SENTIMENT ANALYSIS OF PUBLIC OPINION BASED ON INDIHOME TELKOM USING IMPROVED K-NEAREST NEIGHBOR METHOD

Herdiawan[1]

[1]Informatic Engineering – Indonesian Computer University
112-114-116 Dipati Ukur Street, 40132 Bandung, Indonesia
Email: if.herdiawan@gmail.com[1]

ABSTRACT

Indihome is the latest internet service product of PT. Telkom. The Indihome users reach 300 thousand users until now. Since Indihome users are increasing excessively, PT. Telkom wants to provide the service in the form of product assessment feedback, in order to know the response of consumers to the Indihome. Many consumers discusses Indihome in social media, especially twitter. They share about the quality and shortcoming of Internet service. Unfortunately, twitter does not have the ability to aggregate information derived from the conversation that leads to a conclusion.

One of the ways to make a conclusion from the results of aggregation is using text mining. Improved K-Nearest Neighbor is one of algorithms that can be used for the implementation of classification. The resolution process of ImprovedK Nearest Neighbor algorithm is started from preprocessing consisted of Convert emoticons, Cleansing, Case Folding, Convert negation, Tokenizing, Filtering, and Stemming. The next process is weighting word, then the categorization consisted of calculating the cosine similarity, k-value values and classifying the sentiment in the form of graphs, so the result of this sentiment analysis can be used as an evaluation in determining the further business steps or the quality improvement.

Keywords: sentiment analysis, text mining, classification, ImprovedK-Nearest Neighbor, Indihome

1. INTRODUCTION

Internet users in Indonesia grew a lot. According to the Ministry of communications and Informatics in the year 2015, the number of Internet users in Indonesia has reached the figure of 150 million people, or approximately 61% off all population [1].

PT. Telekomunikasi Indonesia, Tbk as service providers by combining IndiHome some services into one, currently has a promotion service that is able to attract many consumers. Start a large number of users making PT. Telekomunikasi Indonesia, Tbk wish to provide services in the form of feedback assessment of IndiHome products. Mr. Sony Budi Winarso as Manager Marketing Integration Reg-3 plan wanted to know how the response of the consumer there is a products of Indihome social media because he thinks many consumers Telkom which provide commentary on Telkom's products in social media twitter.

See these problems it is necessary the presence of a way how classifying information public sentiment towards Telkom IndiHome of existing public opinion on social media, for get information from the results of classification data through social media twitter in the form of percentage of consumer satisfaction results that can be used as material for the evaluation of Telkom Indihome in order to further improve the quality of its services so that it can fix and determine the next steps the business better.

1.1. Sentiment Analysis

Sentiment analysis or opinion mining is the process of understanding, extract and manipulate textual data automatically to get information the sentiments contained in one sentence opinion to produce a new opinion[6]. Sentiment analysis is conducted to see the opinion or the tendency of opinion against an issue or object by someone, whether sighted or tends to be negative or positive opined. Sentiment analysis is usually conducted to monitor the development of the market or to view a response to a problem, one example of the use of sentiment analysis in the real world is the identification of market tendencies and market opinion against an object products [7].

Sentiment analysis is essentially a classification, but it is not as easy as the usual classification process because the associated use of language that is constantly growing. Where is the media used in this case is an ambiguous text because there is no intonation in a text [8].
Benefits of sentiment analysis to the development of a very large business, so many companies implement analysis of sentiment as the media to look at market developments in defining business measures taken as material considerations such firms.

1.2. Text Mining

Text Mining is data of text where the data source is usually obtained from dokuman, aimed at searching for words that can mewaliki the contents of the document so that it can be done analysis of the connectedness between documents [9].

Text mining is done by a computer to get something new, something previously unknown or reinvent the information implicitly implied, that comes from information that is extracted automatically from text data sources vary.

1.3. Regular Expression

Regular Expression or commonly abbreviated as regex is a special text to describe a search pattern. Regex used for searching or manipulation of text. Regex is supported by many basaha programming, such as Java, PHP, C # and many other programming languages. Here are the rules of writing of the Regex in the Java programming language [10].

1.4. Preprocessing

Text Preprocessing which is the initial stage of text mining that will process data training and data testing. Text Preprocessing aims to prepare unstructured text documents into structured data that is ready to use for the next process. Text Preprocessing stages in this study include:

1. *Convert Negation*

Emoticons are a combination of "emotion" and "icons" which means the icon that is used to express emotions in a written statement, and can change and improve interpretation against the text.

2. *Cleansing*

This stage will remove all characters other than alphabetical order with the aim to reduce the nois. As it known that this emoticon symbol with a combination of special characters and numbers, so that also these emoticons are not erased. In addition, special characters, URLs, the hashtag (#), username (@username), a comma (,), a trailing period (.), exclamation point (!), a semicolon (;), colon (:), hyphen (-), ellipses (...), the question mark (?), parentheses (()), brackets ({...}), quotation marks ("..."), single quotation marks (' '), a slash (/) and (\), and apostrophe (') will be eliminated.

