SENTIMENT ANALYSIS OF CUSTOMER OPINION IN KARAWITAN PHD METHOD USING K - NEAREST NEIGHBOR

Taufik Hidayat Supratman

Informatics Engineering – Indonesia Computer University
Dipatiukur Street No. 112-114 Bandung
E-mail : akasia_a7x@yahoo.com

ABSTRACT

PHD (Pizza hut delivery) provides a forum to share their customer experiences in the form of an assessment of the PHD who visited the outlet can be channeled through the website www.halopizzaindonesia.com. Opinion received will be distributed to each outlet including an outlet PHD PHD Karawitan. Outlet Manager assigned to calculate the percentage of customer satisfaction assessment by studying the opinions submitted by the customer and group assessment into several classifications are satisfied opinion and opinions are not satisfied. Number of ratings that many have difficulty making the outlet manager must classify opinion because each one of the many opinions there and then calculate the percentage of customer satisfaction manually.

To facilitate the work, especially outlet manager specialized in grouping assessment such as customer opinion, we need a method that can automatically analyze opinion. One way that can be used to categorize the opinion that by using sentiment analysis. Sentiment analysis methods were used in this research is using the K-nearest algorithm neigbor. Consists of the preprocessing stage, weighting word, and classification k-nearest neigbor.

Based on the results of the implementation and testing that has been performed k-nearest neigbor algorithm can be applied to a case study in PHD Karawitan customer and can help simplify the outlet manager with grouping assessment satisfied opinion and opinions are not satisfied automatically, and can calculate the percentage of that opinion.

Keywords: K-nearest neigbor, Opinion, Pizza Hut delivery, PHD.

1. INTRODUCTION

PHD (Pizza hut delivery) is a company engaged in the culinary field. To determine the services provided from each outlet, then PHD provides a forum to share their customer experiences in the form of an assessment of the PHD outlets were visited. Assessment customers can be routed via the website www.halopizzaindonesia.com. All ratings stored in the website will be distributed to the respective outlets throughout Indonesia PHD. Outlet Manager assigned to calculate the percentage of how much the assessment of incoming customer satisfaction in the can by studying the opinion delivered by the customer

In this research, PHD outlets that will be a place of research that Karawitan PHD outlet. Number of ratings that many have difficulty making the outlet manager must classify opinion because each one of the many opinions there and then calculate the percentage of customer satisfaction manually. One way that can be used to categorize the opinion that by using sentiment analysis.

Sentiment analysis is a method in analyzing opinions, judgments, and attitudes to an entity such as products, services, events, and topics. [1] Based on previous research on the classification of text that tries to compare the accuracy of three algorithms direct the method Naive Bayes, method of k-nearest neighbor (KNN), the method of support vector machine (SVM) found that the method of K-NN has a greater degree of accuracy compared with the Naive Bayes method and the method of support vector machine (SVM). [2].

1.1 Method Development Sentiment Analysis

Sentiment analysis system that will be built will pass through the stages of the stages consisting of:

a. Analysis of Data Sources

The data used is the data obtained from the website www.halopizzaindonesia.com already distributed to outlets PHD Karawitan. .xls Format data files and data processing for the next stage will be the process of import data.

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, tokenizing, stopword removal and stemming.

c. Weighting word

At this stage it will be done using the keyword extraction process value TF-IDF (Term Frequency - Inverse Document Frequency). Term (words) taken from the results of the last prepocessing are stemming. The value of the weighting results will be used as a classification stage using k-nearest neigbor.

d. Sentiment Classification

The next step is the classification process will be processed using the k-nearest neigbor to determine which ones included satisfied opinion and which included opinions are not satisfied.

e. Visualization

Results of sentiment classification phase will be described in the form of pie charts. Data taken from the number of satisfied opinion and opinions are not satisfied then displayed as a percentage of the diagram

1.2 Method of Software Development

Software development methods used in this study is the waterfall method according to Ian Sommerville reference [3] . Broadly speaking the method waterfall looks like Figure 1.1 below :

