EMITTER International Journal of Engineering Technology
Vol. 5, No. 2, December 2017
ISSN: 2443-1168

Classification of Radical Web Content in Indonesia using Web Content Mining and k-Nearest Neighbor Algorithm

Muh.Subhan, Amang Sudarsono, Ali Ridho Barakbah

Electronics Engineering Polytechnic Institute of Surabaya
Jl. Raya ITS – Kampus ITS Sukolilo Surabaya 601111, Indonesia
Telp: 6231 5947280 Fax: 6231 5946114
Email: subhan@polinef.id, {amang, ridho}@pens.ac.id

Abstract

Radical content in procedural meaning is content which have provoke the violence, spread the hatred and anti nationalism. Radical definition for each country is different, especially in Indonesia. Radical content is more identical with provocation issue, ethnic and religious hatred that is called SARA in Indonesian languange. SARA content is very difficult to detect due to the large number, unstructure system and many noise can be caused multiple interpretations. This problem can threat the unity and harmony of the religion. According to this condition, it is required a system that can distinguish the radical content or not. In this system, we propose text mining approach using DF threshold and Human Brain as the feature extraction. The system is divided into several steps, those are collecting data which is including at preprocessing part, text mining, selection features, classification for grouping the data with class label, simillarity calculation of data training, and visualization to the radical content or non radical content. The experimental result show that using combination from 10-cross validation and k-Nearest Neighbor (kNN) as the classification methods achieve 66.37% accuracy performance with 7 k value of kNN method [1].

Keywords: K-NN, Nearest Neighbour, Radical Content, Indonesia

1. INTRODUCTION

The negative side of the development of information technology is the emergence of crimes committed using the means of information technology. The number of crimes each day continues to increase. Now began to emerge various types of crime with a new dimension that increasingly difficult to overcome such as computer abuse, banking crime and others. Internet abuse for terrorist acts is known as cyber terrorism. Cyber terrorism [2] is a form of politically motivated terrorism that uses technology and information, computer networks and technical infrastructure to undermine. Due to the importance of internet networks, some experts argue that cyber terrorism is more dangerous than traditional terrorists [2][3][4].

The definition of radical content handling in every country is very different, until now still a debate of multi-disciplinary researchers, making it difficult to define the label criteria of radicalism itself. In Indonesia, radical content is more identical and often associated with Ethnic and Religious issues (SARA). The emergence of government policies to block sites that are considered radical cause a response from organizations, politicians, religions, governments, and all Indonesian citizens who feel the impact of blocking sites. Radical content in the procedural sense provokes violence, spreads hatred and anti nationalism. According to the National Agency for the Elimination of Terrorism (BNPT), radical content is encouraging, provoking people to commit violence in the name of religion, interpreting jihad as a suicide bomb and saving the lives of others and excluding people outside their group.

Online detection of online radicalization is a challenging technical problem because of the large amount of data, unstructured, dynamic, noisy, and the difficulty of getting features that match to the radical criteria itself [3][4][5][6]. In this paper, we propose a web-based system using a web content mining approach, how to classify web content that can be categorized as radical content (SARA). Case studies utilize content that is blocked by the Positive Trust of the Ministry of Communication and Information of the Republic of Indonesia.

The contribution of this paper is that we apply the Nearest Neighbour Classification Algorithm combined with DF Threshold to group users based on similarity criteria of text search by measuring how closely the search results of the text with radical related content. Moreover, it could be also as a government preventive measure to prevent, detect early development web content related to radical content.

2. RELATED WORKS

Methods for detecting and analyzing online radicalism using Link Base Bootstrap Algorithm (LBB) semi-automatic approach techniques detect radicals have studied [7]. The workings of the LBB content on the modalities on the internet such as blogs, online forums identified the URL of sourced from authoritative sources. Then the URL is expanded using back search links and favorite links to accumulate related URLs. The idea is that extremist sites (or forums) link to each other and form a kind of community structure. Then expanded the search again by trying to find the URL that has been in the fruit domain. Crawlers to download and collect content. Flick web dark. Until the processing of text mining with pre-existing methods.

Chaurasia et al [4][8] found a new algorithm for network terrorization. The study about efficiency algorithm for destabilization of terrorises network as a system of prevention against terror attacks. In this research, the new method of approach (algorithm) is more effective than ever. The algorithm found in this claim can reveal the hidden hierarchy of network terrories by utilizing social network analysis and involving two methods of centrality massures (page rank, katz, defence centrality measures (DC).

A research done by building the system used an XGraphticClus method has been developed by combining mining hyperlink and web Content [9]. The workings of the developed system is to visualize web navigation to understand the behavior of users while visiting the web

Case detection linked to terrorist website using system detection terrorist interface (ATDS) is a development study of a user access detection system linked to a terrorist website development from the previous system about intrusion detection performed by the same researcher [10]. The workings of this system are operated in two modes: training mode and detection mode, traning mode determines the typical model of user interest or group on the web page accessed by the user over the time limit. While the detection mode works real time monitor and analyzes content within web pages the user access. The system works online capturing the user's Ip address, analyzing the content pages, extracting web pages, clustering, and classifying the clustered content by calculating the proximity of the centroid using the K-Means Algorithm and then defining a normal categorized and abnormal combining user.

3. ORIGINALITY

This research proposes a new approach to finding the most optimal features that can describe radical content features in Indonesia. Recommended content obtained by crawling web content indicated radical by the Ministry of communications and information Republic Indonesia (positive trust). In the field research, it get 33 web URLs that are indicated as radical news. Furthermore, text mining is done to obtain information representing data, from the text mining process obtained as many as 116 content and 29,638 the number of words as a feature and as many as 4 class of labels consisting of Red, Yellow, Green, and White. Furthermore, the dominant feature feature is used to obtain words or keywords that can represent data. Selection Feature using DF Threshold. The next process is classification by k-Nearest Neighbor method with Euclidean distance to measure the level of data resemblance.

4. SYSTEM DESIGN

The proposed system consists of 7 stages: (1) data collection and preprocessing, (2) Text Mining (3) Feature Selection, (4) Training process data (5) Testing process data (6) Classification (7) calculates similarities. system design is shown in Fig. 1, where each design stage is described in Section 4.1 - 4.7.