3. *Case Folding*

It is the stages of change all enter the letters into small letters (lower case). Because the system will be built using the java programming language, then likened first into the same form, in this case into lower case.

4. *Convert Negation*

Convert negation is done if there is a negation Word before the word is positive, then the Word will change its value becomes negative and likewise vice-versa. Words that are the negation of such “bukan”, “bkn”, “tidak”, “enggak”, “g”, “gak”,“tidak”, “tdk”, “enggak”, “engga”, “ga”, “gk”, “jangan”, “jgn”, “nggak”, “tak” dan “gak”.

5. *Tokenizing*

Tokenizing is the stage of cutting text documents on every word that composed it. The word pieces called tokens or term. At this stage would do checking of tweets from the first character to the last character.

6. *Filtering*

Filtering plays to throw out words that often appear and are less public, indicate relevance to the text. This process will eliminate the words that often appear but has no pengaruhapapun in the extraction of the sentiments of a tweet.

7. *Stemming*

The stage is a stage looking for Stemming the root word of each word filtering results. Words that appear in the documents often contain suffixes. Therefore, any left over from the process results filtering stages are formed into the basic Word by way of eliminating his affixes.

Stemming algorithms used in this study i.e. Stripping Confix Stemmer algorithm. The algorithm is to add additional algorithms to overcome a mistake breaks the suffix should not be done.

1.5. Confix Stripping Stemmer

Confix stripping Stemmer is a method of stemming on the Indonesian Language introduced by Dainty Asian which is the development of methods stemming made by Nazief and Adriani (1996). Words that appear in the documents often contain suffixes. Therefore, any left over from the process results filtering stages are formed into the base with the words how to remove imbuhannya.

Basically, the algorithm is often grouped into several categories as follows:

1. Inflection Suffixes suffix groups i.e., that does not change the basic Word form. This group can be divided into two:
   a. Particle (P) or particles, including particle "-lah", "-kah", "-tah", dan "-pun".
   b. Possessive Pronoun (PP) or pronoun belongs, including "-ku" , "-mu", dan "-nya".

2. *Derivation Suffixes* (DS) is the set of suffixes that can be added directly on the base said. Included in this is the suffix type "-i", "-kan", dan "-an".

3. Stages of the Derivation with Prefixes which (DP) is a collection of prefix can be directly given on the basis of pure words, or on the basis of the words already get the addition of up to 2 a prefix. Included in it is a prefix that can be morfologi ("me-", "be-", "pe-", dan "te-") and no prefix morphology ("di-", "ke-" dan "se-"). On the basis of classification suffixes-affixes, forms of the word in the language of Indonesia can be modeled as follows:

   [DP+[DP + [DP+]]] Kata Dasar [[+DS][+PP][+P]] (1)

   With the following limitations:

   a. Not all combinations are allowed. The combination of affixes which prohibited can be seen in table 1.
   b. The use of the same prefix repeatedly is not allowed.
   c. If a Word consists of only one or two letters, then the process of stemming is not done.
   d. The addition of a specific prefix can change the form of the original Word, or prefix that has been previously granted on the basis of the word in question (morphologhy).

Table 1 a combination of Prefix and sfix

Awalan (Prefiks)	Akhiran (Suffiks)
be-	-i
di-	-an
ke-	-i –kan
me-	-an
se-	-i –kan
te-	-an

Table 2 basic word Decay Rules

Aturan	Bentuk Awalan	Peluruhan
1	berV…	ber-V… | be-rV…
2	belajar…	bel-ajar
3	beC$_1$erC$_2$…	be-C$_1$erC$_2$…dimana C$_1$={'r'|'l'}
4	terV…	ter-V…|te-rV…
5	terCer…	ter-Cer…dimana C!='r'
6	teC$_1$erC$_2$	te-C$_1$erC$_2$…dimana C$_1$!='r'
7	me{l|r|w|y}V…	me-{l|r|w|y}V…
8	mem{b|f|v}…	mem-{b|f|v}…
9	mempe…	mem-pe…
10	mem{rV|V}…	me-m{rV|V}…|me-p{rV|V}…
11	men{c|d|j|z}…	men-{c|d|j|z}…
12	menV…	me-nV…|me-tV…
13	meng{g|h|q|k }…	meng-{g|h|q|k}…
14	mengV…	meng-V…|meng-kV…
15	mengeC	menge-C
16	menyV…	me-ny…| meny-sV…
17	mempV…	mem-pV…
18	pe{w|y}V…	pe-{w|y}V…
19	perV…	per-V…|pe-rV…
20	pem{b|f|v}…	pem-{b|f|v}…
21	pem{rV|V}…	pe-m{rV|V}…|pe-p{rV|V}
22	pen{c|d|j|z}…	pen-{c|d|j|z}…
23	penV…	pe-nV…|pe-tV…
24	peng{g|h|q}	peng-{g|h|q}
25	pengV	peng-V|peng-kV
26	penyV…	pe-nya|peny-sV
27	pelV..	pe-lV…; kecuali untuk kata "pelajar" menjadi "ajar"
28	peCP	pe-CP…dimana C!={r|w|y|l|m|n} dan P!='er'
29	perCerV	Per-CerV… dimana C!={r|w|y|l|m|n}

