Microblogging Analysis for Determining Public Policy Priority Based on Public Opinion Using Naïve Bayes and Analytical Hierarchy Process Algorithm

Mohammad Khoiron
Telematika-CIO, Electrical Engineering Department
Faculty of Industrial Technology
Institut Teknologi Sepuluh Nopember Surabaya
e-mail:m.khoiron.w@gmail.com

Surya Sumpeno
Multimedia and Networking Engineering Department
Faculty of Industrial Technology
Institut Teknologi Sepuluh Nopember Surabaya
e-mail:surya@ee.its.ac.id

Adhi Dharma Wibawa
Multimedia and Networking Engineering Department
Faculty of Industrial Technology
Institut Teknologi Sepuluh Nopember Surabaya
e-mail:adhiosa@te.its.ac.id

Abstract
The main task of a government is making and implementing public policy, and also evaluating the public policies that have been made . Often all three tasks can not satisfy the expectations of the wider community because it is arranged not based on the aspirations of a society where the government is located. Determination of public policy is more likely to consider the political aspects and the interests of a certain elite.
By seeing that problems, it is necessary to find the rapid and inexpensive solution for obtain data about what expectations is desired by the community towards a public policy. This can be obtained from the microblogging analysis, by monitoring issues of public policy that are discussed by people in the media microblogging, within a certain time.
Analysis was performed using Naïve Bayes algorithm to classify whether an opinion delivered by the public through the microblogging has a negative, positive, or neutral sentiment. Results from the classification used to determine the priority of public policy using Analytical Hierarchy Process (AHP) algorithm, which became the reference for making a public policy that is expected to satisfy the justice and public expectations.

Key Word: *Public Policy, Public Policy Priority, Sentiment Analysis, Clasification, Naïve Bayes, Analytical Hierarchy Process*

a. Introduction

Characteristic of democratic modern society is the involvement of the community in taking a public policy. The community involvement began since the government planning until implementing the public policy. Community involvement is necessary because public policy will affect their daily lives. Therefore, a democratic government should always involve the community in determining public policy.

In Indonesia now, people look more enthusiastic in discussing a public policy generated by the government. Such enthusiasm is very positive as far as to provide another perspective for the benefit of society. Public debate marks the dynamics of a society. The amount of community involvement can not be separated from the reform era that is still kept rolling with a wide range of dynamics and risks.

One kind of media that used frequently to express public opinion is microblogging social media. At this time microblogging site such as Twitter, Tumblr, and Facebook has become a very popular means of communication among Internet users, where millions of messages appear every day.

Free message format and ease of access from various platforms, making Internet users tend to switch from blogs or mailing to the microblogging service. This has caused many users are posting about a product and services that they use, to express their views on politics and religion, also criticize a public policy.

Twitter as a microblogging site with over 500 million users and 400 million tweets per day, allowing users to share the message using short text called tweets. Twitter can be a data source of the opinion and public sentiment, and then that data can be used efficiently for marketing or social studies.

In this paper will be discussed about twitter microblogging sentiment analysis using Naïve Bayes algorithm which may be utilized as consideration for determining the priority of public policy by using Analytical Hierarchy Process algorithms, so the quality of the policy are expected to fulfill the expectations and desires of the community.

b. Literature Review

- Public Policy

Public policy as a part of the political decision is a rules made by the government to solve the various problems and issues in society. Public policy is also a decision made by the government to perform certain actions between to do or no to to do something.

In a society that is in the jurisdiction of a country often occurs various problems, and the government which holds full responsibility for the lives of the people should be able to resolve these issues. Public policy which is made and issued by the state is expected to be a solution to these problems. Public policy is a decision made to overcome the problems in a particular activity undertaken by the government in the framework of governance (Mustopadidjaja , 2002).

- Naïve Bayes Classifier

Naive Bayes classifier is an algorithm used to find the value of the highest probability to classify the test data to the most appropriate category (Feldman and Sanger 2007). In this research, the test data is a Tweet documents. There are two stages in document classification. The first stage is the training of the documents that have been known the category, and then the second stage is the process of classifying documents of unknown category.

