CLASSIFICATION OF EMOTIONS IN INDONESIAN TEXT USING THE K-NEAREST NEIGHBOR METHOD WITH WIDF WEIGHTING

Wildan Abdul Gani

Teknik Informatika – Universitas Komputer Indonesia
Jl. Dipatiukur 112-114 Bandung
wildan12gani@gmail.com

ABSTRAK

Emotion is an expression of an event related to the feeling of someone who is influenced by a variety of factors, both internal and external. In expressing a person's emotions typically use a variety of ways one of which in his writing. An inscription containing the emotions are very difficult to be interpreted by the reader because it is influenced by the different points of view between the writer and the reader. Therefore to make it easier for the reader to know the emotions of piece of writing can use the method of classification.

On the research methods used in classifying emotion in text that is K-Nearest Neighbor method, the selection is based on research conducted which explains that this method has a good performance with accuracy results up to 86%. In previous research that has been done with the stages of the process from the start preprocessing, weighting TF.IDF and the classification accuracy of obtained results is still lacking around 60%, it is influenced by some of the weaknesses of one in the method of weighting is used. Therefore, in this study the method of weighting to be used is the WIDF which has good performance and development of the TF. IDF.

Based on the test results of this research are implemented in Indonesia language lyrics classify emotions with the K-Nearest Neighbor method and weighting WIDF into five categories, namely happy, fear, anger, sadness and guilt, produce the accuracy of 73.3% on the value of K = 5. To value the relevance of each category of emotions with the use of precision and recall the average value produces a best in category "Fear" for precision, and the category of "angry" for a recall.

Keywords: Emotion, Classification, K-Nearest Neighbor, WIDF Weighting.

1. INTRODUCTION

1.1 Background

Language in an article is a communication tool to translate a thought expressed by an author. In an article a writer not only convey a description of the information, but also provides information about human behavior, including emotions [1]. Emotions are hidden behind the writing is very difficult to be interpreted by the reader due to influence by a different point between writers and readers [2]. Therefore to make it easier for readers to know the emotions of a word can use the method of classification.

Many methods can be used in the text of which is the method of classification Naive Bayes, k-nearest neighbor (KNN), support vector machine (SVM) decision tree, neural network and so forth. Each - each method has advantages and disadvantages of each - each and accuracy vary. As in the study entitled "Review on Comporation Between Text Classification Algorithms" of the comparison algorithm Naive Bayes, k-nearest neighbor (KNN), and support vector machine (SVM), it is known that the algorithm K-nearest neighbor (KNN) has performance great value accuracy up to 86% [3].

In previous studies of text processing in term weighting method frequently used method TF.IDF. This method has a drawback as mentioned in the study entitled "Text Categorization Based on Weighted Inverse Document Frequency" explaining that all the text that contain certain terms are treated equally, so that the IDF does not take into account the number of occurrences of a term in a document. So for term weighting method that will be used in this research is the method WIDF (Weighted Inverse Document Frequency), this method has better performance than TF.IDF by providing increased accuracy in the IDF by 7.4% Maximum [4].

Based on an idea that has been described, in this study used methods of K-nearest neighbor (KNN) with WIDF term weighting method to classify the emotion in the Indonesian language text and know the great accuracy obtained.

1.2 Purpose and objectives

Based on the problems to be studied, the purpose of this study is to classify the emotions in the Indonesian language text using K-nearest neighbor method with weighting WIDF

The goal of the simulation of the introduction of the emotion in the Indonesian language text is:
1. To classify the emotion in the Indonesian language text
2. To determine the great accuracy obtained from the classification of emotions in the Indonesian language text using the K-nearest neighbor weighted WIDF.

1.3 Stages of Development System

Emotion classification system to be built will pass through the stages as follows:
a. Data Sources Analysis
Training data used comes from the data ISEAR (International Survey On Emotion antecedents and Reaction) who speak English and then translated into Indonesian without reducing the meaning of the sentence - the sentence in ISEAR.
As for the test data to be used is the data speak Indonesian song format (.txt)
b. Preprocessing
At this stage, the collected data will be processed so that the data obtained become more structured and easy to prepare. Step - a preprocessing step consists of folding case, convert negation, tokenizing, stopword removal and stemming.
c. Weighting words
At this stage the keyword extraction process will be conducted using values WIDF (Weighted Inverse Document Frequency). Term (words) taken from the results of the last prepocessing are stemming. The value of the weighting of the results will be used as a stage classification using k-nearest neigbor.
d. Classification
The next step is the classification process will be processed using k-nearest neighbor method to determine the emotional content of the test data entered.