CS stemmer algorithm works as follows:

1. The word on stemming first sought in the dictionary. If found, it means that the word is the Word Basic, if not then step 2 is done.

2. Check the rule precedence. When a Word has the prefix-suffix "be-lah", "be-an", "me-i", "di-i", "pe-i", or "te-i" then the next step stemming is (5, 6, 3, 4, 7). If the Word does not have a prefix-suffix pair, stemming the normal steps (3, 4, 5, 6, 7).

3. Eliminate inflectional particle P ("-lah", "-kah", "-tah", "-pun") and a pronoun or possessive pronoun belongs to PP *pronoun PP* ("-ku", "-mu", "-nya").

4. Eliminate Derivation Suffixes *DS* ("-i", "-kan", or "-an").

5. Elimitane Derivational Prefixes *DP* {"di-","ke-","se-","me-","be-","pe", "te-"} By iteration is three times the maximum:
   a. 5 step this stop if:
      1. Occurring combination forbidden

2. A prefix that are detected by of current equal to a prefix being omitted previously
3. Three prefix has been omitted.

b. dentify a prefix and clear.There are two types of a prefix:
   1. Standard: “di-”, “ke-”, “se-” That can directly omitted from the word.
   2. Kompleks: “me-”, “be-”, “pe”, “te-” is the prefix type morfologi that can be in accordance with the base said that followed .Because of it , use the rules on the table ii-14 to get proper beheading.

c. What the word that has been omitted awalannya this in a dictionary .If not found , then step back repeated 5 .If found , then the whole process was stopped.

6. If after five basic steps said it was not found the process of recoding done with reference to the rules on tables ii-14.Recoding done by adding recoding characters in the word was decapitated.Ii-14 on a chart, the character is the character after recoding (’-’) and sometimes being before parentheses.For example, in a “menangkap” (aturan 15), rules after being severed “nangkap”.The invalid, then recoding done and produce “tangkap”.The rule should be 22 not found in the fairest Jelita Asian.

7. If all measures fail , said that input and tested on an algorithm is regarded as a basic .

If the word will stemming found hyphens (’-’), Hence the possibility of a word to stemming is said to repeated .Stemming to a word repeated done by breaking up the word into two parts is part of the left and right based on the position of hyphens ’-’) And do stemming 1-7 step in the two words .If the results of both of them stemming the same , then the basic been obtained .

1.6. Term Weighting

Weighting term weighting is a technique in any term or word .This stage most of the weighting in the text mining technique using tf.idf . Tf.idf apply weighting of the multiplication of the weighting of a combination of both frequency and local term global (global weight inverse document frequency). [13] A method of tf-idf can be formulated as follows:

$$idf = \log\left(\frac{N}{df}\right) \quad (2)$$

Where :
N = Of all of the data
df = document frequency

$$w(t, d) = tf(t, d) * IDF \quad (3)$$

Where :
tf = term frequency
IDF = Inverse Document Frequency
d = Document into-d
t = said into-t of keywords
w(t,d) = The weighting of documents into-d to the word into-t

1.7. Improved K-Nearest Neighbor

The determination needed to get proper k-values high accuracy of test categorization documents in the process .Improved algorithms k-nearest k-values neighbors do a modification in the determination .Where the determination of k-values be done , just having different k-values each category .Differences in each category k-values owned besar-kecilnya big or small the adapted to the number of documents trainer owned by the category .So when k-values getting high , the results of categories not affected in the category of having a larger number of documents trainer .

To compute similaritas between the two documents using the cosine similarity (CosSim).Seen as a measure (similarity measure) between vector document (d) with a vector query (q).The same document with a vector vector query the document could be considered more appropriate with queries. [13] The formula used to calculate cosine similarity is as follows:

$$Cos(\theta_{QD}) = \frac{\sum_{i=1}^{n} Q_i \cdot D_i}{\sqrt{\sum_{i=1}^{n}(Q_i)^2} \cdot \sqrt{\sum_{i=1}^{n}(D_i)^2}} \quad (4)$$

Where :
$Cos(\theta_{QD})$ = Resemblance documents Q terhadap D
Q = Data Testing
D = Data Training
n = Of all of the data

An algorithmic k-values on the improved k-nearest neighbor was done using equation (4) the first rank in the reckoning similaritas decline in each category.

