Classification Content in Indonesian Website Da’wah using Text Mining for Detecting Islamic Radical Understanding

Nahed Nuwairah[1*], Munsyi Munsyi[2]

[1] *Faculty of Da’wah and Communication Sciences, Islamic State University Antasari Banjarmasin, Indonesia*
[2] *Faculty of Da’wah and Communication Sciences, Islamic State University Antasari Banjarmasin, Indonesia*
**Corresponding author. Email: nuwairah1975@yahoo.com*

ABSTRACT
The Islamic radical content in procedural meaning is content that has provoked the violence, spread the hatred and against nationalism through Islamic da'wah in Indonesian website. The radical definition for each country is different, especially in Indonesia. Radical content is identical with provocation issues and ethnic and religious hatred called SARA (*Suku, Agama, Ras, Antargolongan*). SARA content is challenging to detect due to the large number, unstructured system, and much noise that can be caused by multiple interpretations. This problem can threaten the unity and harmony of the religion. According to this condition, a system is required to distinguish the radical content or not. We propose a text mining approach using the DF threshold and the Human Brain as the feature extraction in this system. The system is divided into several steps are collecting data which is including at pre-processing, text mining, selection features, classification for grouping the data with the class label, similarity calculation for processing data training, and visualization to the radical Islamic content or not radical content. This research is expected to be literate for users who access websites exposed to radical understanding to suppress the spread of provocation on websites in Indonesia. The experimental result shows that using a 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 by collected data using web scrapping in the website blocked by the Ministry of Information and Communication Indonesia (Menkominfo).

Keywords: *Islamic Radical Content, SARA, Text Mining, k-NN Algorithm.*

1. Introduction

The behavior of human beings, when they use the Internet for terrorist acts is known as cyber terrorism. Cyber terrorism, it is a form of politically-motivated terrorism that uses technology and information, computer networks, and technical infrastructure to destroy [1]. Due to the importance of internet networks, some experts argue that cyber terrorism is more dangerous than traditional terrorists [2]. Online radicalization [3], also called Cyber-Terrorism or Extremism, or Cyber-Racism, or Cyber Hate, has become a significant concern of society, government, and law enforcement agencies around the world [4][5]. The definitions of radical treatment in each country are different. Hence it has become a debate among multi-disciplinary researchers, making it difficult to determine the criteria that radicalism itself says. Radical content in a procedural sense, namely provoking violence, spreading hatred, and anti-nationalism [6][7] For BNPT, radical content encourages and provokes people to commit violence in the name of religion, interprets jihad as suicide bombings, and takes the lives of the others.

Various platforms on the Internet (easy to publish content, allow anonymity, provide exposure to millions of users and potential from very quickly and widely) such as YouTube, Twitter, Facebook, and Instagram are easily misused for malicious purposes [4]. Such platforms are used to form hate groups, racist societies, spread extremist views, incite anger or violence, promote radicalization, recruit members and create virtual organizations and communities. It became a vital problem for researchers, the difficulty of getting features that match the criteria of radical itself, this is because the content is very secretive to avoid traditional web crawlers [5]. The solution offered in this research is to try the Text Mining method using Web Scraping to classify radical content. This research proposes a new approach to the best classification technique in classifying web content as related to radicals and not radicals, as well as proposing a new approach to detect the behavior of internet users who access the web related to radical content to group data to detect content related to radicals.

2. Material and Method

For supporting this research, our project is necessary to put forward cover the basic theory and radical definitions in terms of various views (politics, sociology, culture, economics, and IT) and theories about web content mining. For example, radicalism in the Indonesian Dictionary, second edition, Balai Pustaka is a school or ideology that wants social and political change or renewal through violence or drastic means. Meanwhile, in the popular scientific dictionary by M. Dahlan al Barry, the term radicalism is defined as the political ideology of the state that demands major changes and reforms as a way to achieve progress.

2.1. Scheme of Islamic Radical Content

This study's data mining methods and techniques explain the stages or flow of creating radical content detection applications to help users about radical and non-radical content in Indonesia. The analysis results are in the form of keywords [8]. This keyword became a reference for the feature selection process [9]. In the following process, these keywords are stored in the database. The process is the stage after the selection feature is carried out. This classification determines which keywords or features correspond to the label class according to the label class given by the Ministry of Communication and Information Technology. The system design in this study can be seen in Figure 1.

As in figure 1, the system's flow starts from taking news content data from the positive Internet of the Ministry of Communication and Information; then, the web content extraction process is carried out. Web content extraction removes the HTML tag and leaves text content. Then, the next process is Text Mining. Text mining is the process of extracting and retrieving words that represent what is in a news document, usually called keywords. The text mining process consists of several stages, namely: case folding, tokenizing, filtering, stemming, tagging, and the last process of text mining is analysis. The analysis results are in the form of keywords [10]. This keyword became a reference for the feature selection process. In the next process, these keywords are stored in the database. At the same time, the features will be selected and saved back into the database. After that, feature selection is made to reduce the dimensional space of the data. Then the classification process. The process is the stage after the selection feature is carried out. It determines which keywords or features correspond to the label class of the Ministry of Communication and Information Technology.

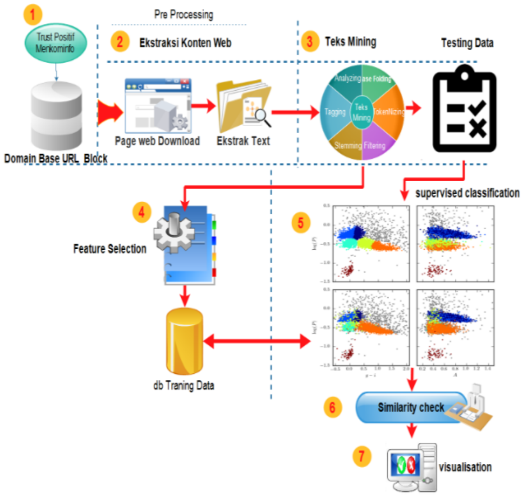


Figure 1 Radical Content Research System Design.