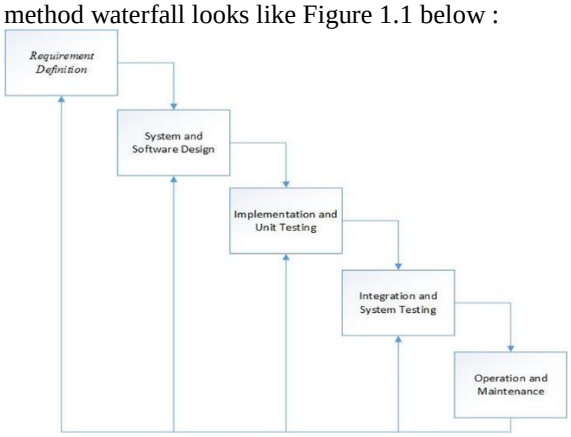


Figure 1.1 waterfall method according to Ian Sommerville

2 CONTENTS RESEARCH

2.1 Analysis Sentiment

Sentiment analysis is a way to analyze the opinions, judgments, and attitudes to an entity such as a product , the character of people , services , events and topics . One example of the use of sentiment analysis in the real world is the identification of market trends and market opinion to an object goods . [4]

2.2.1 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 the preprocessing phase, can be explained by using a flowchart Flowchart manifold system, which is a chart that shows the workflow or what is being done in the overall system and explain the sequence of procedures that exist in the system [11].

1. Case Folding

At this stage, all the letters will be converted to lowercase.

measures in case of folding stages are as follows :

1. Check the size of each character from the beginning to the end of the character .
2. If it is found that the character using capital letters (uppercase) , then the letter will be converted to lowercase (lowercase) . The following flowchart folding case shown in Figure 2.1

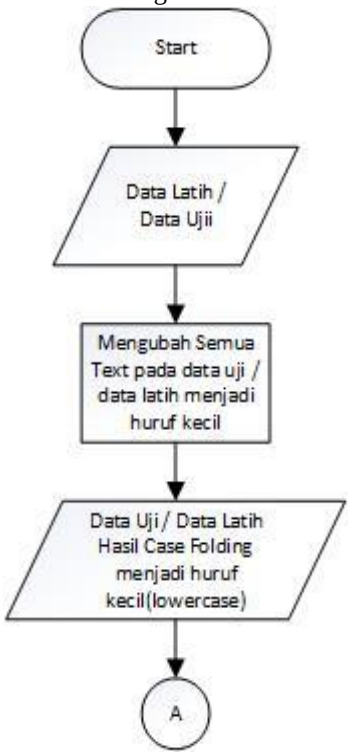


Figure 2.1 flowchart case folding

2. Convert negation

Convert negation is the process of converting the words of negation that there is an opinion, because 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 “bukan”, “tidak”, “tak”, “ga”,”gak”, “enggak”, “jangan”, dan ”nggak”.

Steps to convert negation stage are as follows :

1. The word used is the result of a case of folding
2. If it is found that contains the word opinion - says it will put together the negation of the negation word by word after word of the negation .

Convert negation following flowchart shown in Figure 2.2



Figure 2.2 Flowchart Convert Negation

3. Tokenizing

Tokenizing is the stage of cutting a sentence based on every word that constitute it. This process did decomposition original description in the form of sentences into words and removes symbols such as a period (.), Exclamation mark (!), A question mark (?), Comma (,), space, emoticons.

Measures on tokenizing stages are as follows :

1. The word used is the result of the negation convert .
2. Cut each word in the sentence based on the word separator is a space .
3. Eliminate symbols such as a period (.) , Exclamation mark (!) , A question mark (?) , Comma (,) , space , emoticons .