Next on improved algorithms k-nearest neighbor, k-values new called by n. equation (4) explaining of the percentage of the determination of k-values (n) in all categories.

$$n = \left[\frac{k * N(c_m)}{maks\{N(c_m) \mid j=1.....N_c\}}\right] \quad (5)$$

Where :
n = New k-values
k = k-values Set
$N(c_m)$ = The amount of data training in the category / category m
$maks\{N(c_m) \mid j=1.....N_c\}$ = The amount of data on training most of all categories.

A number of n documents selected in each category is a top n documents or document top is a document that has most similaritas in each of the category

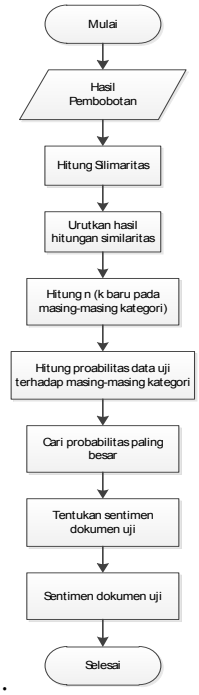


Image 1 *Flowchart Improved K-Nearest Neighbor*

1.8. Precision, Recall dan F-Measure

A system of gathering information back to return a bunch of documents as the answer to queries users .There are two categories of documents produced by a system of common ground back information related to query processing , that is relevant documents (relevant documents with queries) and documents retrieved (documents received by the user). A common measure used to measure the quality of data retrieval is a combination of precision and recall .

Precision evaluate the ability of the system of gathering information back to find back data top-ranked most relevant , and is defined as the percentage of the data returned really relevant to queries users .Precision is the proportion of a set of obtained relevant .Precision can be formulated the equation (6).

Table 3 table contingency

	Predicted Class		
		+	-
Actual Class	+	TP	FN
	-	FP	TN

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

To describe the third up so we can get the equation (6), and (7) in order to obtain the value of precision and recall.It is true that the number of positive documents that made the application according to the document given by the experts.FP is false positive that the document to be considered by the experts wrong application is true (the undesirable).FN is false negative that the document for the experts are right and wrong as by the application of (missing result).

A combination of precision and recall combined as ordinary harmonic mean , commonly called f-measure which can be in formulasikan as an equation (8).

$$F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (8)$$

F-measure system commonly used in the field of gathering information back to measure the classification of the search query classification of documents and performance. Previous research focused on f-measure to calculate the value of, but as with the development of large scale search engine, now more emphasis on performance f-measure precision and recall itself. So that more can be seen on the application as a whole.

2. THE CONTENT OF RESEARCH

2.1. Analysis of The Problem

The problem of this research is how classified information from social media particularly twitter with the consumers of telkom indihome into two classes are negative and positive.Then, the result of those served in graphical form.

2.2. System Analysis which will be built

The system which will be on the application of this research is used for analysis sentiment against

telkom indihome.Thus, the groove or proceedings of the system which will be built are as follows:

1. Taking Process Data
   The recovery of data testing and data training.The data is taken from social media twitter
2. Preprocessing Process
   Data training and data testing going through the process of text preprocessing who belongs to an early stage of text mining. Text processing it is aimed at preparing a document text that is not structured be structured data that ready to use for the next process.
3. Term Weighting Process
   Through a process preprocessing data obtained going through the weightings of the stage
4. Classification Process
   The various stages in the classification was intended to divide the data entering the class which have been determined so as to produce the results of sentiment analysis.

2.3. The Withdrawal of The Data Analysis

The tweets in this research data obtained by using API that provided by twitter .By using the API made an application to take data from twitter then stored the tweets into a database .

In the data, tweets API search, then bringing keyword-keyword associated with the Telkom Indihome combined with the sentiment.

Table 4 Example Words Sentiment

Kata Negatif	Lelet, lemot, lambat, parah, jelek, buruk, kecewa, payah
Kata Positif	Bisa, bagus, lancar, cepat, cepet, mantap, mantep, baik, kenceng, kencang, lancar, senang, canggih
Kata Netral	ikutan, selamat, kompetisi, dapatkan, gratis, kesempatan, info, nikmati
Emoticon Negatif	:-(, :(, =(, ;(
Emoticon Positif	:-), :), =), :D
Nama produk	Indihome, speedy, useetv, telkomcare