2. CONTENTS OF RESEARCH

2.1 Classification

Classification is a process to find a model that explain or differentiate the concept or class of data, with the aim to be able to estimate the class of an object whose class is unknown. In the classification given number of records called training set, which consists of several attributes, attributes can be either continuous or categorical, one of which shows the attributes for the class record.

2.2 Preprocessing

Preprocessing the digging process, process, organize the information by analyzing the relationship, the rules that exist in the textual data semi-structured or unstructured. part - part of a preprocessing stage, can be explained by using a flowchart Flowchart manifold system, which is a chart showing the workflow or what is being done within the overall system and explain the sequence of procedures that exist in the system.

1. Case Folding
At this stage, all the letters will be converted to lowercase.
2. Convert negation
Negation Converter is a conversion process of negation words contained in a statement, because of the word negation have an influence in changing the value of sentiment in a tweet. If there is a negation said it will be merged with the word thereafter. Said - said the negation includes the word "not", "no", "no", "ga", "not", "baseball", "do not" and "no".
3. Tokenizing
Tokenizing a sentence based on the cutting stage that compose each word. This process did the original description decomposition in the form sentences into words and removes symbols like a period (.), Exclamation mark (!), The question mark (?), Comma (,), space, emoticons.
4. Removal Stopword
Stopword defined as a term that is not related to the main subject of the database even though the word is often present in the document and words that are considered not to provide influence in determining a category sentiment.
5. Stemming
Stemming is the stage to transform the words contained in a document to word roots (root word) by using certain rules. By using stemming can reduce the variation of the word has the same root. One stemming algorithms are algorithms Nazief and Adriani.

2.3 WIDF Weighting

Word weighting method used in this system is Weighted Inverse Document Frequency (WIDF). WIDF is a method of the development of the method of the IDF. Inverse Document Frequency (IDF) is the process of measuring term which is rarely present in the corpus. Assessment using the trainer used the entire document. If a term frequently appears, then the term could not be considered a term that represents the document. Conversely, if the term rarely appears on the corpus, then the term is said to have ties with the document. The inverse Document Frequency is a calculation of the total number of documents (D) divided by the frequency of the document (DF) of the word (it), can be expressed with the following equations.

$$\text{IDF} = \log\left(\frac{N}{Dfi}\right)$$

Description:
IDF = inverse document frequency
N = the number of sentences that contain the term (t)
DFI = the number of occurrences of a term of d.
A drawback of the method the IDF is all text containing a specific term is treated the same. This means that the IDF does not account for the number of occurrences of a term in a document. For example

in table 2.1 below, where d (column) is text and t (row) is the term you're looking for, the meeting between in and tj is the termfrequency of the term tj in the text in.

	d_1	d_2	d_3	d_4	d_5
⋮	⋮	⋮	⋮	⋮	⋮
t_x	2	50	3	2	4
t_y	3	2	3	2	3
⋮	⋮	⋮	⋮	⋮	⋮

Figure 1. Example Calculation Of The Emergence Of The Term

It can be seen from equation (2.1) that IDF (tx) and IDF (ty) each is worth 0, because the search term appears on all the text. Frequency distribution, while the tx and ty against d1 to d5 is different, the IDF cannot identify the difference, because the binary nature of the counting. So, unable to determine the number of frequency of term.

IDF assumes that the term is inversely proportional to the amount of text that contain the term. Therefore, an important part of the equation (1) above is a factor 1/df (t). This method of calculating WIDF factor with the term frequency. For example, 1/df (t) from table 2.1 above into

$$\frac{1}{1+1+1+1+1}$$

for all text. Then the number "1" is replaced with the frequency of each term, be

$$\frac{50}{2+50+3+2+4}$$

in the case of d2. It is this factor which is called WIDF. Unlike the IDF, WIDF can distinguish each text d1 ... d5. So, with the term t WIDF in text d can be written as an equation (2) below:

$$WIDF(d,t) = \frac{TF(d,t)}{\sum_{i \epsilon D} TF(i,t)}$$

where TF (d, t) is the term frequency of the term in the text t d and i stated the amount of text. Term of WIDF t add up all term frequency of collection of all text. In other words the WIDF is a form of normalization term frequency of all batches of text [4].