In a naïve Bayes classifier algorithm each document is represented by a pair of attributes "x1, x2, x3, ... xn" which is x1 is the first word, x2 is the second word, and so on, while V is the set of Tweet categories.

In the process of classification algorithm will search for the highest probability of all the document categories that were tested (VMAP), where the equation is as follows:

$$V_{MAP} = \underset{V\,jeV}{\arg max} \frac{P(x_1, x_2, x_3, \dots x_n|V_j)P(V_j)}{P(x_1, x_2, x_3, \dots x_n)} \qquad \text{(b.1)}$$

For P (x1, x2, x3, ... xn) is constant for all categories (Vj) so that the equation can be written as follows:

$$V_{MAP} = \underset{V\,jeV}{\arg max} P(x_1, x_2, x_3, \dots x_n|V_j)P(V_j) \qquad \text{(b.2)}$$

The equation can be simplified as follows:

$$V_{MAP} = \underset{VejV}{arg\,max} \prod_{i=1}^{n} P(x_i\,|V_j)P(V_j) \qquad \text{(b.3)}$$

Description:

Vj : Tweet category j =1, 2, 3,…n, which in the research

j1	: negative sentiment tweet category
j2	: positive sentiment tweet category
j3	: neutral sentiment tweet category
P(xi|Vj)	: xi probability in Vj category
P(Vj)	: Vj probability

For P (Vj) and P (xi|Vj) that calculated at the time of training, the equation is as follows:

$$P(V_j) = \frac{|docs\ j|}{|contoh|} \qquad \text{(b.4)}$$

$$P(x_i|V_j) = \frac{n_k+1}{n+|kosakata|} \qquad \text{(b.5)}$$

Description:

|docs j|	: the number of document at each j category
|contoh|	: the number of document of all category
nk	: the number of occurence frequency of each word
n	: the number of occurence frequency of each word from each category
|kosakata|	: the number of words from all categories

- **Analytical Hierarchy Process**

AHP (Analytical Hierarchy Process) is a decision support system that decompose a complex multi-factor problem into a hierarchy, where each level is formed of specific elements. The main equipment AHP is a functional hierarchy with the main input is human perception. The existence of a hierarchy allows complex or unstructured problem is divided into sub- problems, then compile them into a form of hierarchy (Kusrini, 2007).

Decision makers involved to provide consideration in determining the relative importance of these factors. The general objective of the decision to be taken is located on the top of the hierarchy, while the criteria and alternative decision at a lower level sequentially. The AHP stages are as follows:

1. The establishment of a hierarchy
   Hierarchy is a structure tree that is used to represent the spread of influences ranging from goals down to the structure located at the most basic level
2. Pairwise Comparison
   Step in AHP involves estimating the weighting priority of a set of criteria or alternatives of a square matrix used in pairwise comparisons A = [aij], in which the weight value must be positive and if policies regarding pairwise comparison is completely consistent then made a reverse comparison of that value, for example: aij = 1/aij for all i, j = 1, 2, 3, ..., n.
   Furthermore , the final weight of the wi as a i-th factor that has been normalized, is as follows:

$$w_{ij} = a_{ij} / \left(\sum_{i=1}^{n} a_{ij} \right) \qquad \forall\, i = 1, 2, \dots, n \qquad \text{(b.6)}$$

Pairwise comparisons scale for the relative importance is assessing in a comparative degree of importance between an element with another element. A comparative scale used in AHP according Kusrini are:

Value	Description
1	Criteria / alternative A as important as the criteria / alternative B
3	A little more important than B
5	A clearly more important than B
7	A very clearly more important than B
9	A absolutely more important than B
2,4,6,8	When hesitating between two adjacent values

Table 1. Comparison Scale (Source: Kusrini, 2007:134)

3. Consistency checking
   Check whether the pairwise comparisons were made based on a policy decision remains within specified limits or not. Consistency measurement naturally or deviation of consistency called consistency index (CI), which is defined as follows:

$$CI = \frac{\lambda_{\max} - n}{n-1} \qquad \text{(b.7)}$$

Consistency Index of a inverse comparison matrix from scale 1 to 9 which is generated randomly, with the inverteed comparison results, for each size of the matrix is called the Random Index (RI) shown in the following table:

