International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT206658

Users' Emotions Analysis based on Hybrid Feature Extraction Techniques

Sulis Sandiwarno
Department of Computer Science, Universitas Mercu BuanaUniversity, Indonesia
sulis.sandiwarno@mercubuana.ac.id

Article Info
Volume 6, Issue 6
Page Number: 291-296
Publication Issue :
November-December-2020

Article History
Accepted : 10 Dec 2020
Published : 24 Dec 2020

ABSTRACT

In order to solve some problems of importance of words and missing relations of semantic between words in the emotional analysis of e-learning systems, the TF-IWF algorithm weighted Word2vec algorithm model was proposed as a feature extraction algorithm. Moreover, to support this study, we employ Multinomial Naïve Bayes (MNB) to obtain more accurate results. There are three mainly steps, firstly, TF-IWF is employed used to compute the weight of word. Second, Word2vec algorithm is adopted to compute the vector of words, Third, we concatenate first and second steps. Finally, the users' opinions data is trained and classified through several machine learning classifiers especially MNB classifier. The experimental results indicate that the proposed method outperformed against previous approaches in terms of precision, recall, F-Score, and accuracy.

Keywords : TF-IWF, MNB, Word2Vec, emotions, e-learning

I. INTRODUCTION

Nowadays, text classification is widely seen as a task of supervised learning concept which is defined as the categories' identification of new documents [1, 2]. As the amount of available new documents textual information available online increases, managing to classify them properly becomes more difficult. This is due to the ability to effectively retrieve the correct categories for new documents relies heavily upon the amount of labelled documents already available for reference. Traditional document representation involves classification adopting techniques of information retrieval such as Term Frequency Inverse Document Frequency (TF-IDF) algorithm that have been widely employed in Natural Language Processing (NLP). Where, these techniques can help in providing a simplified documents representation via various features. TF-IDF by reflecting the word importance to a particular document in a documents' collection [3]. Whereas, Word2Vec algorithm is an efficient tool that introduced by Google to represent the word as a real value vector in 2013 [4]. By adopting deep learnings' idea, Word2Vec algorithm can be simplified text content processing into K-dimensional vector space calculation and for representing the semantics similarity the text.

The application of information technology today is the focus of life that leads to the advancement of

science education. The development of information technology is supported by the lives of everyone who wants to use this information technology to help solve existing problems [1, 2, 5]. Information technology development in the world of health has a very big role in everyday life in solving existing problems.

II. RELATED WORK

Text emotional analysis is also often mentioned to as opinion or emotional minings, that refers to the subjective objectivity mining and analysis, views, emotions and polarity of the text via the computing technology [6]. The previous approaches have been widely used machine learning classifiers to analyze the polarity of users' opinions in education field such as e-learning systems [7–9]. On other hand, in analyzing users' emotions in e-learning systems, the aforementioned has been adopted TF-IDF or Wrod2Vec algorithms as feature extraction techniques [9, 10]. Although, the previous approaches obtained good results in analyzing users' emotions by employing machine learning classifiers based on TF-IDF algorithm, but there are also have some problems. If there is a text large amount of the same type in the users' opinions corpus, the weight of this type of keywords will become lower. Therefore, in order to tackle this problem, we are adopting Term Frequency Inverse Word Frequency (TF-IWF) algorithm [11].

III. PROPOSED MODEL

Figure 1 shows the process of building a user model, which includes text preprocessing, feature extraction, and classification.