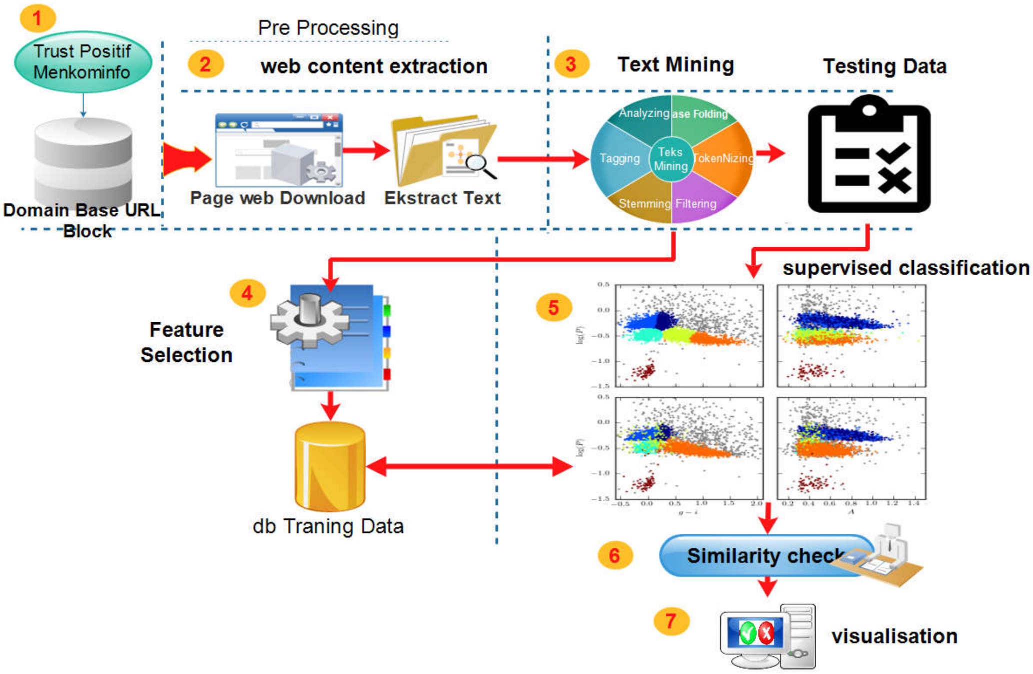

Figure 1. The system design overview of our proposed research for classfication web content radical

4.1 Data Collection and Preprocessing

The data collection is done by a searching method with radical content search index, web extremes, terror and all content related to SARA (Tribe, Religion and Taste Group) issues such as: humiliation content, hate spread. As well as data domain base content that is considered contribute negatively by the Government in Ministry of Information and Communication (Menkominfo) on positive links trust. The positive trust category content from Menkominfo is presented in Table 1.

Table 1. List of blocked website categories

Category	Number of website
Pornography	753,497
Radicalism	37
SARA	23
Violence	0
Child	3
Fraud	452
Gambling	1,162
Security	1
Copyright	48
Etc	11,412
Normalization	248

Source: Trust Positif, Menkominfo, 2016

The amount of data that successfully crawling is 33 URLs that are indicated to contain radical content with 116 amount of news or content and consists of 296,398 words. The list of blocked websites are presented in Table 2.

Table 2. List of blocked websites

URL Address
http://ajirulfirdaus.tumblr.com/
http://batalyontauhidwassunnahwaljihad.blogspot.co.id/
http://anshoruttauhidwassunnahwaljihad.blogspot.co.id/
https://jalanallah.wordpress.com/
https://religionofallah.wordpress.com/
http://daulahislamiyyah.is-great.org/
http://ummatanwahidatan.is-great.org/
http://metromininews.blogspot.co.id/
http://al-khattab1.blogspot.co.id/
http://fadliistiqomah.blogspot.co.id/
https://daulah4islam.wordpress.com/
www.muharridh.com
https://abdulloh7.wordpress.com/
http://ruju-ilalhaq.blogspot.co.id/
http://fursansyahadah.blogspot.co.id/
https://karawangbertawhid.wordpress.com/
http://terapkan-tauhid.blogspot.co.id/
https://arrhaziemedia.wordpress.com/
http://syamtodaynews.com/
https://anshardaulahislamiyahnusantara.wordpress.com/
http://jihadsabiluna-dakwah.blogspot.co.id/
http://kupastajam.blogspot.co.id/
https://mabesdim.wordpress.com/
http://anshorullah.com/
http://azzam.in
http://bahrunnaim.co
http://dawlahislamiyyah.wordpress.com
http://keabsahankhilafah.blogspot.co.id
http://khilafahdaulahislamiyyah.wordpress.com
http://tapaktimba.tumblr.com
http://mahabbatiloveislam.blogspot.co.id
http://thoriquna.wordpress.com
http://tauhiddjihat.blogspot.co.id

The blocked websites are classified into 4 class labels that determine the importance level of blocked news content. the definition of each class used in this study is based on the Positive Trust of the Ministry of Communication and Informatics of the Republic of Indonesia (Table 3).

Table 3. Classification of radical content

Label	Description Textual
Red	Written words or sentences on specific actions, mobilization of funds, people
Yellow	Written words or sentences in a certain attitude, provocation
Green	Spreading false news to certain groups
White	Normal news does not fall into the three previous categories.

4.2 Keyword Extraction

Keyword extraction is one step in preprocessing that can not be abandoned. Keyword extraction or commonly called case folding is a step of removing HTML tags that are considered not representative of the downloaded content.

4.3 Text Mining

Text mining process is done in 5 stages, namely Tokenizing, Filtering, Stemming, Tagging, and Analyzing. The five stages were done in sequence and interconnected [22]. The five stages in text mining process is described in Fig. 2.

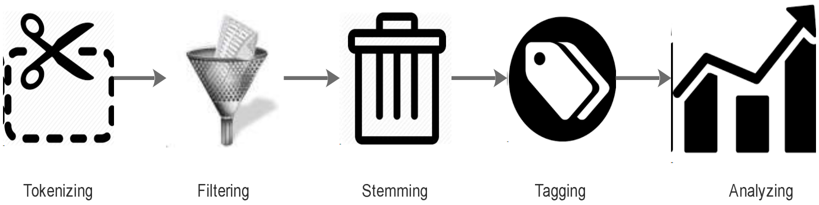

Figure 2. Text mining process

4.3.1 Tokenizing

The first stage in text mining is tokenizing. Tokenizing is the process of cutting words into tokens or words [11]. The process of cutting the text is done based on a space on each word.