The following flowchart shown in figure 2.3 Tokenizing

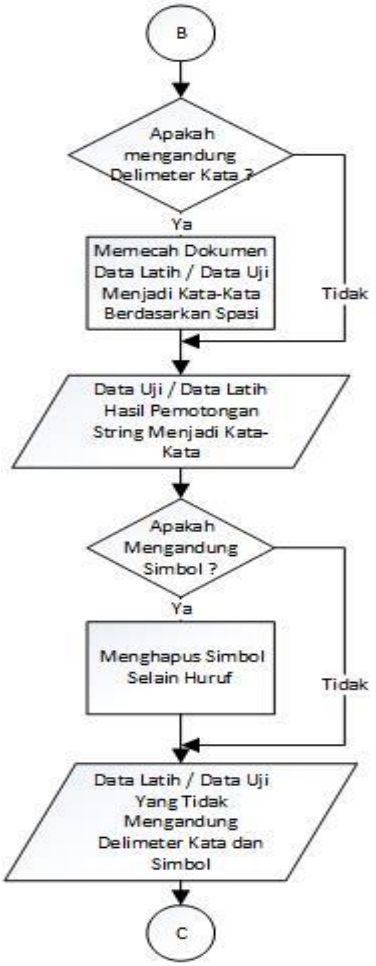


Figure 2.3 Flowchart Tokenizing

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 unable to give effect to determine a category sentiment.

The words are inserted into the stopword list which is usually in the form of :

1. pronoun people. Can only be used to replace the noun person , the person's name , or anything else that personified . For example : he , you , Mr. , Mrs. , Mr. , Mrs. , MBA , Mr. , Mrs., employee , worker , employee , etc.
2. pronoun pen. For example : what, when , why , who, how , what , where , to where, etc.
3. pronoun instructions . For example : this , that , etc.
4. pronoun liaison. For example : the , and , or etc.
5. The word is irrelevant . For example : one , because , really , well , somewhat , with , should , from , with , dg , which , okay , etc.

Steps in stopword removal is as follows :

1. The word tokenizing results will be compared with a list of stopword.Dilakukan checking whether a word at the stopword list or not .

2. If the word is the same as that of the stopword list , then it will be removed .
Removal Stopword following flowchart shown in figure 2.4

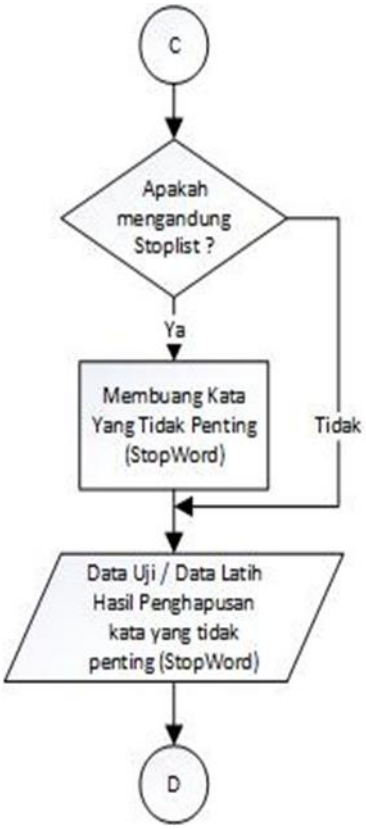


Figure 2.4 *flowchart StopWord Removal*

6. 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 [5].
Measures stemming using Nazief and Adriani algorithm is as follows:
1. The word that is not in the dictionary look-stemming. If the word was immediately found, the word is the word basic. The word is restored and the algorithm is stopped.
2. Remove inflectional suffixes first. If this is successful and the suffix particle ("lah" or "kah"), this step is performed again to remove inflectional suffixes possessive pronoun ("I", "you" or "his").
3. Derivational suffix ("-i", "-an" and "right") and then removed. Then this step is continued again to check whether there derivational suffix is left, if there is then eliminated. If nothing else then do the next step.
4. Then derivational prefix ("in -" "to -", "se -", "te -", "be -", "him" and "PER" ") are removed. Then this step is continued again to check whether there derivational prefix is left, if there is then eliminated. If nothing else then do the next step.
5. After no more affixes left, then the algorithm is terminated then the basic words are searched in the dictionary, if it is to meet the basic word means algorithm is successful but if it does not meet the basic words in the dictionary, then the recoding.
6. If all steps have been done, but the basic word is not found in the dictionary as well, then the algorithm returns the original words before stemming.