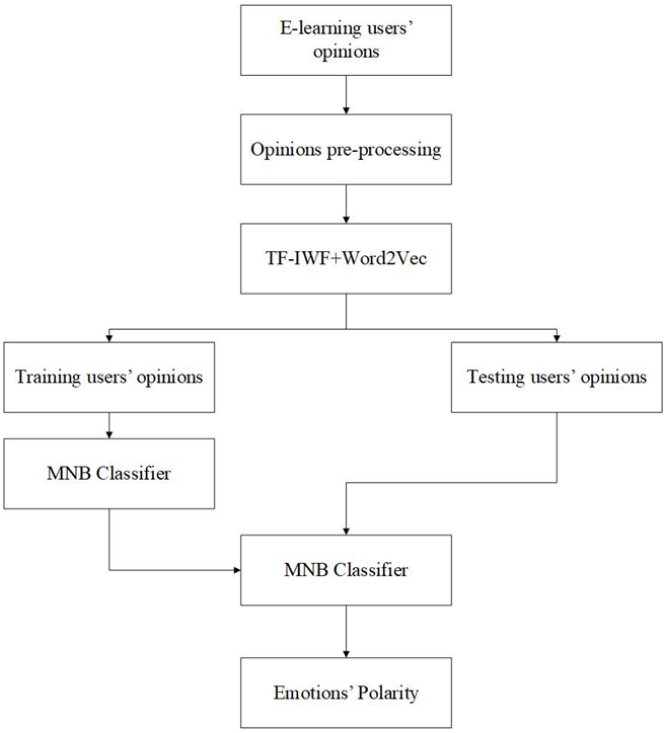

Figure 2. The framework of Proposed Model

3.1 Opinions pre-processing

This section removes some identities from a text, where the identity is HTML decoding, remove stop words, and remove bad characters in opinion.

3.2 Proposed Model

3.2.1 TF-IWF

Term Frequency Inverse Word Frequency (TF-IWF) algorithm is the improvement of TF-IDF algorithm due to TF-IDF ignores two importance keys TF (Term Frequency) and IWF (Inverse World Frequency) [12]. The mainly idea of TF-IWF algorithm is that a word has a high occurrences number in a single text and the occurrences number of that word in the general corpus is lower than that of all words in the corpus. Then, TF-IWF algorithm is considered that it has a good classifications' function and can be adopted for text classification. Generally, the basic of TF-IWF depicted as:

$$TF = \frac{N}{\sum_l N_{l,m}} (1)$$

where, N is defined as the occurrences number of a word wi in the users' opinions document ND, $\sum_l N_{l,m}$ represents the sum of occurrences number of all word wi in the ND.

$$IWF = log\ \frac{\sum_{i=1}^{p} N_c}{N_c} \quad (2)$$

where, $\sum_{i=1}^{p} N$ is depicted as the sum of occurrences number of all words is the users' opinions corpus and the N_c is the occurrences number of a word wi c in the users' opinions corpus. Then, TF-IWF algorithm is defined as:

$$TF - IWF = TF\ x\ IWF \quad (3)$$

3.2.2 Word2Vec

In the NLP, computers need to deal with a large natural language amount, due to the computers cannot be read like humans, they need for the convert natural language into digital language. Unfortunately, even after the computer "understands" natural language, it still cannot understand its semantic relations. The core idea of Word2vec algorithm is mapping the sentences composed by words that the computer think are duplicated to each other into a higher-dimensional matrix, and replace the relations of semantic between words with the mathematical relations in the matrix [13].

3.2.3 TF-IWF+Word2Vec

Suppose the word wi in the dictionary sentiment is ds, the vector of ds is vocab voc, and users' opinions is ND, = <w1, w2, ... ,Wn> :

$$voc = \{w_i XiE1 \ldots . Nv\} \quad (4)$$

where, Nv is defined as the dimension of vector of the word wi.

1. First, using TF-IWF algorithm to weight the word in the users' opinions corpus to achieve the word weight
2. Second, employing the Skip-gram algorithm to train the users' opinions corpus, and compute the vector of each word, compute the vector of each word in the users' opinions document ND, to obtain the vector of sentence of the users' opinions $Svec(ND)$;

$$vec(ND) = \sum_w Word2Vec(w_i), w_i \in ND \quad (5)$$

3. By multiplying the word weight by its vector of corresponding, then the vector of weighted word can be achieved, compute the weighed vector of each word wi in the users' opinions document ND to obtain the weighted vector of sentence of the e-learning users' opinions $W_{sen}(ND)$

$$W_{sen}(ND) = step1\ x\ step2 \quad (6)$$

3.3 Multinomial Naïve Bayes (MNB)

Given the test description of the document d of an opinion represented by the vector $< w_1, w_2, \ldots\ldots, w_m >$, to classify the document d, MNB is defined as:

$$C_{MNB^{(d)}} = P(c) \prod_{i=1}^{n} P(w_i|c)^{fi} \quad (7)$$

where, $P(c)$ is a prior probability that a document d belongs to class c, n is a number of the features, $P(w_i|c)$ is the conditional probability that a word w_i occurring in the class c, w_i is the word feature occurred in d, fi is the number of frequency count of a word w_i in reporting d, and $C_{MNB^{(d)}}$ is the class label of d predicted by the classifier [14].