4.3.2 Filtering

The next process is filtering, filtering is the process of removing the words that are not important or so-called stop list. Unrepresentative words can be conjunctions, auxiliaries, conjunctions, etc. Apart from the word, in general, the author also makes a list of words that enter the stoplist is a liaison in the Arabic language in Indonesia language.

There are 1,430 Arabic words in Indonesian that are not in accordance with the Indonesian rules that the author collects manually by reading the news content in the collection. The radical content stoplist term is presented in Table 4.

Table 4. Radical content stoplist term

No.	Term	No.	Term	No.	Term
1.	Azza	14.	Wa	27.	Badu
2.	La	15.	Ilaha	28.	Lata
3.	Ilallah	16.	Ibn	29.	Uzza
4.	Abd	17.	Asy	30.	Ardli
5.	Ar	18.	Saw	31.	In
6.	Qs	19.	An	32.	Ahlul
7.	Al	20.	Hr	33.	Aqdi
8.	Sd	21.	Ain	34.	Anna
9.	Ar	22.	Ied	35.	Ahlal
10.	Dar	23.	Wal	36.	Qura
11.	Hajj	24.	Amma	37.	Wattaqau
12.	Bin	25.	Shaf	38.	Gin
13.	Baz	26.	El	39.	Att
40	……				

4.3.3 Stemming

The next process is stemming, stemming is to remove the word particles such as affixes that can not represent or represent the document [11]. The formation of basic words with the steaming process in the document in Indonesia still has constraints where not all words can be truncated properly. This study uses the Zamief Nasri algorithm [16] which has been picketed in a sastrawi PHP library. Table 5 shows the list of Indonesian affixes.

Table 5. List of affixes

#	Prefix
Prefix	_kan, _pun, _i, _nya, _in, _is, _isme, _wan, _ah, _wi
Suffix	ber_, per_, me_, di_, ter_, ke_, se_, pe_, pem_, peng_

The stemming process is done per word. Stemming was started from checking suffix. If the word is not in the database and has a suffix, the suffix is deleted. If not, it will automatically go to the next stage to find the prefix. To find the prefix, the way is also the same as looking for suffixes, but the checking was started from the first character. The last new word is then defined and proceeded to the next process. The process of checking the word into the database is done continuously until the word becomes pure word.

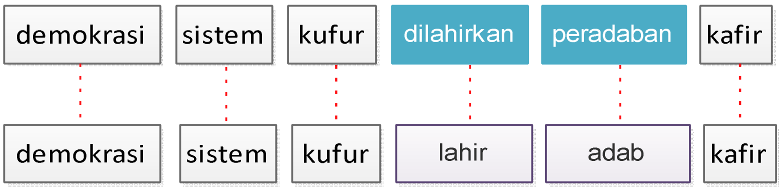

Figure 3. Illustration of radical content stemming process

The Figure 3 above illustrate of radical content stemming process. how the stemming process is done, the blue-labeled word is the word with the affix. The word "dilahirkan" has the prefix "di" and the suffix "kan" and "peradaban" have the prefix "per" and the "an" suffix. These two words are then in stemming so that the word "dilahirkan" becomes "lahir" and the word "peradaban" becomes "adab".

Stemming will generate a clean word from affix. However, the results of stemming still have not made the word into a word, it is because there are some words that still not correct writing after the affix cutting. In addition, non-standard words are still not justified, so the tagging process should be done.

4.3.4 Tagging

The next process is tagging, tagging is the process of returning the results of stemming to a basic word or just words that are not true writing. The tagging process is also used to convert non-standard words into defaults. Because all the news content on the block is the character of the blog. News that publish on blogs do not go through the editorial process like in news sites in general, this allows news uploaders can provide writing without any special language rules. So there are still words that are not in accordance with or not in the Indonesian dictionary. Abbreviated words, regional languages, foreign languages, or slang languages commonly used by news uploaders. The nonstandard words that must be processed are by tagging. The example of not standard word are presented in Table 6.

Table 6. Word list not standard

Word not standard	Word standard
alloh	Allah
fardlu	Fardhu
hadis, hadits	Hadist
haqq, haq	Hak
ibadat, ibada	ibadah
……	……

4.3.5 Analyzing

After preprocessing (keyword extraction), each document is presented as a Vector Space Model (VSM)[[12][13][14]. Each term in a document is a representation of a different feature. So the greater the document it will produce more and more features. Fig. 4 shows the aggregation process.

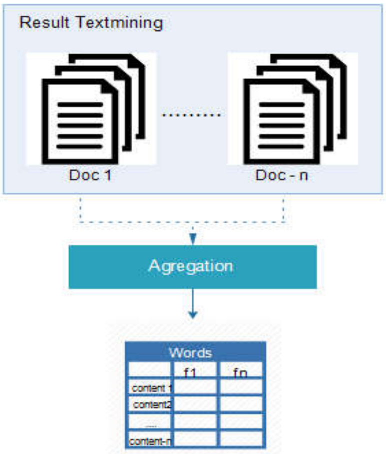

Figure 4. Aggregation process

There are three ways to calculate the value of the term frequency (TF) which are calculating the frequency as the weight, calculating the probability of occurrence as a weight (TF without normalization) and calculating the logarithm of the number of occurrences of term (normalized TF)[14][17]. The normalized TF is calculated using Eq. 1[12][13][14].

$$f_{i,j} = \frac{tf_{i,j}}{\max tf_{i,j}} \tag{1}$$

where : $tf_{i,j}$ is the normalized frequency $tf_{i,j}$ is the frequency of the data to i in document j, $\max tf_{i,j}$ is the maksimum frequency of the i term in document *j*. Table 7 showa the illustration keyword of radical aggregation matrix.