Order Matrix	RI value	Order Matrix	RI value	Order Matrix	RI value
1,2	0,00	5	1,12	8	1,41
3	0,58	6	1,24	9	1,45
4	0,90	7	1,32	10	1,49

Table 2. List of Random Index (Source: Kusrini, 2007:136)

So that the consistency ratio (CR) is defined as the ratio between the CI and RI for the same order matrix

CR = CI/ RI (b.8)

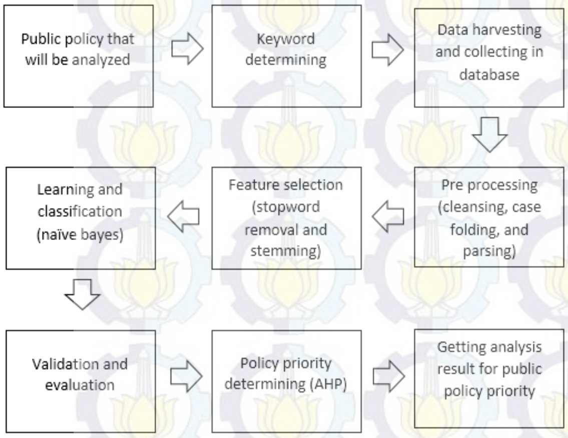
CR < 0.1 then the policy is acceptable. If the CR value more than 0.1, the leader necessary to review the measures taken.

4. Overall weight evaluation
   Weighting of each critera that has been obtained is multiplied by the value of the criteria for each alternative so the best alternative is an alternative that has the highest priority

5. Group decision-making / establishing policies
   To produce policy outcomes of the group, each member of the group makes its own policies to copy model they have and then combining the results

c. Methodology

- Flowchart of Research Methodology


Public policy that will be analyzed
Keyword determining
Data harvesting and collecting in database
Pre processing (cleansing, case folding, and parsing)
Feature selection (stopword removal and stemming)
Learning and classification (naïve bayes)
Validation and evaluation
Policy priority determining (AHP)
Getting analysis result for public policy priority

Picture 2. Flowchart of Research Methodology

- Determining Public Policy

Public policy has a very broad sphere, so it is necessary for an example of public policy that can be used to simulate the prioritization of public policies on the terms of public opinion that comes from Twitter. For example, the priority of public policy that will be made in this research are the MDGs (Millennium Development Goals) that has eight goals.

- Determining Keywords

After determining public policy priorities that will be made, the next step is selecting the keywords that can represent each predetermined policy. Keywords used to search the public opinion via Twitter which are expected consistent with the public policy that has been set. Here is a list of keywords and the public policies represented in Bahasa Indonesia:

No	Public Policy based on MDGs	Keywords
1.	Memberantas Kemiskinan dan Kelaparan Ekstrem	Kemiskinan Kelaparan
2.	Mewujudkan Pendidikan Dasar untuk Semua	Pendidikan Buta huruf
3.	Mendorong Kesetaraan Gender dan Pemberdayaan Perempuan	Kesetaraan gender Pemberdayaan perempuan
4.	Menurunkan Angka Kematian Anak	Kematian bayi Imunisasi
5.	Meningkatkan Kesehatan Ibu	Kesehatan ibu Kesehatan reproduksi
6.	Memerangi HIV dan AIDS Malaria Serta Penyakit Lainnya	Cegah HIV Cegah penyakit
7.	Memastikan Kelestarian Lingkungan	Keanekaragaman hayati Kelestarian lingkungan
8.	Mengembangkan Kemitraan Global untuk Pembangunan	Akses internet Perdagangan bebas

Table 3. List of Keywords

- Data Harvesting

The process tweet data harvesting done by utilizing the Twitter Streaming APIs. Searching and collecting of public opinion in Twitter made within two mobths based on keywords that are predefined. Data obtained from the results of harvesting are stored into a database.