IV. EXPERIMENTAL AND RESULTS

The main content of this section are divided into three categories: (1) data source of users' opinions, (2) evaluations metrics, and (3) classification results.

4.1 Data sources

In this study, we are crawling users' opinions data from social media such as Twitter and Facebook. Where, we obtained 80.000 users' opinions for each dataset and set the users' opinions 20% for training and 80% for testing. The distribution of users' opinions dataset is shown in Table 1.

TABLE 1. THE DISTRIBUTION OF USERS' OPINIONS DATASET

Dataset Name	Amount	Percentage (%)
Twitter	40.000	50
Facebook	40.000	50
Total	80.000	100

4.2 Evaluation Metrics

In supporting analysis our study, we employ several evaluation metrics such as precision (pre), Recall (rec), and F1 (F1), and Accuracy (acc). The precision is defined as the correct result proportion in all the results; the recall represents the proportion of the data amount with results of classification in all the opinions dataset. Suppose there are A data of the original users' opinions dataset, and Nt results are classified, among which n results are correctly classified. The formula for pre and rec are as follows:

$$pre = \frac{n}{Nt} \quad (8)$$

$$rec = \frac{N}{M} \quad (9)$$

$$F1 = \frac{2pre\ x\ rec}{pre\ x\ rec} \quad (10)$$

4.3 Evaluation Metrics

In this experiment, the original TF-IWF and Word2vec algorithms and the improved TF-IWF weighted Word2vec algorithms were employed to classify users' emotions in order to performance test of the new method. The results are shown in Table 2 and Table 3.

TABLE 2. THE EXPERIMENTAL RESULTS OF ANALYZING USERS' EMOTIONS (DATASET – 1)

Features	Polarity	pre	rec	F1	acc
TF-IDF	Positive	0.692	0.689	0.688	0.698
	Negative	0.614	0.621	0.612	0.611
TF-IWF	Positive	0.732	0.711	0.732	0.734
	Negative	0.681	0.677	0.689	0.688
Word2Vec	Positive	0.787	0.781	0.778	0.792
	Negative	0.732	0.722	0.733	0.752
Proposed	Positive	0.846	0.852	0.845	0.856
	Negative	0.782	0.777	0.781	0.788

It can be seen from the experimental results that the acc of the TF-IDF algorithm is only 0.698 (Dataset – 1) and 0.736 (Dataset – 2), which is the worst performance since it ignores the relationship of semantic between words, the Word2vec algorithm takes into account the words semantics, but it ignores the word importance, so it also obtained the poorly results. The proposed method outperforms in emotion classification, due to which considers both the words importance and the relations of semantic between words.

TABLE 3. THE EXPERIMENTAL RESULTS OF ANALYZING USERS' EMOTIONS (DATASET – 2)

Features	Polarity	pre	rec	F1	acc
TF-IDF	Positive	0.692	0.689	0.688	0.698
	Negative	0.614	0.621	0.612	0.611
TF-IWF	Positive	0.732	0.711	0.732	0.734
	Negative	0.681	0.677	0.689	0.688
Word2Vec	Positive	0.787	0.781	0.778	0.792
	Negative	0.732	0.722	0.733	0.752
Proposed	Positive	0.846	0.852	0.845	0.856
	Negative	0.782	0.777	0.781	0.788

Based on the research results applied on two users' opinions dataset, it was concluded that 4.05% (Dataset - 1) and 2.68% (Dataset - 2) had no emotions in the users' opinions. From the two datasets used, it can be concluded that Dataset-2 is better than Dataset-1 which is seen from the number of positive emotions that are 69.65% and 62.74% as shown in Figure 2.