Table 7. Illustration keyword of radical aggregation matrix

Content	**Kafir**	**Demokrasi**	**Daulah**	**Khalifa**	**...**
Content1	2	3	2	2	
Content2	3	1	2	1	
Content3	1	0	1	4	
Content4	4	1	0	5	
……					

4.4 Fitur Selection

The selection of features used in this study is Document Frequency (DF) Thresholding. Document Frequency is the number of documents containing a certain term. Each term will be calculated the value of its DF and then the term is selected based on the number of DF values. If the DF value is below the specified threshold less then 3, as well as the DF of more then half document length. Then the term will be discarded [15][16][17][18]. The less visible DF term does not have much effect in the process of grouping documents. The rare term discharges in each of these documents can reduce the large feature dimensions of a document. In addition to the above methods in this system also apply DF Threshold Concentration degree [15][16][17][19].

$$\frac{DF(t,c_i)}{(1+\sum_{j=1|j\neq i}^{n} DF(t,c_j))} \tag{2}$$

Expectation Crossing Entropy (ECE) [20] is a relationship between the appearance of features and classes. Through the calculation of information, a feature appears in a class. A simplified equation such as Eq. 3.

$$P(C_i,t)\log\frac{P(C_i\mid t)}{p(c_j)} \tag{3}$$

Eq. 4 as the second selection feature method is used to reduce the feature based on the 3 principles expressed [20]:

$$DFM(t,c_j)=\frac{DF(t,c_i)}{(1+\sum_{j=1|j\neq i}^{n} DF(t,c_j))}+p(t\mid c_j)+P(C_i,t)\log\frac{P(C_i\mid t)}{p(c_j)} \tag{4}$$

4.5 Data Learning Process

The process of learning data is done with data consisting of 116 news content from 33 websites that have been blocked. This data is then trained to identify class labels by classifying them into 4 class labels (Red, Yellow, Green, and White). Classification is done by k-Nearest Neighbor Classifier method. The input used for this training is the radical content that has been assigned a class label for each news item. Class labeling is done manually using the human brain by reading in detail the news one by one then the words in the news that are considered to represent a news is selected and separated from a set of news. Then collected and labeled the class then the next process of text mining and selection features. The following learning process data flow on the system offered. Figure 5 shows the learning data process.

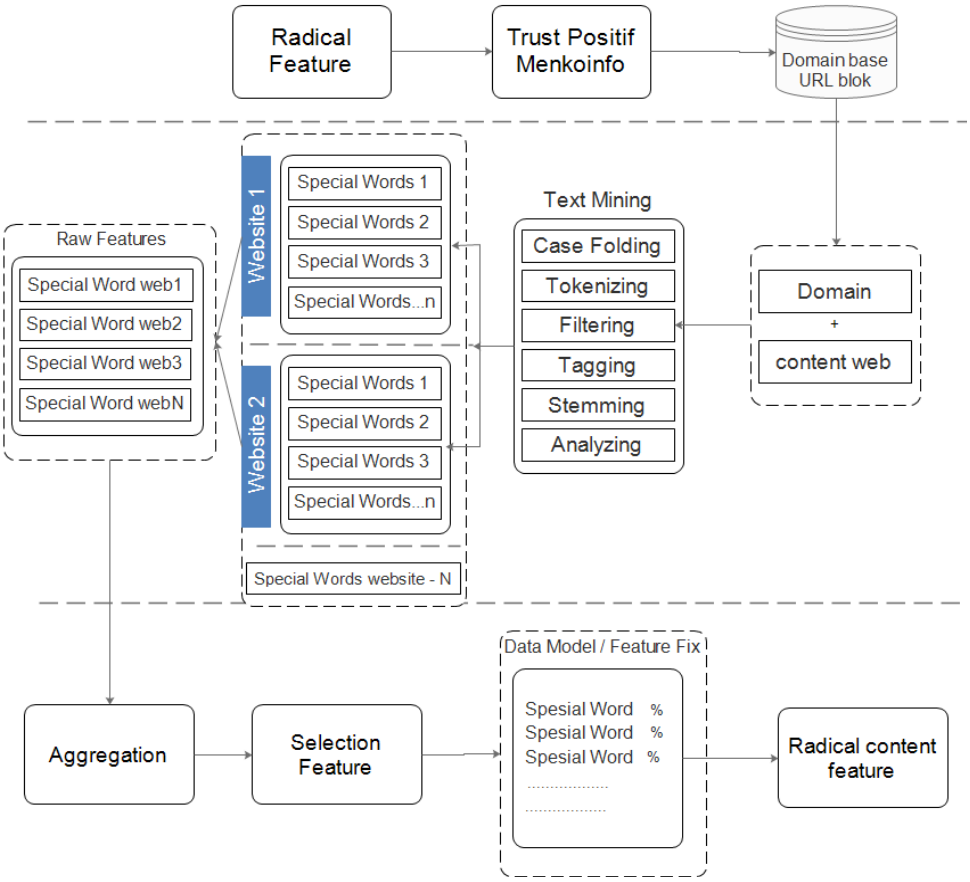

Figure 5. Data learning process

Figure 5 shows the data flow diagram of the process of data conducted by the system. The first data are taken from the positive trust minister still in the form of HTML document and link URL, the data is then done preprocessing process by removing the HTML tag, then left only text content. After the process of text mining and the final selection process features of the final process is done, the data then stored in the database

4.6 Testing Data Process

The data training method that is performed refers to the measurement performed by measuring the accuracy level of the k-Nearest Neighboard algorithm based on the radical dataset which is divided into several decision variables or attributes. A System built using Web searching application. Fig 6 shows the test has done.