- Pre Processing

Before doing the feature selection process of the tweet has been obtained and to obtain more accurate results for tweet sentiment analysis, preprocessing of the exixsting tweet data need to be done, which includes:

1. Cleansing
   Things done in the cleansing process includes the removal of a URL , @mention , #hashtags and delimiter (alphanumeric characters and symbols)
2. Case Folding
   At this stage, all uppercase characters converted to lowercase
3. Parsing
   This is the stage where a tweet or a sentence is separated into words

- Feature Selection

Feature selection is done before the process of learning and classification. There are two processes at this stage, namely:

1. Stop Word Removal
   Elimination of vocabulary that is not a characteristic (unique word) of a document (eg: "di", "oleh", "pada", "sebuah", "karena")
2. Stemming
   Process mapping and decomposition of various forms (variants) of a word to its basic word (stem), by removing the particle-particle whether it be prefixes , suffixes , and infixes that exist in every word.

- **Learning and Classification**

From the feature selection that has been done, the next thing is learning process and classification using Naïve Bayes algorithm which is divided into two stages:

1. First stage
   Training of tweet documents that have been known the category (negative or positive sentiment, or neutral).
2. Second stage
   The process of document classification with the unknown categories (negative or positive sentiment, or neutral).

- **Validation and Evaluation**

This stage is necessary to validate and evaluate the extent of the learning process and classification accuracy by using Naïve Bayes algorithm that has been done.

- **Determining Priorities of Policy**

From the analysis of tweet sentiment using Naïve Bayes algorithm which has been obtained, the next process is determining the priority of public policy by using Analytical Hierarchy Process (AHP) algorithm, which the hierarchical structure is formed of a number of positive sentiment tweets, the number of negative sentiment tweets, the number of neutral tweets, the number of retweets, the number of tweets in the form of questions, and tweet that is not a retweet and question (direct tweet) for each public policy.

d. Result and Analysis

- **Data Harvesting**

Tweet Data were collected between June and July 2015 with the following results:

No.	Keywords	Number of tweets
1.	Kemiskinan	50226
	Kelaparan	72404
2.	Pendidikan	96060
	Buta huruf	10350
3.	Kesetaraan gender	1636
	Pemberdayaan Perempuan	7444
4.	Kematian bayi	1766
	Imunisasi	11620
5.	Kesehatan ibu	3609
	Kesehatan reproduksi	2712
6.	Cegah HIV	345
	Cegah penyakit	4082
7.	Keanekaragaman hayati	2509
	Kelestarian lingkungan	1703
8.	Akses internet	12804
	Perdagangan bebas	2854
	Total Tweet	**282124**

Table 4. Number of tweets on each keyword

- **Training Data**

From the result of tweet harvesting, will be taken 3000 tweet that will be used as training data. Retrieving training data doing by considering the percentage of acquisition of each keyword so that there are elements of representation. Furthermore, the training data is labeled manually to classify in a tweet that has a negative or positive sentiment, or neutral.

- **Pre Processing**

From the training data as much as 3,000 tweets, pre processing stage need to be done with the following stages:

1. Cleansing
2. Case Folding
3. Parsing

- **Feature Selection**

The next step is selecting a feature on the training data that has been through the pre processing stage. The process at this stage is:

1. Stop Word Removal
   In this process a list of words that have no meaning will be removed from a training data tweet document. A list of words that have no meaning obtained from the research results of Tala (*Tala, F. Z. (2003)*)

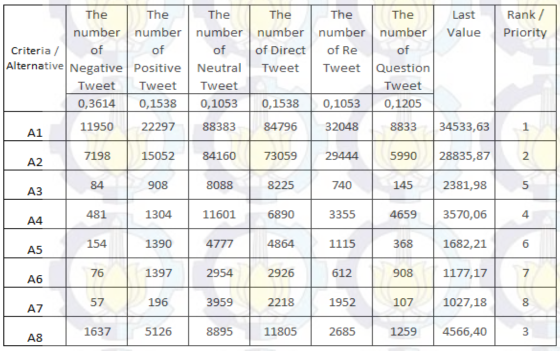
2. Stemming
   Training data tweet document that have been through the process of stop word removal is processed using PHP library of Sastrawi which is based on stemming algorithm of Nazief and Andriani.

