless of the emotions detected. This is probably due to the hierarchical classification employed in the study whereby an initial ED was performed using Emolex, which is then refined using a second level of classification using string-based MNB. The improvement is also credited to the reduced dimensionality that helps to match attributes within the text by removing attributes that are otherwise considered useless. For example, in this study, the focus was on words that can be associated with emotions, hence attributes like “appear, pharmaceuticals etc.” were dropped by the algorithm in order to reduce overfitting and produce more accurate results.

Table 2: Emotion detection effectiveness

Emotion	Emolex		Naïve Bayes		MNB		MNB-string	
	Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure
Trust	0.615	0.762	0.611	0.706	0.623	0.776	0.805	0.884
Surprise	0.520	0.684	0.538	0.682	0.552	0.754	0.752	0.854
Sadness	0.572	0.728	0.507	0.601	0.557	0.641	0.605	0.744
Joy	0.748	0.856	0.701	0.725	0.687	0.762	0.820	0.870
Fear	0.827	0.905	0.796	0.811	0.803	0.817	0.891	0.870
Disgust	0.679	0.809	0.569	0.675	0.569	0.790	0.796	0.849
Anger	0.558	0.716	0.601	0.677	0.556	0.669	0.656	0.771
Anticipation	0.430	0.602	0.501	0.571	0.513	0.582	0.595	0.740
Average	***0.62***	***0.76***	***0.60***	***0.68***	***0.61***	***0.72***	***0.74***	***0.82***

In most cases, the MNB algorithm estimates the probability of a word occurring taking into account the number of times that particular word appears in the document. In other words, an increase in occurrence adds weightage to the word. In the context of this experimentation, when words related to a particular emotion appears multiple times within a sentence, it is naïvely accepted as the overall emotion of said text. For example,

*“I am **enraged** that the nurses at school took so long to give my son his EpiPen and the administration are **disregarding** it like it was nothing!! As a mother I am deeply **disappointed**, **hurt** and **angry**”*

The above shows multiple words mapped to the emotion anger (enraged, disregarding, angry). The MNB algorithm works by using those multiple terms and assumes it to have a feature vector that causes it to have large dimensionality issues. This in the long run causes the accuracy to dwindle specifically when it comes to classifying longer text such as text extracted from social media sites such as Facebook. Therefore, by converting those vectors into strings, the detection effectiveness has been improved for the string-based MNB compared to Naïve Bayes and MNB.

A further analysis of the specific emotions based on the diabetes dataset shows the highest accuracy was recorded for fear (i.e. 0.891) whereas the lowest was for anticipation (i.e. 0.595). Also, a count of posts classified based on the emotions indicate the top positive emotion to be joy, whereas top three negative emotions were fear, sadness and anticipation (Fig. 2). Figures 3 and 4 show the text visualization for fear and joy, respectively. Looking at Fig. 3, it is clear that major concerns are related to "sugar, blood, diabetes and eat" whereas for joy, popular keywords seem to be "carbs, good, diabetes."

Overall, the illustration shows that most of the communication was laden with negative emotion, indicating that this particular group is highly concerned regarding the issue on hand, that is, diabetes. The lowest count for surprise could be due to the fact that the emotion can be described to be "neutral", that is, it can be used to reflect a positive surprise, or a negative surprise, and thus the performance of the ED could have been affected. For example,

"*My latest numbers truly shocked me. I knew the treatment was working but this well! Everyone should give Chinese herbs a try!*" – A positive surprise

"*I am in awe of the way the pharma's are pretending like we are guinea pigs for them to try new drugs on!!*" – A negative surprise