Tabel 5 Example *Tweet*

Tweet
Koneksinya speedy galau nih :(gimana ini @TelkomCare
@TelkomCare Jaringan WIFI speedy kenapa lemot sudah dua hari, mengganggu aktivitas saya.
Asik nih UseeTV bisa playback acara hari kemarin :D
Mau hujan pun, streaming youtube lancar, speedy emang mantap :)
Download film di ganool.com pake speedy cepet banget.
Gak Asik dari kemarin koneksi speedy lelet parah :(#speedyLELET
@TelkomPromo Raih silatuhrahmi 100% di hari yg fitri dgn GRATIS nelpon 1000 menit lokal/Interlokal indihome. Info: http://indihome.co.id .
Mau dimana2 jaringan speedy selalu lancar :')
@TelkomCare Speedy koneksi tidak bagus. gak bisa dengan restart modem saja.
Mau dimana2 jaringan speedy selalu lancar :')

2.4. Term Weighting Analysis

This stage is part of the weightings , that is done after the process of preprocessing.The weightings of a method that is used is a method of tf.idf . On this method term frequency (TF) to be multiplied by inverse document frequency (IDF).The formula used to express bobor (d) of the various documents on to the word documents a lock is on similarities (2) and (3) .

Table 6 Data Training Are Known

Kode	Data Latih	
	Data Latih	*Kelas*
D1	[koneksi] [galau] [kecewa]	Negatif
D2	[jaringan] [wifi] [lelet] [ganggu]	Negatif
D3	[download] [film] [lancar] [banget] [senang]	Positif
D4	[streaming] [youtube] [lancar] [mantap] [senang]	Positif
D5	[koneksi] [selalu] [lancar] [senang]	Positif

Table 7 Data Testing That Will Be Analyzed

Kode	Data Uji	
	Data Uji	*Kelas*
D6	[gak_asik] [koneksi] [lelet] [parah] [kecewa] [speedylelet]	?

Based on table 6 and 7 of the table, D1 to D6 is data that we will test the weight of the documents. D1 until D5 data is already known his class, while the D6 data not yet known to his class and to be tested. To determine what D6 into class. The first count the weighting of each term

Table 8 An Example Case The Application of The Term Weighting Stage

Term	TF						IDF	Wdt=TF.IDF					
	D6	D1	D2	D3	D4	D5	DF	LOG (n/df)	D6	D1	D2	D3	D4	D5
gak_asik	1						1	0.778	0.778	0	0	0	0	0
Koneksi	1	1				1	3	0.301	0.301	0.301	0	0	0	0.301
Lelet	1		1				2	0.477	0.477	0	0.477	0	0	0
Parah	1						1	0.778	0.778	0	0	0	0	0
Kecewa	1	1					2	0.477	0.477	0.477	0	0	0	0
speedylelet	1						1	0.778	0.778	0	0	0	0	0
Galau		1					1	0.778	0	0.778	0	0	0	0
Jaringan			1				1	0.778	0	0	0.778	0	0	0
Wifi			1				1	0.778	0	0	0.778	0	0	0
Ganggu			1				1	0.778	0	0	0.778	0	0	0
download				1			1	0.778	0	0	0	0.778	0	0
film				1			1	0.778	0	0	0	0.778	0	0
Lancer				1	1	1	3	0.301	0	0	0	0.301	0.301	0.301
Banget				1			1	0.778	0	0	0	0.778	0	0
Senang				1	1	1	3	0.301	0	0	0	0.301	0.301	0.301
Streaming					1		1	0.778	0	0	0	0	0.778	0
Youtube					1		1	0.778	0	0	0	0	0.778	0
Mantap					1		1	0.778	0	0	0	0	0.778	0
Selalu						1	1	0.778	0	0	0	0	0	0.778
									3.59	1.56	2.81	2.94	2.93	1.68

2.5. Application Analysis *Improve K-Nearest Neighbor*

After going through the process of document the weightings of going through the stage the classifications, in this process will be used algorithms improved k-nearest neighbor.The steps his steps are as follow:

Counting similaritas between two documents using methods cosine similarity (CosSim). Count resemblance a vector D6 document with every a document already team (D1, D2, D3, D4 and D5). Resemblance between documents can use cosine similarity.The formula is as follows:

$$Cos\left(\theta_{QD}\right) = \frac{\sum_{i=1}^{n} Q_i \cdot D_i}{\sqrt{\sum_{i=1}^{n}(Q_i)^2} \cdot \sqrt{\sum_{i=1}^{n}(D_i)^2}} \qquad (4)$$

Where :
$Cos(\theta_{QD})$ = Resemblance documents Q terhadap D
Q = Data Testing
D = Data Training
n = Of all of the data

To settle equation (4) could be divided into two the following steps:

1. Count the result of a scalar between D6 and D5 document that has been team.The multiplication of the results of each document with D6 totaled using formula equation (4) the upper part $\sum_{i=1}^{n} Q_i \cdot D_i$
2. Any count long documents including D6.The way is do weights every term in any documents the number of right squared the value and then root with using formulas equation (4) the lower part $\sqrt{\sum_{i=1}^{n}(Q_i)^2} \cdot \sqrt{\sum_{i=1}^{n}(D_i)^2}$

The left (WD6 * WD1) to the 9 represents the first step which wd6 it w of pembobotan equation (3), WD1 the trainer when pembobotan (3) and the right-hand side (of vectors) showed the second step.