Process stemming in this study is the final process of the preprocessing stage, after opinion preprocessing results already done so weighted opinion said that could be classified using KNN method.
Stemming the following flowchart shown in Figure 2.5

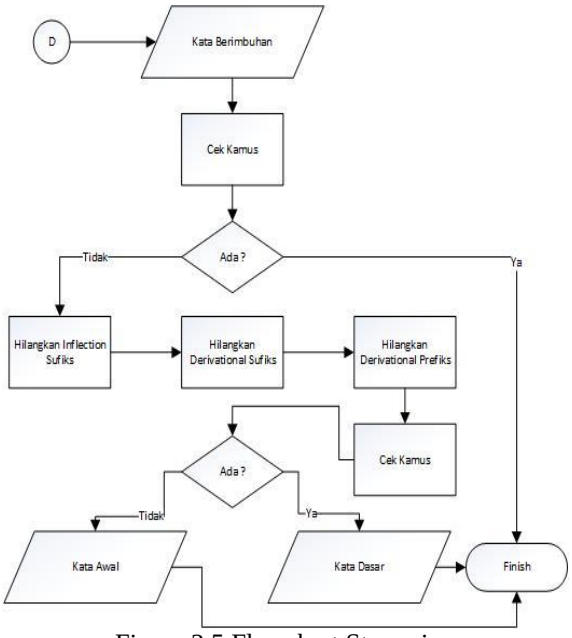


Figure 2.5 Flowchart Stemming

2.2.2 Weighting word (TF IDF)

Term weighting is a stage to give a value / weight on the terms contained in a document after passing through preprocessing .

$$Idf = \log\left(\frac{N}{Dfi}\right) \quad (2.2)$$

IDF = inverse document frequency
N = The number of sentences containing the term (t)
Dfi = The number of occurrences of the term D

Said weighting is done after preprocessing stage , the value of the weighting results then the word will be used to calculate the similarity between documents (Cosine Similarity) which is the stage in

which the opinion classification using KNN method .

Here is a step in the weighting :

1. Create a composition of a record of the word - the word good opinion training data or test data , words - words that appear more than one written only 1x
2. Fill in the numbers 1 when he appeared in every (tf) either training data or test data , if it does not appear browse the numbers 0
3. Calculate the number of words - words that appear in all the training data and test (df)
4. Calculate Idf using equation (2.2)

Multiply each - each W with each - each Tf in accordance with her partner . Example : Tf (i) * W (i)

Sample case

k1 : layanan cepat ramah

k2 : layanan kurang muas kurang sopan

Table 2.1 step in the weighting

Kata	Tfk1	Tf k2	Df	Idf	Wk1	Wk2
layanan	1	1	2	0	0	0
cepat	1	0	1	0.30	0.30	0
ramah	1	0	1	0.30	0.30	0
kurang	0	2	2	0.30	0	0.60
muas	0	1	1	0.30	0	0.30
sopan	0	1	1	0.30	0	0.30

Results of the weighting value will be used for the calculation of cosine simmilirity the KNN method .

2.2.3 Classification of K-Nearest neigbor (KNN)

K - Nearest Neighbor (KNN) is a method that uses an algorithm supervised instance where the results of a new query based on the majority of categories diklasifikan on KNN . KNN has several advantages , namely resilience training data that has a lot of noise and effective training if the data is large . Meanwhile , the weakness of KNN is KNN need to determine the value of the parameter k (the number of nearest neighbors) , training based on distance is not clear about the type of distance what to use and which attributes should be used to get the best results , and computational cost is quite high due to the necessary calculation distance from each query instance in the entire training sample. [6] .