2.4 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a method that uses a supervised algorithm which results from query instance new remedies based on the majority of categories on KNN. KNN has several advantages, namely toughness against training data that has a lot of noise and effective when training his big data. Meanwhile, the weakness of KNN is KNN need to determine the value of the parameter k (the number of nearest neighbors), based on distance training was not clear about the kind of distance is what should be used and which attributes should be used to get the best results, and computational cost is quite high because the required calculation of the distance of each query instance on the overall training sample.

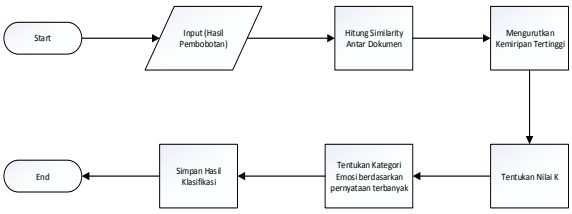


Figure 2. Flowchart Of K-Nearest Neighbor

Steps-steps: KNN method

1. Calculate the distance between the sample data (test data) with data training that has been built. One equation in calculating the distance of the proximity of the Cosine Similarity.

$$Cos(i,k) = \frac{\sum_k (d_i d_k)}{\sqrt{\sum_k d^2_{ik}} \sqrt{\sum_k d^2_{jk}}}$$

Description:

$\sum_k (d_i d_k)$: Vector dot product of i and k

$\sqrt{\sum_k d^2_{ik}}$: The length of the vector i

$\sqrt{\sum_k d^2_{jk}}$: The length of the vector k

I : test data to-i

K : training data to a-j

2. Determine the value of the parameter k = the number of the nearest tetanggaan.
3. Sort the smallest distance from the test data
4. Attach the category in accordance with the conformity

Search for the most of the nearby tetanggaan and then set the category.

2.5 Testing

Testing in this study was done by testing as many as 30 song lyrics discussion of Indonesia with a value of K is 2, 3, 4, 5 and category classification prescribed is happy, sad, emotional tahut, anger, and guilt.

Based on the testing that has been done then accuracy is obtained as follows:

No	Nilai K	Akurasi
1	2	60%
2	3	63.3%
3	4	70%
4	5	73.3%

As for the relevance of the value of each category of emotions by using testing precision and recall.

Following the results of testing the relevance of each category:

	2	3	4	5	Rata-rata
Senang	0.4	0.45	0.56	0.56	0.49
Takut	1	1	1	1	1
Marah	0.5	0.5	0.6	0.6	0.55
Sedih	0.75	0.72	0.75	0.81	0.75
Bersalah	0.6	1	1	1	0.9

	2	3	4	5	Rata-rata
Senang	0.4	0.45	0.56	0.56	0.49
Takut	1	1	1	1	1
Marah	0.5	0.5	0.6	0.6	0.55
Sedih	0.75	0.72	0.75	0.81	0.75
Bersalah	0.6	1	1	1	0.9

3. FINAL

3.1 Conclusions

Based on research that has been done on the classification of emotions in the Indonesian language text using k-nearest neighbor method with weighting WIDF can be summarized as follows;

1. Simulator made for the classification of emotions in the Indonesian language text using k-nearest neighbor method by implementing the song lyrics in Indonesian language the test was successful, by providing emotional categories that correspond to the content of the lyrics of the song.
2. The success of the classification of emotions in the lyrics of songs influenced by the value of K in the k-nearest neighbor and training data are used..
3. Using the methods of weighting WIDF in the process of classifying emotions provide increased accuracy values from previous studies with great accuracy obtained amounted to 73.3%.
4. To value the relevance of each category of emotion by using precision and recall generate maximum value of 100%. The value obtained in the category of "Fear" for precision, and the category of "Angry" to recallnya.

3.2 Suggestion

Based on the research that has been done there is a shortage obtained that the classification process takes a long time, so the authors provide suggestions for future research in order to further enhance and training data used to be able to use the original data in Indonesian language so that the results obtained are more relevant

BIBLIOGRAPHY

[1] C. Z.-j. H.W, "Multimodal Emotion Recognation From Speech," p. 1, 2004.

[2] S. M. Kim, "Recognising Emotions and Sentiments in Text," Sydney, The University of Sydney, 2011, p. 26.

[3] J. A. Vaibhav C.Gandhi, "Review on Comparison between Text Classification Algorithms," International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) , vol. 1, no. 3, 2012.

[4] M. I. Takenobu Tokunage, "Text Categorization Based on Weighted Inverse Document Frequency," in ISSN 0918-2802, Ookayama, Meguro Tokyo Japang, Department Of Computer Science Tokyo Institute Of Technology , 1994.