Table 9 The Completion of Cosine Similarity

WD6*Wdi					Panjang Vektor					
D6XD1	D6XD2	D6XD3	D6XD4	D6XD5	D6	D1	D2	D3	D4	D5
0	0	0	0	0	0.606	0	0	0	0	0
0.0906	0	0	0	0.091	0.091	0.091	0	0	0	0.091
0	0.2276	0	0	0	0.228	0	0.228	0	0	0
0	0	0	0	0	0.606	0	0	0	0	0
0.2276	0	0	0	0	0.228	0.228	0	0	0	0
0	0	0	0	0	0.606	0	0	0	0	0
0	0	0	0	0	0	0.606	0	0	0	0
0	0	0	0	0	0	0	0.606	0	0	0
0	0	0	0	0	0	0	0.606	0	0	0
0	0	0	0	0	0	0	0.606	0	0	0
0	0	0	0	0	0	0	0	0.606	0	0
0	0	0	0	0	0	0	0	0.606	0	0
0	0	0	0	0	0	0	0	0.091	0.091	0.091
0	0	0	0	0	0	0	0	0.606	0	0
0	0	0	0	0	0	0	0	0.091	0.091	0.091
0	0	0	0	0	0	0	0	0	0.606	0
0	0	0	0	0	0	0	0	0	0.606	0
0	0	0	0	0	0	0	0	0	0.606	0
0	0	0	0	0	0	0	0	0	0	0.606
0.31826	0.22764	0	0	0.0906	2.3625	0.9238	2.0442	1.9978	1.9978	0.8774
					1.537	0.961	1.43	1.413	1.413	0.937
						0.215	**0.104**	**0**	**0**	**0.063**

Of calculation table 9 known the value of the Cosine Similiarity of D1, D2, D3, D4 and D5 is:

Tabel 10 The Values of Cosine Similiarity

D1	D2	D3	D4	D5
0.215	0.104	0	0	0.063

The next step is to 're sequencing the level of the resemblance of the data were obtained:

Table 11 The Order The Level of Resemblance

1	2	3	4	5
D1	D2	D5	D3	D4

Next on improved algorithms k-nearest neighbor, new k-values called with n. equation (5) explain about the proportion of the determination of k-values (n) in each category.

$$n = \left[\frac{k * N(c_m)}{maks\{N(c_m) \mid j=1.....N_c\}}\right] \qquad (5)$$

Where :
n = New k-values
k = k-values Set
$N(c_m)$ = The amount of data training in the category / category m
$maks\{N(c_m) \mid j=1.....N_c\}$ = The amount of data on training most of all categories.

The results of calculation of the value of n:

Table 12 Number of Data Training

Data Latih		
Positif	Negatif	Jumlah
3	2	5

$$n = \frac{3\ x\ 3}{3} = \frac{9}{3} = 3$$

$$n = \frac{3\ x\ 2}{3} = \frac{6}{3} = 2$$

Table 13 The Value of n (k-New)

	n (k baru)	
Nilai k	n Positif	n Negatif
3	3	2

A number of n documents selected in each category is a top n documents or document is a document that has top similaritas most in each of the category.

Known the order after the level of its resemblance take as many as k-values new (n) the most high levels of its resemblance to D6 class and set of D6. The results:

Table 14 The final result of the resemblance

1	2	3
D1	D2	D5
0.215	0.104	0.063
Negatif	Negatif	Positif

Finally, is set based on class D6 class that appear most. Because the majority of classes are emerging negative, then D6 in the a negative.

If there is a special case in which the value k taken have fulfilled and the same class that appears is, the documents noted the resemblance to a class of having the value of the most high.

2.6. Testing System

Testing the method is a process of testing on an algorithm classifications. The purpose of this testing to know whether there is a mistake while implement logic improved algorithms k-nearest neighbor.