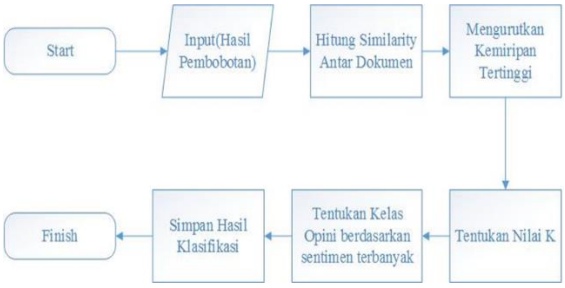


Figure 2.6 Classification of K-Nearest neigbor (KNN)

Steps method KNN :

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

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

Description :

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

$\sqrt{\sum_k d^2{}_{ik}}$: Length vector i

$\sqrt{\sum_k d^2{}_{jk}}$: Length vector i

i : test data to - i

k : training data to - j

2. Determining the value parameter k = number next door to the nearest free .
3. Sort the smallest distance from the test data
4. Fit the categories according to appropriateness

Find the highest number of next door to the nearest Then set category . [7]

2.2.4 Testing Confusion Matrix

Confusion Matrix [8] contains information about the actual classification and the prediction made by the classification system . System performance is generally evaluated using the data in the matrix . The following table shows the confusion matrix for classification of two classes.

		Diklasifikasikan sebagai (*Predicted Class*)	
		Puas	Tidak Puas
Klasifikasi yang benar (*Actual Class*)	Puas	True Puas	False Tidak Puas
	Tidak Puas	False Puas	True Tidak Puas

Table 2.2 Testing Confusion Matrix

2.2.5 Testing *K-Fold Cross-Validation*

In the approach K-fold cross-validation, each record is used several times in the same amount for proper training and once for testing. To illustrate this method, count as a data partitioned into two subsets of equal size. First, select one of the second subset for training and another for testing. Then the exchange of the function of a subset such that a subset previously as a training set into a test set and vice versa. This approach is called a two-fold cross validation. Total error is obtained by summing the error -Error tersebut.Setiap record for the second process is used exactly once for training and one for testing. Methods k-fold cross-validation generalize this approach by segmenting data into k partitions of equal size. During the process, one of the partition chosen for testing, while the rest is used for training. This procedure is repeated k times so that each partition is used for testing exactly once. Total error is determined by summing the error for all k process [9].

Methods k-fold cross-validation set k = N, the size of the data set. This method of approach is called a leave-one-out, each test set contains only one record. This approach has the advantage in the use of as much data for training. Test sets are mutually exclusive and can effectively cover the entire data set. Disadvantages of this approach is the number of computing to repeat the procedure as much as N times.

K-fold cross-validation is one of the techniques to evaluate the accuracy of the model, with characteristics [10]:

1. Partitioning the data randomly into k sets / fold namely D1, D2, ..Dk. Each group has nearly the same amount.
2. At iteration i, used in a set of test data and the other as the training data Example:

a. At iteration 1: D1 and D2 as test data sd .dk as training data
b. In the 2nd iteration: as the test data D2 and D1, D3 sd Dk as the data pelatihanan onwards
c. Conduct training and testing as much as k times
d. Calculate the accuracy of the formula.

As an illustration, if we do a 3-Fold Cross-Validation eksperimenya then design the data is as follows:

Dataset :

Tabel **Error! No text of specified style in document.**.3 Data Set

K1	K2	K3

Data Subset :

Tabel **Error! No text of specified style in document.**.4 Data Subset

Experiments to	Training	Testing
1.	K2, K3	K1
2.	K1, K3	K2
3.	K1, K2	K3

2.2.6 Testing Blackbox

Black box testing has some kind . Blackbox testing to be used in this study is the functional testing . In this test , the software is tested for functional requirements . Tests conducted in written form to check whether the application is running as expected . Functional testing covers how well the system is performing its functions . [12].

blackbox test case examples :

Black box testing is done on an application made is divided into two parts, the test scenarios , test cases and results .

1. Scenario Testing

Test scenario describes the testing of the system that existed at this sentiment analysis applications . Scenario testing to be performed on this application can.