- **Learning and Classification**

From the feature selection that has been done, the next step is doing learning process and classification using naïve Bayes algorithm which is divided into two stages:

1. First stage
   By using the WEKA software, training of tweet document training data that has been known the categories obtained 73.8 % accuracy using Naïve Bayes algorithm and features of the TF - IDF.
2. Second stage
   - Furthermore, the unknown category tweet document will be classified.
   - To get a direct tweet, retweet and tweet question conducted by filtering based on the characters '?' and 'RT @'

Results from the overall classification and filtering of tweets shown in the table:



Criteria / Alternative	The number of Negative Tweet	The number of Positive Tweet	The number of Neutral Tweet	The number of Direct Tweet	The number of Re Tweet	The number of Question Tweet	Last Value	Rank / Priority
	0,3614	0,1538	0,1053	0,1538	0,1053	0,1205		
A1	11950	22297	88383	84796	32048	8833	34533,63	1
A2	7198	15052	84160	73059	29444	5990	28835,87	2
A3	84	908	8088	8225	740	145	2381,98	5
A4	481	1304	11601	6890	3355	4659	3570,06	4
A5	154	1390	4777	4864	1115	368	1682,21	6
A6	76	1397	2954	2926	612	908	1177,17	7
A7	57	196	3959	2218	1952	107	1027,18	8
A8	1637	5126	8895	11805	2685	1259	4566,40	3

Table 5. Classification and Filtering Results

Description:

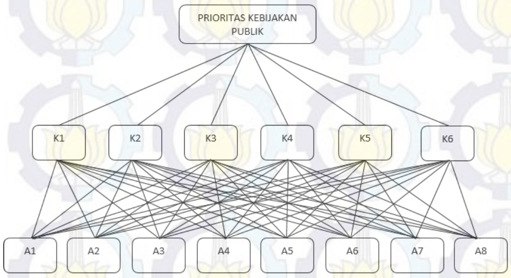
A1	Memberantas Kemiskinan dan Kelaparan Ekstrem
A2	Mewujudkan Pendidikan Dasar untuk Semua
A3	Mendorong Kesetaraan Gender dan Pemberdayaan Perempuan
A4	Menurunkan Angka Kematian Anak
A5	Meningkatkan Kesehatan Ibu
A6	Memerangi HIV dan AIDS Malaria Serta Penyakit Lainnya
A7	Memastikan Kelestarian Lingkungan
A8	Mengembangkan Kemitraan Global untuk Pembangunan

Table 6. Policy Priorities Alternative

- **Determination of Public Policy Priorities**

From the data that has been obtained in the preceding stage, determining public policy priorities algorithms using Analytical Hierarchy Process (AHP) can be done with the steps as below:

1. **The establishment of a hierarchy**



Picture 7. The Establishment of A Hierarchy

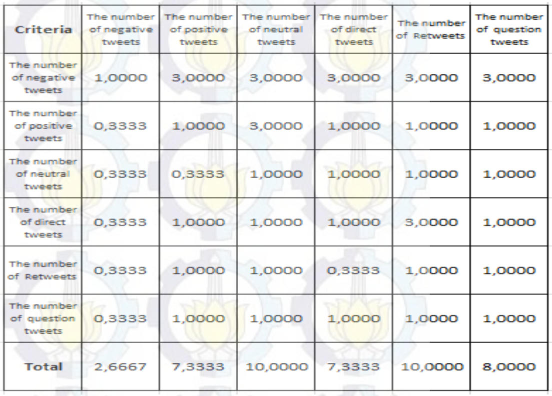
Description:
K1= The number of negative tweets
K2= The number of positive tweets
K3= The number of neutral tweets
K4= The number of direct tweets
K5= The number of re-tweets
K6= The number of question tweets

2. **Pairwise Comparison**

The main objective of this study is, to make a ranking of public policy based on public opinion towards a public policy that is most negative. Hereafter devised pairwise comparison matrix with the following criteria:

a) The number of negative tweets little more important than the number of positive tweets.
b) The number of negative tweets little more important than the number of neutral tweets.
c) The number of negative tweets little more important than the number of direct tweets.
d) The number of negative tweets little more important than the number of re-tweets.
e) The number of negative tweets little more important than the number of question tweets.
f) The number of positive tweets little more important than the number of neutral tweets.
g) The number of direct tweets little more important than the number of question tweets.