Testing accuracy classifications tweets held to find out the level of accuracy of classification that tweets be done manually tweets having a classification which is done by the system by using improved k-nearest neighbor. Testing was done using confusion matrix which is that an matrik of a prediction will be compared with the class who is a native from the data put. Testing carried out using 20 sample tweets. To more scenario he explained will be presented in table the following:

Table 15 Sample Testing Classifications The Tweets

No	Tweet	Sentimen	
		Kasifikasi Manual	Klasifikasi Sistem
1	@TelkomCare min saya indihome 10 mbps fiber kok lemot ya ? tiba tiba jadi ke 2 mbps ? kenapa mmin ? ping pun jadi 70 ,	N	N
2	Dan janji telemarketer 147 buat naikin kecepatan internet dgn ganti pake indihome jg ga kejadian. Di test masi lemot juga ni wifi □	N	N
3	@TelkomCare indihome saya bermasalah terus beberapa hari ini, sinyal lambat dan koneksi internet sering putus.	N	P
4	@TelkomCare Mohon segera perbaiki koneksi Indihome kami, sudah satu minggu lebih tidak bisa digunakan.	N	P
5	@wifi_id min saya mau register wifi.id tapi bingung no internet indihome saya langganan spedy ada koneksi wifi.id nya penjelasannya	P	N
6	Ga bisa pake periscope dgn koneksi indihome, pake 3G malah bisa... @TelkomCare	N	N
7	@TelkomCare indihome seharian jelek nih.. kadang nyambung kadang ngga.. lampu indikator internet jg merah mulu..	N	N
8	@ardynp saya pernah pake @transvision_ind sebelum pindah ke indihome @TelkomPromo @UseeTV_Cable , oke juga.	P	P
9	Tadi blm nonton . Coba ILK dulu ah di Tipiiii. Pakai fitur playback @UseeTVcom nya IndiHome.#IamIndihome #KipasSakti @IndihomeFiber	P	N
10	@TelkomCare tolong dong dicek kan indihome gw, dari sore sampe sekarang gangguan nauzubilah, -_- daerah pontianak	N	N
11	@kih00ns Iya sama2 Kak, semoga lancar terus dan tidak ada gangguan lagi. ^AY	P	P
12	@TelkomCare internetnya lancar jaya	P	P
13	@TelkomSpeedy ga **bagus** penanganan gangguannya, gmn mau masuk ke market **indohome**?! Haha nanganin gangguan rusak modem aja sampe 4hr ga jelas!	N	N
14	Penasaran sama speedy. Uda bayar tapi koneksi antara lemot atau putus nyambung. Mau banting HP, sayang Hp blm ada sebulan umumta :'(	N	N
15	Sedikit kecewa dengan pelayanan telkom speedy	N	N
16	@Adityaa_Prtm IndiHome masup di daerah lo nggak? temen2 gw pada bilang **bagus** sih	P	P
17	Layanan Indihome dari @TelkomIndonesia sangat bagus, puas deh :)	P	P
18	Sejauh ini puas dengan @TelkomPromo yaitu indihome. Internet cepat plus tv yang mumpuni. Masakinian banget dehh.. Makasih telkom :)	P	P
19	Sejauh ini puas dengan @TelkomPromo yaitu indihome. Internet cepat plus tv yang mumpuni. Masakinian banget dehh.. Makasih telkom :)	P	p
20	@TelkomCare kalau telepon tdk tau knp tidak dipasang, tp usee tv msh berjalan dgn lancar	P	P

Desc: P (Positive), N (Negative)

The following table of confuion the matrix:

Table 16 *Confusion Matrix*

	Positive	Negative
Positive	8	2
Negative	2	8

After the system do classifications, then count precision, recall and its accuracy based on equation (6) and (7).

$$Precision = \frac{8}{8+2} = 0{,}80$$

$$Recall = \frac{8}{8+2} = 0{,}80$$

$$F - Measure = 2 * \frac{0{,}80 * 0{,}80}{0{,}80 + 0{,}80} = 0{,}80$$

The testing used in the 15 uses sample tweet, tweet about 20 The test which has been done can see that there some precision analysis of influencing sentiment by using methods improved k-nearest neighbor. Based on test precision, recall and f-measure, we get the results of the analysis f-measure the tweets sentiment using improved k-nearest neighbor as much as 80 % of 80 %, with precision and recall by 80 %.

2.7. Implementation of Interface

Following a display interface is on this application.