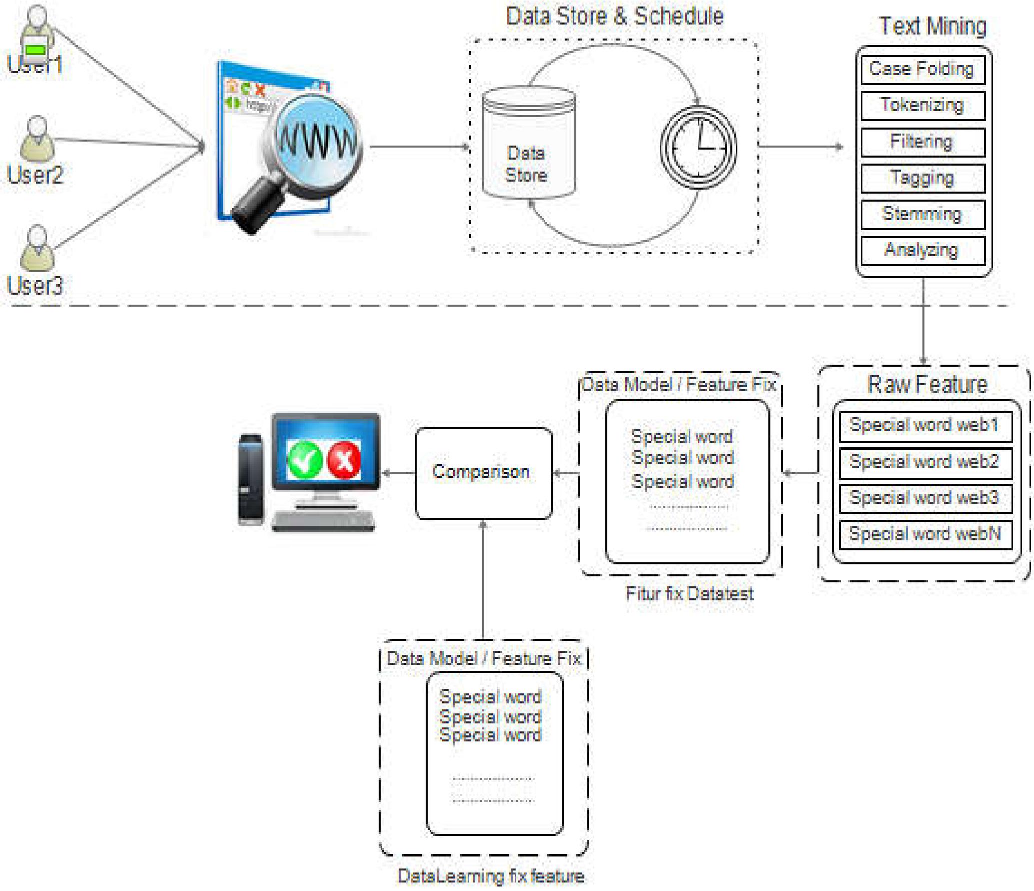

Figure 6. Testing data Process

The Fig. 6 shows the process of testing that can be done on the system offered, there are 2 modes that can be used namely input mode and database mode. Fig. 6 shows the system tested using the input mode by the user, the system user enters the news search by entering keywords that will in search. The next step is text mining process and aggregation of new data. The data will be compared with the data set that has been in the learning stored in the database, dan it classified to measure how similar the new data with data learning, then the system will issue text visualization of new data entered in the category that has been specified (red, yellow, green, white).

4.7 Classification Process

For the classification process is using previously selected features. The data is transformed into a vector (VSM) [12][13][14]. The classification process consists of two stages: the learning stage and the testing stage (described in the previous process). Each training content is assumed to have been owned by a particular class or called a class attribute. The process further is testing the model with new data or with old data that deliberately omitted its label class. It aims to measure the accuracy of model performance developed. In the system offered, we use a k-Nearest Neighbor algorithm for the process of classification. Table 8 shows the training data.

Table 8. Sample Training Data

Class	Features
Red	Zionis, war, kafir, slander, din, crusader, terrorist, conspiracy
Yellow	Khilafah, capitalist, jihad, kufr, vanity, secular, hypocrite
Green	Jews, war, conspiracy, atheist, ahlul, ahmadiyah, misra, gulam,
White	Religion, Islam, charity, worship, alms, etc.

The following k-NN work diagram on the system that is on offer. The illustration of set k value is presented in Fig. 7.

1. Vector term Document

Before entering the process of classification, document or test content first done preprocessing and converted into vector format (X_1,X_2,X_3....X_n).

2. Determine the value of k

The K value must not exceed the value of data testing, k value is the number of nearest neighbor documents. For example k-3, it means that 3 documents will be taken that have the closest distance to document testing.

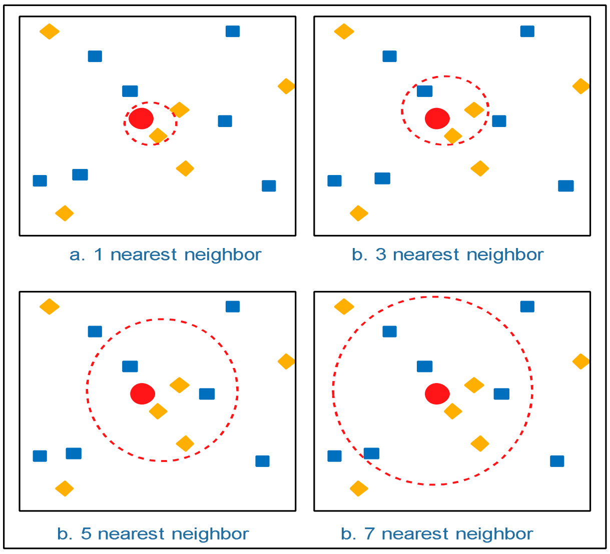

Figure 7. Illustration of set k value

3. Calculating the distance between the new data in each data label (Euclidean distance).

To calculate the degree of similarity in this project is using Euclidean distance.The variable that needs to be taken is the value of weight TF or term frequency. Each word of the test document and all sample documents on all content are calculated the distance of each weight. The calculation is using Eq. 5.

$$D(a,b) = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2} \tag{5}$$

Where: D is the distance between the two points, b is the number of nth news content and a is the initial news content.