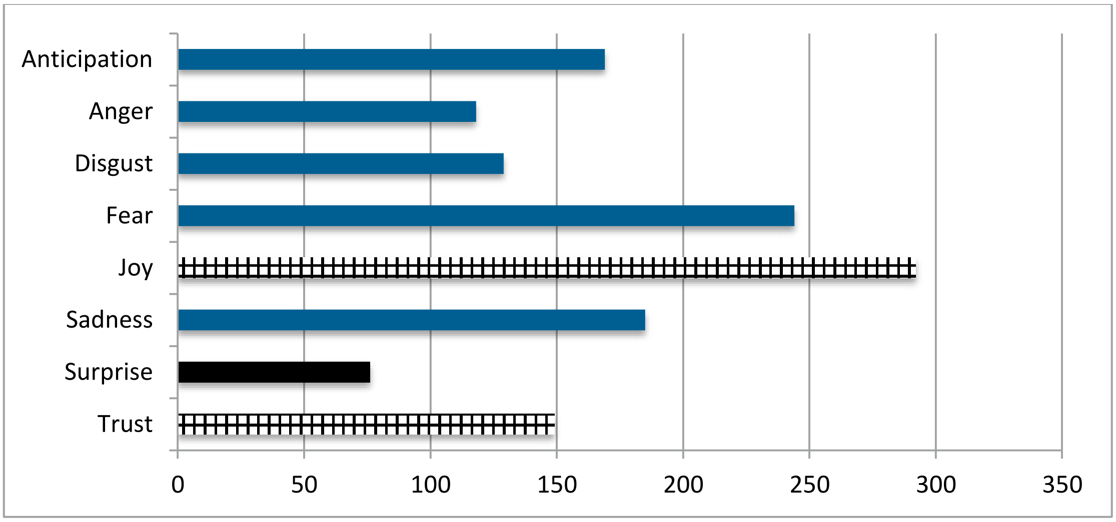

Fig. 2. Emotion frequencies

Fig. 3. Text visualization for fear

Fig. 4. Text visualization for joy

5. Conclusion

The study analysed human emotions based on online communication among diabetes community using a string-vector based MNB. Results generally indicate the proposed method to outperform all the compared models, that is, Emolex, Naïve Bayes and Multinomial Naïve Bayes with an average accuracy of 74% (versus 62% for Emolex; 60% for Naïve Bayes; 61% for MNB) and an average F-score of 82% (versus 76% for Emolex; 68% for Naïve Bayes; 72% for MNB). Additionally, the majority of the posts were found to be joyful in nature, followed by those categorized as fearful and sadness. The findings of the study show that emotion detection using a string-based vector is better than solely depending on the numerical vectors.

Acknowledgements

The authors would like to express their gratitude to University of Malaya Grant [RP059C-17SBS] for supporting this study, and also to the human experts who assisted in manually annotating the dataset.

References

[1] Al-Anzi, F. S., & AbuZeina, D. (2017). "Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing." *Journal of King Saud University-Computer and Information Sciences, **29(2),** 189-195.*

[2] Bickmore, T., & Giorgino, T. (2006). "Health dialog systems for patients and consumers." *Journal of Biomedical Informatics, **39(5),** 556-571.*

[3] Blanch-Hartigan, D. (2013). "Patient satisfaction with physician errors in detecting and identifying patient emotion cues." *Patient Education and Counseling, 93, 56–62*

[4] Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). "A practical guide to sentiment analysis." *(Vol. 5): Springer*

[5] Harrag, F., & Al-Qawasmah, E. (2010). "Improving Arabic Text Categorization Using Neural Network with SVD." *JDIM, 8(4), 233-239.*

[6] Jo, T. (2017a). "Long text segmentation by string vector based KNN." *Paper presented at the 19th International Conference on.Advanced Communication Technology (ICACT), 2017 (pp. 805-810). IEEE*

[7] Jo, T. (2017b). "String vector based AHC for text clustering." *Paper presented at the 19th International Conference on Advanced Communication Technology (ICACT), 2017 (pp. 673-678). IEEE*

[8] Liu, Y., Bi, J.-W., & Fan, Z.-P. (2017). "Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms." *Expert Systems with Applications, 80, 323-339*

[9] Mohammad, S. M., & Turney, P. D. (2013). "Crowdsourcing a word–emotion association lexicon." *Computational Intelligence, **29(3),** 436-465.*

[10] Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2015). "Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment." *Expert Systems with Applications, **42(1),** 306-324.*

[11] Oeldorf-Hirsch, A., & Sundar, S. S. (2015). "Posting, commenting, and tagging: Effects of sharing news stories on Facebook." *Computers in Human Behavior, 44, 240-249.*

[12] Plutchik, R. (2003). "Emotions and life: Perspectives from psychology, biology, and evolution", *American Psychological Association*.

[13] Shalin, L. V. A., & Prasadh, K. (2016)." "An Effective Multi-clustering Anonymization Approach Using Discrete Component Task for Non Binary High Dimensional Data Spaces." *Procedia Technology, 25, 208-215.*

[14] Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). "Current state of text sentiment analysis from opinion to emotion mining." *ACM Computing Surveys (CSUR), **50(2),** 25.*

[15] Medhat, W., Hassan, A., & Korashy, H. (2014). "Sentiment analysis algorithms and applications: A survey." *Ain Shams Engineering Journal, **5(4),** 1093-1113.*

[16] Ogurtsova, K., da Rocha Fernandes, J., Huang, Y., Linnenkamp, U., Guariguata, L., Cho, N., . . . Makaroff, L. (2017). "IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040." *Diabetes research and clinical practice, 128, 40-50.*

[17] Tromp, E. E., & Pechenizkiy, M. M. (2014). "Rule-based emotion detection on social media : putting tweets on Plutchik's wheel." *Retrieved from http://repository.tue.nl/802156*