1. **Interface Home**



Image 2 Display Interface Home

2. **Interface Crawl Tweet**

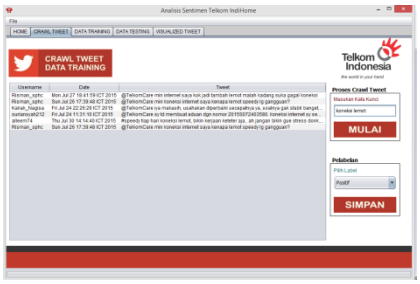


Image 3 Display Interface Crawl Tweet

3. **Interface Data Training**

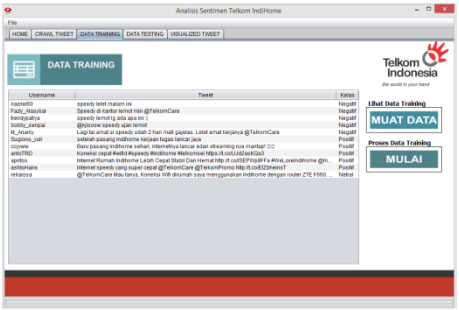


Image 4 Display Interface Data Training

4. **Interface Data Testing**

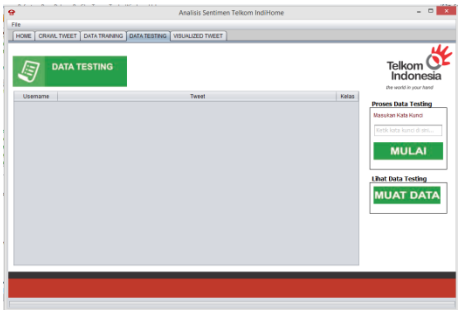


Image 4 Display Interface Data Testing

5. **Interface Visualized Tweet**

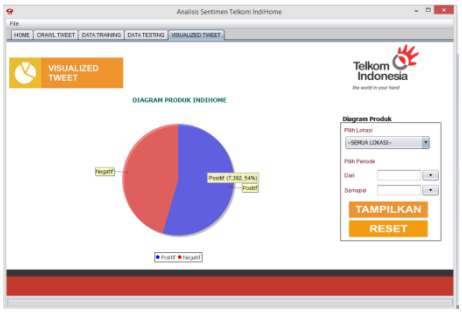


Image 4 Display Interface Visualized Tweet

3. CLOSING

3.1. Conclusion

From the research that has been carried out it can be seen that improved algorithms k-nearest neighbor can classify a opinions in the form of the tweets into two classes that is positive and negative accurately. The level of accuracy from such classification strongly influenced by the process of training. So that we can conclude from the results of the classifications presented in the form of a chart in visualized the tweets can be seen clearly information public sentiment against a product indihome and can be used as an evaluation Telkom Indihome in order to further improve the quality of its service so that it can improve and determine the steps business afterward that is better.

3.2. Suggestion

As for the advice of this research is as follows:

1. Needed further research or research development to use a method of analysis the classification of another such sentiments weighted k-nearest neighbor or other methods combine improved method by method k-nearest neighbor who can better k-nearest neighbor of a method that improved analysis the classification obtained the results of sentiments better and more accurate.
2. On further research is expected can recognize sentence sarcasm as “Connection indihome “lancaaarr” once until just know it will be difficult to browse :)”.
3. When doing the weightings in this research , counting system based on said resemblance frequency of occurrence, so as to have optimal results should be used a system that can check said synonyms.

BIBLIOGRAPHY

[1] https://dailysocial.net/post/kemenkominfo-targetkan-pengguna-internet-di-indonesia-tahun-2015-capai-150-juta-orang
[2] http://tekno.liputan6.com/read/2164377/pengguna-internet-indonesia-kuasai-media-sosial-di-2015
[3] http://tekno.liputan6.com/read/2164377/pengguna-internet-indonesia-kuasai-media-sosial-di-2015?p=1
[4] Iwan Arif, Text Mining http://lecturer.eepis-its.edu/~iwanarif/kuliah/dm/6Text%20Mining.pdf
[6] B. P. a. L. Lee, "Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval," vol. 2, no. 1-2, pp. 1-135, 2008.
[7] Fahrur Rozi Imam, "Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi", 2012
[8] Yusuf Nur Muhammad dan Santika D. Diaz "ANALISIS SENTIMEN PADA DOKUMEN BERBAHASA INDONESIA DENGAN PENDEKATAN SUPPORT VECTOR MACHINE" 2011
[9] Raymon J. Mooney. CS, Machine Learning Text Categorozation, 2006
[10] L. Vogel, "Java Regex - Tutorial, Vogella,," 14 Januari 2014.
[11] Sunni Ismail "Analisis Sentimen dan Ekstraksi Topik PenentuSentimen pada Opini Terhadap Tokoh Publik" volume 1, nomor 2, 2012
[12] Utomo manalu Boy, "Analisis Sentimen Pada Twitter Menggunakan teks mining" 2014
[13] Arfianda Putri Prima "IMPLEMENTASI METODE IMPROVED K-NEAREST NEIGHBOR PADA ANALISIS SENTIMEN TWITTER BERBAHASA INDONESIA"
[14] Kroenke M. David "Database Processing Jilid 1" edisi 9, 2005
[15] Prodase Labolarotium, "Object-Oriented Programming Module" 2013/2014
[16] Dwiyoga Tahitoe Andita “Implementasi Modifikasi Enhanced Confix Stripping Stemmer Untuk Bahasa Indonesia Dengan Metode Corpus Based Stemming”,
[17] Ngesti Waluyo Catur, “Confix Stripping Stemmer”, 2012.