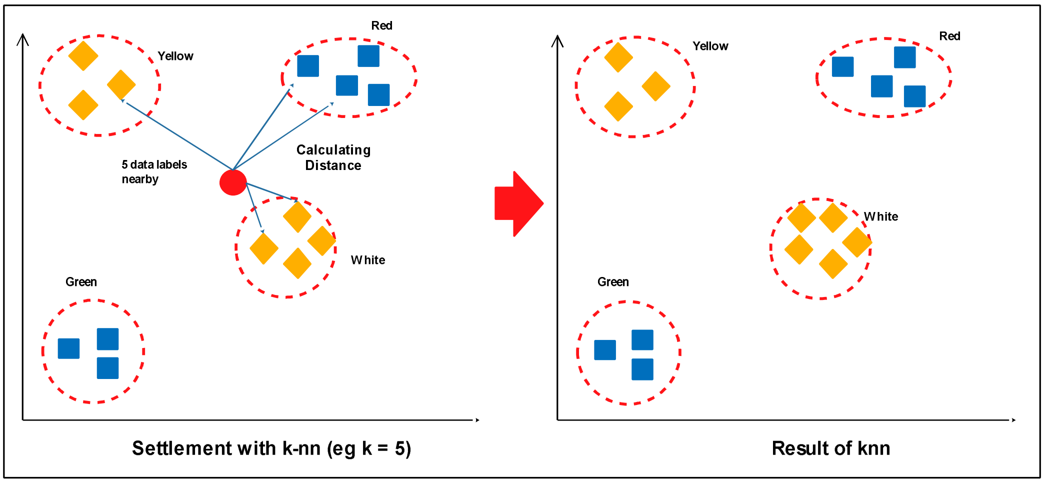

Figure 8. Illustration of k-nn algorithm process

4. Determine the k label of data that has the minimum distance

The results of distance calculation, then in the rank or in the rank of bringing proximity.

$$P(X,C_j) = \sum_{d_{i \in KNN}} D(X,b_i), y(b_i,C_j) \tag{6}$$

Where $y(b_i,C)$ is a category attribute function that satisfies the equation.

$$y(b_i,C_j) = \left\{\begin{matrix} 1, b_{i \in C_j} \\ 0, b_{i \in C_j} \end{matrix}\right\} \tag{7}$$

5. Classify new data into major data labels

Search for a majority label is using a predefined reference to determine the classification result by looking at the largest number of classes obtained between the nearest k documents predicted to know the class of the test content and to see the largest number of classes obtained between the k content of the nearest.

4. EXPERIMENT AND ANALYSIS

To obtain the best accuracy we tried to play the k parameters of the k-NN algorithm. In the experiments, the k values were set from 1 to 30, but each k value was not always good, based on the observations the optimum k value was obtained when the value of k = 6. Table 9 shows the accuracy percentage by playing k k-NN algorithm value:

Table 9. Percentage accuracy

k	Accuracy %	Error Ratio %
1	50.17	49.83
2	52.59	47.41
3	58.62	41.38
4	62.07	37.93
5	58.62	41.38
6	66.37	33.63
7	66.37	33.63
8	66.37	33.63
9	66.37	33.63

Table 9 shows the optimum and stable k values obtained with k = 6. In the graph below shows the best accuracy process to the optimum k value, from the graph the k value looks stable, we deliberately added the values of k 8 and 9 to ensure the value to the stability of k. Figure 9 shows the chart parameter of optimum k value.

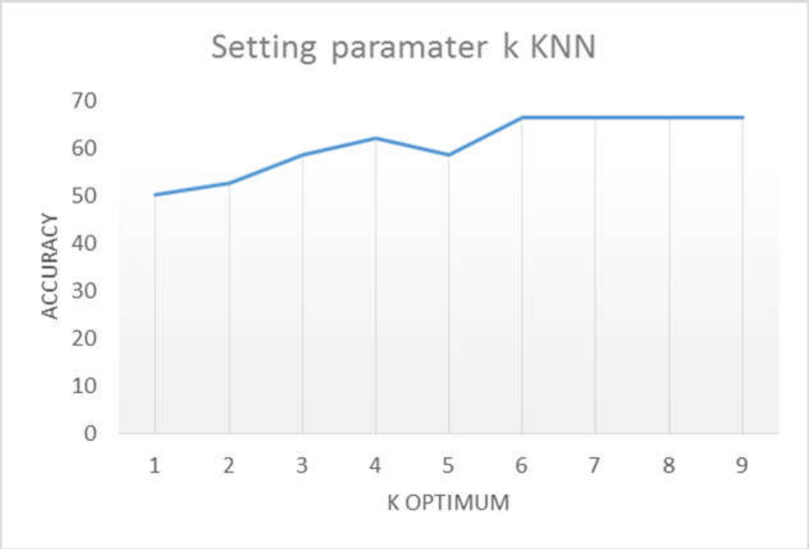

Figure 9. Chart parameter of optimum k

Confution matrix is used to evaluate the performance of the system by counting the accuracy. Generally, to evaluate the performance of classification is using F-Measure, recall and precision of each k generated. Precision and recall is used for measuring the performance of classification test. Recall is the percentage of positive labeled which are correctly recognized. Table 10 explain the definition and description of confussion matrix, such as: True Positive (TP), False Negative (FN), True Negative (TN), and False Negative (FP).

Table 10. Cofution Matrix Measurement

		Class Result Prediction	
		Positive	**Negative**
Original Class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

True Positive (TP) occurs if the message is true value in diagnosing the correct data. False Positive (FP) occurs if the message is true value in diagnosing the wrong data. False Negative (FN) occurs if the message is incorrect value in diagnosing the correct data. True Negative (TN) occurs if the message is incorrect value in diagnosing the wrong data

Precision is the precentage of correct positive predictions. It was calculated to know how precise relationships between classes. F-Measure is accuracy prediction using combination of recall and precision. The followings equations (Eq. 8 – 11) describe how to calculate the F-Measure, recall, precision and accuracy. Table 11 and 12 show the precision and recall results of each class.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$F - Measure = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

Table 11. Precision of each class

Class	K						
	1	2	3	4	5	6	7
Red	0.133	0.214	0.143	0.143	0.233	0.457	0.457
Yellow	0.385	0.538	0.583	0.471	0.385	0.571	0.571
Green	0.089	0.078	0.089	0.128	0.178	0.210	0.210
White	0.632	0.701	0.690	0.690	0.632	0.741	0.741

Based on Table 11, it appears that the "Green" class accuracy value of all k values has the lowest value. While the other classes tend to be stable. The value of k > 5 tends to be good.