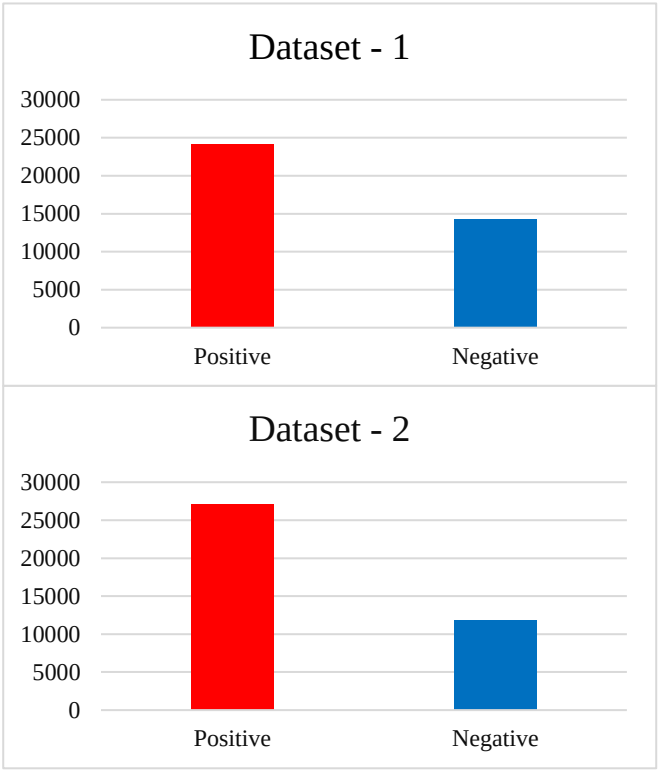

Figure 2. The emotions classification results

V. CONCLUSION

This paper analyzes the advantages of the TF-IWF algorithm, explains the principle of the Word2vec algorithm, and implements in e-learning systems users based on the TF-IWF weighted Word2vec algorithms. Which are, this proposed method aim solves the problem of ignoring the words importance and missing relations of semantic between words in the users' opinions dataset classification.

VI. REFERENCES

[1]. Sadikin M, Fanany MI, Basaruddin T (2016) A New Data Representation Based on Training Data Characteristics to Extract Drug Name Entity in Medical Text. Comput Intell Neurosci. https://doi.org/10.1155/2016/3483528

[2]. Sadikin M (2017) Mining relation extraction based on pattern learning approach. Indones J Electr Eng Comput Sci. https://doi.org/10.11591/ijeecs.v6.i1.pp50-57

[3]. Chang HT, Liu SW, Mishra N (2015) A tracking and summarization system for online Chinese news topics. Aslib J Inf Manag 67:687–699. https://doi.org/10.1108/AJIM-10-2014-0147

[4]. Mikolov T, Sutskever I, Chen K, et al (2013) Distributed representations ofwords and phrases and their compositionality. In: Advances in Neural Information Processing Systems. pp 3111–3119

[5]. Triana YS (2018) Monte Carlo Simulation for Modified Parametric of Sample Selection Models Through Fuzzy Approach. In: IOP Conference Series: Materials Science and Engineering

[6]. Yousif A, Niu Z, Tarus JK, Ahmad A (2019) A survey on sentiment analysis of scientific citations. Artif Intell Rev 52:1805–1838. https://doi.org/10.1007/s10462-017-9597-8

[7]. Arguedas M, Daradoumis T, Xhafa F (2016) Analyzing the effects of emotion management

on time and self-management in computer-based learning. Comput Human Behav 63:517–529. https://doi.org/10.1016/j.chb.2016.05.068

[8]. Buhr EE, Daniels LM, Goegan LD (2019) Cognitive appraisals mediate relationships between two basic psychological needs and emotions in a massive open online course. Comput Human Behav 96:85–94. https://doi.org/10.1016/j.chb.2019.02.009

[9]. Fei H, Li H (2018) The Study of Learners' Emotional Analysis Based on MOOC. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)

[10]. Klenin J, Botov D, Dmitrin Y (2018) Comparison of vector space representations of documents for the task of information retrieval of massive open online courses. In: Communications in Computer and Information Science. pp 156–164

[11]. Liu L, Li B, Bu L, et al (2013) Automatic acquisition of Chinese words' property of times. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp 154–165

[12]. Huang CH, Yin J, Hou F (2011) A text similarity measurement combining word semantic information with TF-IDF method. Jisuanji Xuebao/Chinese J Comput. https://doi.org/10.3724/SP.J.1016.2011.00856

[13]. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings

[14]. Nizamani ZA, Liu H, Chen DM, Niu Z (2017) Automatic approval prediction for software enhancement requests. Autom Softw Eng. https://doi.org/10.1007/s10515-017-0229-y

Cite this article as :

Sulis Sandiwarno, "Users' Emotions Analysis based on Hybrid Feature Extraction Techniques ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 6, pp. 291-296, November-December 2020. Available at doi
: https://doi.org/10.32628/CSEIT206658
Journal URL : http://ijsrcseit.com/CSEIT206658