With reference to the 1-9 scale Saaty, L Thomas, pairwise comparison matrix can be made as shown in the table:

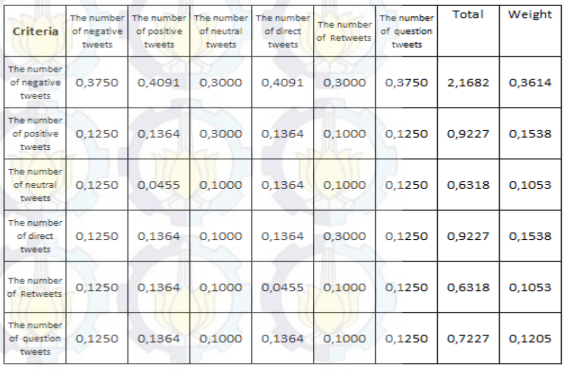


Criteria	The number of negative tweets	The number of positive tweets	The number of neutral tweets	The number of direct tweets	The number of Retweets	The number of question tweets
The number of negative tweets	1,0000	3,0000	3,0000	3,0000	3,0000	3,0000
The number of positive tweets	0,3333	1,0000	3,0000	1,0000	1,0000	1,0000
The number of neutral tweets	0,3333	0,3333	1,0000	1,0000	1,0000	1,0000
The number of direct tweets	0,3333	1,0000	1,0000	1,0000	3,0000	1,0000
The number of Retweets	0,3333	1,0000	1,0000	0,3333	1,0000	1,0000
The number of question tweets	0,3333	1,0000	1,0000	1,0000	1,0000	1,0000
Total	2,6667	7,3333	10,0000	7,3333	10,0000	8,0000

Table 7. Pairwise Comparison Matrix

3. **Pairwise Comparison Matrix Normalization**

The next phase is to normalize the pairwise comparison matrix by dividing each value in the column matrix with the sum of the corresponding column



Criteria	The number of negative tweets	The number of positive tweets	The number of neutral tweets	The number of direct tweets	The number of Retweets	The number of question tweets	Total	Weight
The number of negative tweets	0,3750	0,4091	0,3000	0,4091	0,3000	0,3750	2,1682	0,3614
The number of positive tweets	0,1250	0,1364	0,3000	0,1364	0,1000	0,1250	0,9227	0,1538
The number of neutral tweets	0,1250	0,0455	0,1000	0,1364	0,1000	0,1250	0,6318	0,1053
The number of direct tweets	0,1250	0,1364	0,1000	0,1364	0,3000	0,1250	0,9227	0,1538
The number of Retweets	0,1250	0,1364	0,1000	0,0455	0,1000	0,1250	0,6318	0,1053
The number of question tweets	0,1250	0,1364	0,1000	0,1364	0,1000	0,1250	0,7227	0,1205

Table 8. Pairwise Comparison Matrix Normalization

4. Consistency Ratio Checking (CR)

A consistency check is required to see whether the pairwise matrix that we have created a consistent value. It is fulfilled if the value of CR <= 0.1

Maximum Eigen Value

ƛmaks= 6,2889

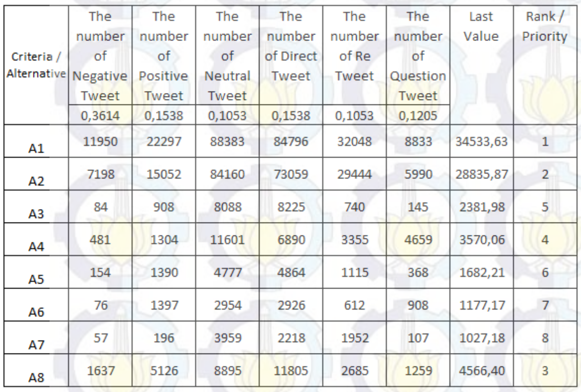
Consistency Index Value (CI)