Table 12. Recall of Each Class

Class	k						
	1	2	3	4	5	6	7
Red	0.118	0.176	0.118	0.118	0.118	0.294	0.294
Yellow	0.208	0.292	0.292	0.333	0.208	0.500	0.500
Green	0.087	0.095	0.192	0.192	0.271	0.333	0.333
White	0.689	0.824	0.797	0.784	0.683	0.811	0.811

In Table 12, the "Red" class value has the lowest value among all classes. And the highest is the white class that tends to be good and stable in every value k. More precisely, the precision and recall values are shown in Fig. 10 and 11.

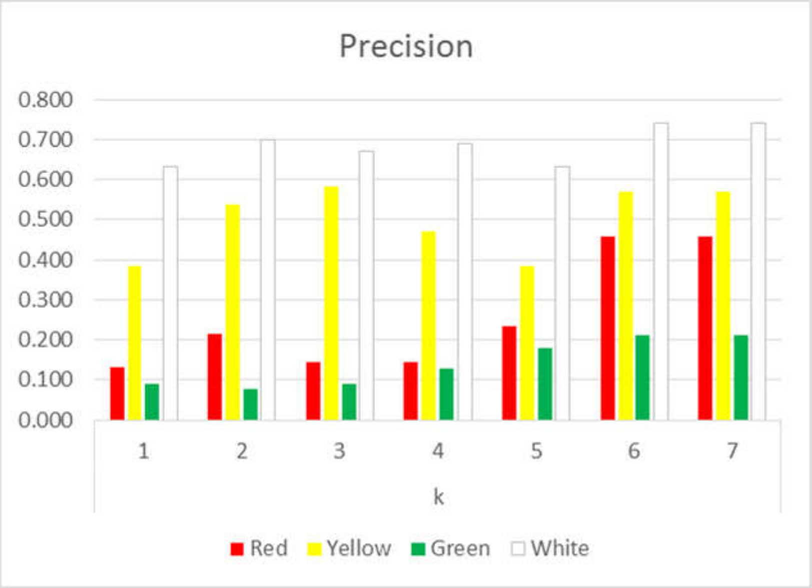

Figure 10. Precision of each class

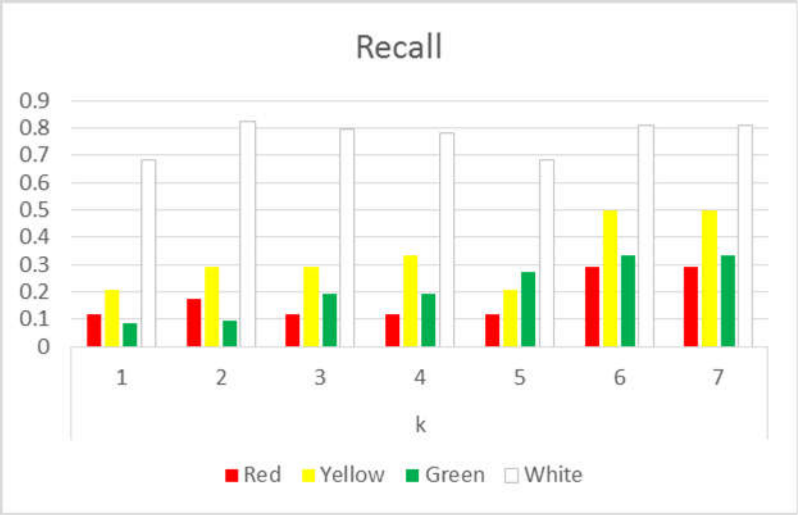

Figure 11. Recall of each class

From the Fig. 10 and 11, the White class has the highest precision and recall value on all k values. To measure the F-Measure is using Rapid Miner tools from k = 1 to 30. The comparation of precision, recall, and F-measure for k = 5 to 7 are shown in Table 13.

Table 13. Comparation of Precision, Recall, F-Measure

Class	k=5			k=6			k=7		
	P	R	F	P	R	F	P	R	F
Red	0.233	0.118	0.125	0.457	0.294	0.323	0.457	0.294	0.323
Yellow	0.385	0.208	0.270	0.571	0.500	0.533	0.571	0.500	0.533
Green	0.178	0.271	0.098	0.210	0.333	0.128	0.210	0.333	0.128
White	0.632	0.683	0.743	0.741	0.811	0.774	0.741	0.811	0.774

(P)=Precision, (R)=Recall,(F)=F-Measure

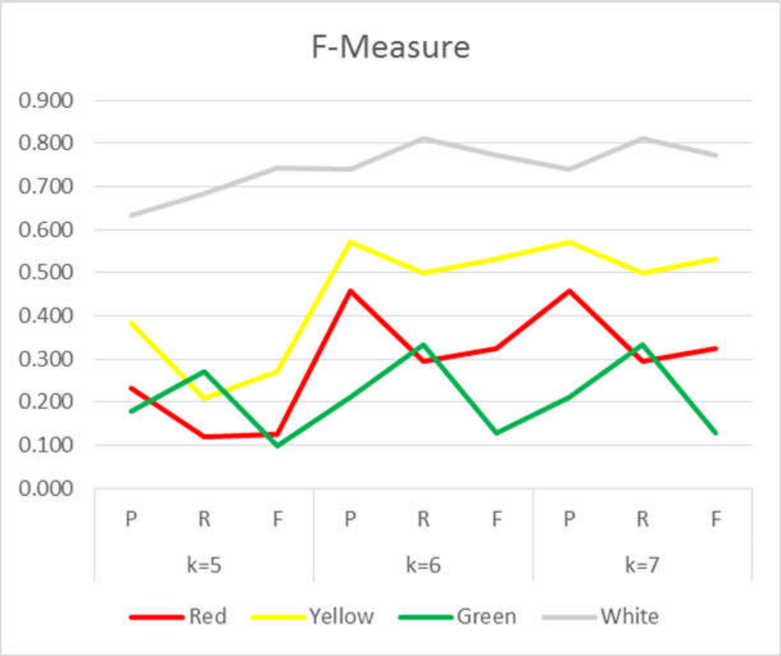

Figure 12. Chart Comparation of Precision, Recall, and F-Measure

From table 13 and Fig. 12, the comparison of F-Measure shows the highest class that is"White" class. This shows the best class of k-NN classification in this system is in the "White" class. The percentage of accuracy obtained at the value of k = 6 is 66.37%. This value represents the content that failed in the classification of 33.63% of the total content of 116 content.