Table 2.5 Scenario Testing blackbox

No.	The components in the test	This type of testing
1.	Proses Pengolahan Data Latih	*Black Box*
2.	Proses Pengolahan Data Uji	*Black Box*
3.	Proses Klasifikasi	*Black Box*
4.	Proses Visualisasi	*Black Box*

2. Cases and Test Results

Cases and test results shows that the exposure of the testing plan has been drawn up on the test scenario . Testing was done in black box with attention to every case that will be tested . Here is the exposure of each.

a. Testing Data Processing

Testing Data Processing Train describes the tests performed on the current user activity will conduct training data retrieval . Testing training data

processing will be done on this application can be seen in Table 2.6

Table 2.6 Testing Data Processing Train

cases Will In Test	Test results	
	Expected results	conclusion
Pengolahan Data Latih	sistem berhasil melakukan penyimpanan data latih yang telah ter-*processing* otomatis pada database	[√] Berhasil [ ] Tidak Berhasil

b. Testing Test Data

Testing Test Data Processing describes tests performed on user activity when going to test data retrieval . Testing test data processing will be done on this application can be seen in Table 2.7

Table 2.7 Testing Test Data

cases Will In Test	Test results	
	Expected results	conclusion
Pengolahan Data Uji	sistem berhasil melakukan penyimpanan data uji pada database	[√] Berhasil [ ] Tidak Berhasil

c. Testing Classification

Testing classification describes tests performed on user activity when going to the classification opinions . Testing the classification will be done on this application can be seen in Table 2.8

Table 2.8 testing Classification

cases Will In Test	Test results	
	Expected results	conclusion
Mengklasifikasikan penilaian *customer* (Data uji)	sistem berhasil mengklasifikasikan penilaian customer(data uji) dalam sentimen puas atau tidak puas	[√] Berhasil [ ] Tidak Berhasil

d. Testing Visualization

Testing visualization describes tests performed on the current user activity will be visualized in a pie chart. Visualization testing will be done on this application can be seen in Table 2.9

Table 2.9 Testing Visualization

cases Will In Test	Test results	
	Expected results	conclusion
Menampilkan Diagram persentase sentimen puas dan sentimen tidak puas	sistem berhasil Menampilkan Diagram persentase sentimen puas dan sentimen tidak puas sesuai dengan data yang sudah diklasifikasikan sebelumnya	[√] Berhasil [ ] Tidak Berhasil

3 CLOSING

3.1 Conclusion

Based on the results of the implementation and testing has been done on the system sentiment analysis using the K - nearest neigbor this it can be concluded that the system analyzes this sentiment may classify the opinions automatically and can provide information on the percentage of sentiment satisfied and sentiment are not satisfied so you can help Outlet Manager job easier , especially in the classification of customer assessment (opinion) .

3.2 Suggestion

Suggestions for further development , as follows :

1. Increasing the number of training data to get a better result when the classification of customer assessment .
2. Optimization of time in the processing system.

REFERENCE

[1] B. Liu, "Sentiment Analysis and Opinion Mining, Morgan & Claypool," 2012.
[2] K. Khamar, "Short Text Classification Using KNN Based on Distance Function," International journal of advanced research in computer and communication engineering, 2013.
[3] I. Sommerville, Software Engineering, jakarta: Erlangga, 2003.
[4] B. P. a. L. Lee, ""Opinion Mining and Sentiment Analysis"," Foundations and Trends in Information Retrieval, vol II, no 1-2, pp. 1-135.
[5] Agusta, "Perbandingan Algoritma Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia," Konferensi Nasional Sistem dan Informatika, 2009.
[6] Evan, Buku TA : K-Nearest Neigbor(KNN), Jakarta: Informatika, 2010.
[7] D. E. Garcia, "The Classic Vector Space Model (Description, Advantages and Limitations of the Classic Vector Space Model)," 2005.
[8] K. a. Provost, "On Applied Research in Machine Learning," In Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, 1998.

[9] J.Han, "Data Mining : Concepts and techniques, Morgan Kaufman,," 2006.
[10] Nugroho, "SVM : Paradigma Baru dalam Softcomputing dan Aplikasinya," Konferensi Nasional Sistem & Informatika, 2008.
[11] Sudarsono, "[Online] Analisis Dan Perancangan Sistem," Available : sdarsono.staff.gunadarma.ac.id/Downloads/files/16512/Flowchart.pdf [Accessed 15 Agustus 2015].
[12]J. Simarmata, rekayasa perangkat lunak, Yogayakarta: Andi.