CI=(λmaks–n)/(n-1)
CI= 0,057777778

Consistency Ratio Value (CR)

RI value taken from the Random Index Table. The value for matrix which has orders for 6 is = 1.24
CR=CI/RI
CR= 0,046594982 (CR value <=0,1 so it is cosistence)

5. Weight Evaluation



Criteria / Alternative	The number of Negative Tweet	The number of Positive Tweet	The number of Neutral Tweet	The number of Direct Tweet	The number of Re Tweet	The number of Question Tweet	Last Value	Rank / Priority
	0,3614	0,1538	0,1053	0,1538	0,1053	0,1205		
A1	11950	22297	88383	84796	32048	8833	34533,63	1
A2	7198	15052	84160	73059	29444	5990	28835,87	2
A3	84	908	8088	8225	740	145	2381,98	5
A4	481	1304	11601	6890	3355	4659	3570,06	4
A5	154	1390	4777	4864	1115	368	1682,21	6
A6	76	1397	2954	2926	612	908	1177,17	7
A7	57	196	3959	2218	1952	107	1027,18	8
A8	1637	5126	8895	11805	2685	1259	4566,40	3

Table 9. Weight Evaluation Table

From the weight evaluation shows that the order or priority of public policy that can be taken is based on Analytical Hierarchy Process (AHP) algorithm is as follows:

1. Eradicate Extrem Poverty and Hunger
2. Achieve Universal Primary Education
3. Global Partnership for Development
4. Reduce Child Mortality
5. Promote Gender Equality and Empower Women
6. Improve Maternal Health
7. Combat HIV/AIDS, Malaria, and other Diseases
8. Ensure Environmental Sustainability.

e. Conclusion

This research proved that microblogging analysis may be taken into consideration and studies to determine the priority of a public policy that is closer to the aspirations and desires of the community.

Besides that, it can be seen also that the public is very easy to give their opinion on matters that affect their daily lives, evidenced by the problem of poverty and education ranked number one and two in the tweet acquisition that correlated with the rating of public policy priorities.

The data presented in this research are preliminary results that could still be improved. The author still want to try to improve classification accuracy by using another methods and features that better.

f. Bibliography

[1] Berry, M.W. & Kogan, J. 2010. Text Mining Aplication and theory. WILEY : United Kingdom.

[2] Feldman, R & Sanger, J. 2007. The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press : New York.

[3] Han, J & Kamber, M. 2006 Data Mining: Concepts and Techniques Second Edition. Morgan Kaufmann publisher : San Francisco.

[4] Kusrini. 2007. Konsep dan Aplikasi Sistem Pendukung Keputusan. Yogyakarta : Andi.

[5] Nazief dan Adriani. 1996. Confix Stripping : Approach to Stemming Algorithm for Bahasa Indonesia.Technical report,Faculty of Computer Science, University of Indonesia,Depok, 1996

[6] Pang, B., Lee, L., & Vithyanathan, S. (2002). SentimentClassification Using Machine Learning Techniques. Dalam Proceedings of The ACL-02 conference on Empirical methods in natural language processing, pp. 79-86. Stroudsburg: Association for computationalLinguistic.

[7] Prasad, S. 2011. Micro-blogging Sentiment Analysis Using Bayesian Classification Methods.

[8] Sunni, I. & Widyantoro, D. H. 2012. Analisis Sentimen dan Ekstraksi Topik PenentuSentimen pada Opini Terhadap Tokoh Publik

[9] Tala, F. Z. (2003). A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. M.S. thesis. M.Sc. Thesis. Master of Logic Project. Institute for Logic, Language and Computation. Universiteti van Amsterdam The Netherlands

[10] Wang, A. H. 20100. Don't Follow Me: Twitter Spam Detection. Proceedings of 5th International Conference on Security and Cryptography (SECRYPT) Athens 2010: pp. 1-10. California:IEEE.