6. CONCLUSION

From the k -nn classification method by plotting k value was obtained k optimum value at k = 6. The accuracy result was 66.37% by using k-Nearest Neighbor algorithm with 64 most important attribute and the error ratio equal to 33.63% with cross validation.

To parse the results obtained, it has done by trying to calculate the inter-class precision result (P) obtained from the value of k = 1 to 7, the minimum grade values of all classes were obtained in the "Green" class and the maximum results were given in the "White" class. For the (R) Recall between classes obtained, the "Red" class is lower than the other class, and class "White" has the highest value among all classes. From the comparison of table (P) Precision (R) Recall and (F) F-Measure, the best class result is class "white". It still need further research to reduce the error ratio by choosing the appropriate features and the use of a combination of classification algorithms such as Decision Tree and Support Vector Machine (SVM) to obtain higher accuracy results.

Acknowledgements

We would like to thank the Positive Trust, the Indonesian Ministry of Communications and Information which has allowed us to use the Crawling database radical in our research.

REFERENCES

[1] Muh.Subhan, Amang Sudarsono, Ali Ridho Barakbah, **Preprocessing of Radicalism Dataset to Predict Radical Content in Indonesia**, *The International Electronics Symposium on Knowledge Creation and Inteligent Computing (IES-KCIC)*, Surabaya, Indonesia, 2017.

[2] Prichard JJ, MacDonald LE. **Cyber terrorism: A study of the extent of coverage in computer Security Textbooks**. *J Inf Technol Educ*. 3, 279–89, 2004.

[3] Edna Reid, Jialun Qin, Yilu Zhou, Guanpi Lai, Marc Sageman, Gabriel Weimann, Hsinchun Chen, **Collecting and Analyzing the Present of Terrorist on the Web: A Case Study of Jihad Websites,** P.Officer et al (Eds): *ISI* 2005, *LNCS* 3495, pp. 402-411, 2005, *Springer-Verlage Berlin Heidelberg*, 2005.

[4] Sonali Vighne, Priyanka Trimbake, Anjali Musmade, Ashwini Merukar, Sandip Pandit, **An Approach to Detect Terror Related Activities on Net,** *International Journal Of Advance Research And Innovative Ideas In Edcuation (IJARIIE)*, Vol. 2 Issue. 1, 2016.

[5] Dongjin Choi, Byeongkyu Ko, Heesun Kim, Pankoo Kim, **Text analysis for detecting terrorism-related articles on the web,** *Journal of Network and Computer Applications* Vol. 38, pp. 16-21, 2014.

[6] Gerstenfeld P., Grant, R. Diana, Chiang Pu-Chau, **Hate Online: A Content Analysis of Extremes Internet Site**, *Analyses of Social Issues and Public Policy*, Vol. 3, No.1, pp.29-44, 2003.

[7] Correa D, Sureka **A. Solutions to Detect and Analyze Online Radicalization : A Survey**. IIITD PhD Comprehensive Report, 2013.

[8] Chaurasia N, Tiwari **A. Efficient Algorithm for Destabilization of Terrorist Networks**. *Int J Inf Technol Comput Sci,* Vol. 5, No. 12, pp. 21–30, 2013.

[9] Jayanthi S, Sasikala M. **XGraphticsCLUS: Web Mining Hyperlinks and Content of Terrorism websites for Homeland Security**. *Int. J. Advanced Networking and Applications,* Vol. 02, Issue. 06, pp. 941-949, 2011.

[10] Yuval Elovci, Bracha Shapira, Mark Last, Omer Zaafrani, Menahem Friedman Mothie Schneider, Abraham Kandel, **Detection of Access to Related Wb site Using an Advanced Terror Detection System(ATDS),** *Journal of The Association for Information Science and Technology,* Vol. 61, Issue 2, pp. 405-418, 2010.

[11] Mustofa Kamal, Ali Ridho Barakbah, Nur Mubtadai. **Temporal Sentiment Analysis for Opinion Mining of ASEAN Free Trade Area**

on Social Media. *The Fifth International Conference on Knowledge Creation and Inteligent Computing (KCIC) 2016-IEEE*, Manado, Indonesia, 2016.

[12] Yuhefizar, Yoyon K Suprapto, Mochamad Hariadi, I Ketut Eddy P, **Preprocessing Data Web Log Untuk Kluster Pengguna Web Menggunakan Algoritma K-Means***, Journal of Eletrical and Electronics Engineering,* Vol. 8, No. 1, 2010.

[13] Titin Winarti, Jati Kerami, and Sunny Arief, **Determining Term on Text Document Clustering using Algorithm of Enhanced Confix Stripping Stemming**, *International Journal of Computer Applications,* Vol 157, No. 9, pp. 8-13, 2017.

[14] Lailil Muflikhah and Baharum Baharudin, **Document Clustering Using Concept Space and Cosine Similarity Measurement**. *Int Conf Comput Technol Dev*. pp. 58–62, 2009.

[15] Berry MW, **Survey of Text Mining : Clustering, Classification, and Retrieval**. *Springer (New York),* pp. 262, 2004.

[16] Baharudin B, Lee LH, Khan K, **A Review of Machine Learning Algorithms for Text-Documents Classification**. *J Adv Inf Technol.* Vol. 1, No. 1, pp. 4–20, 2010.

[17] Ristu Saptono, Sulistyo ME, Trihabsari NS, **Text classification Using Naïve Bayes Updateable**, *Telematika*, Vol. 13, No. 2, 2016.

[18] Gorge Forman, **Chapter: Feature Selection for Text Classification Book: Computational Methods of Feature Selection,** *CRC Press/ Taylor and Francis Group*, 2007.

[19] Wei Zheng, Guohe Feng, **Feature Selection Method Based on Improved Document Frequency**. *Telkomnika,* Vol. 12, No. 4, pp. 905-910, 2014.

[20] Poonkuzhali G, Sarukesi K, Uma G V, **Web content outlier mining through mathematical approach and trust rating**, *ACACOS'11 Proc 10th WSEAS Int Conf Appl Comput Appl Comput Sci*, pp. 77–82, 2011.