The last process measures the level of data similarity between the data and keywords that have been stored in the database using the Euclidean distance. Then the keyword is visualized according to the criteria for the closeness of the data to the labels stored in the database. The discussion in this stage is to get a model from radical content. Making this data model is carried out in several stages, starting from literature study, collection of radical content data, pre-processing starting with text mining, manual class labeling, and feature selection. Therefore, a model of radical content is obtained according to the characteristics and definitions of radicals in Indonesia.

Our system collected the data; then text mining began. Text Mining itself functions to look for words that represent what is in the content so that these words can be analyzed and related. Text mining is the application of data mining concepts and techniques to look for patterns in text, which is analyzing text to find useful information for specific purposes. Based on the irregularity of the text data structure, the text mining process requires several initial stages, which are preparing so that the text can be changed to be more

Table 1. Radical content classification label

Number	Classification	Description
1	Red	Call on specific actions, mobilization of funds, and People
2	Yellow	Invitation to certain attitude, provocation
3	Green	Spread fake news to certain groups

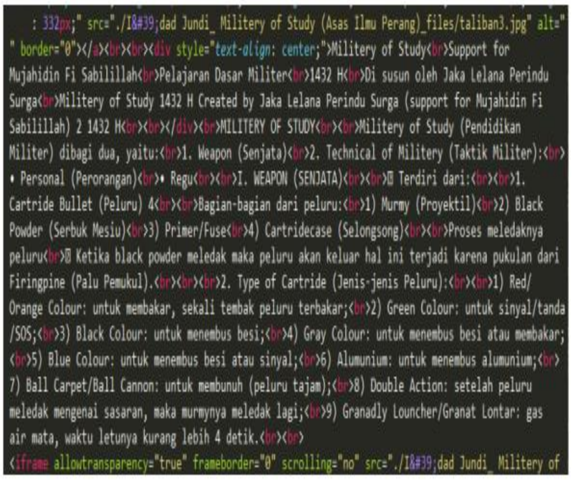
structured. The text mining process is carried out in five stages: tokenizing, filtering, stemming, tagging, and analyzing. The five stages are carried out sequentially and are interconnected. The next process gives four class labels: Green, Yellow, Red, and White. Red Label is symbolized as content with an enormous influence such as (enabling readers to mobilize and raise funds, provoke to oppose the government). Table 1 is a description of each class label content radical obtained from the Menkominfo.

In this research, the literature review explores radical content. This literature study includes the types of radical content and non-radical content, radical research in various countries, characteristics of content categorized as radical from various countries, the content of radical content that can be converted into radical features of Indonesia, as well as studies on methods of determining radical class in various countries. This category is then partly used as a rule for radical content on the system being built. Data collection was carried out by interviewing sources related to radical Islam and search methods with search indexes for radical content, web extremes, terror, and all content related to issues of SARA (ethnicity, religion, and sense of class) such as Contents of Insult, Hate Spread, Content Websites of Religious Organizations (MUI, NU, Hidayatullah, Hisbut Tahrir, PGI, MAGABUDI, PHDI). As well as data Domain base content that is considered negative by the Government, the Ministry of Information and Communication (Menkominfo) on positive links trust. A total of twenty-three content containing SARA, twenty-three content with radical issues, for an explicit explanation of the content of the positive trust category from the Minister of Communication and Information are shown in table 2.

The number of data successfully deleted was thirty-three URLs, indicating that they contained radical content with 116 news or content and obtained 296,398 words. The process at this analysis stage is to obtain a radical content model. Making this data model is carried out in several stages, starting from literature study, collecting data on radical content, pre-processing starting with text.

Table 2. URL positive internet from Menkominfo

URL
1. http://ajirulfirdaus.tumblr.com
2. http://batalyontauhidwassunnahwaljihad.blogspot.co.id/
3. http://anshoruttauhidwassunnahwaljihad.blogspot.co.id/
4. https://jalanallah.wordpress.com/
5. https://religionofallah.wordpress.com/
6. http://daulahislamiyyah.is-great.org/
7. http://ummatanwahidatan.is-great.org/ Access
8. http://metromininews.blogspot.co.id/
9. http://al-khattab1.blogspot.co.id/
10. http://fadliistiqomah.blogspot.co.id/
11. https://daulah4islam.wordpress.com/
12. www.muharridh.com Domain Url Closed
13. https://abdulloh7.wordpress.com/
14. http://ruju-ilalhaq.blogspot.co.id/
15. http://fursansyahadah.blogspot.co.id/
16. https://karawangbertawhid.wordpress.com/
17. http://terapkan-tauhid.blogspot.co.id/
18. https://arrhaziemedia.wordpress.com/
19. http://syamtodaynews.xyz/ Error Access
20. https://anshardaulahislamiyahnusantara.wordpress.com/
21. http://jihadsabiluna-dakwah.blogspot.co.id/
22. http://kupastajam.blogspot.co.id/
23. https://mabesdim.wordpress.com/
24. http://anshorullah.com/
25. http://azzam.in
26. http://bahrunnaim.co
27. http://dawlahislamiyyah.wordpress.com
28. http://keabsahankhilafah.blogspot.co.id
29. http://khilafahdaulahislamiyyah.wordpress.com
30. http://tapaktimba.tumblr.com
31. http://mahabbatiloveislam.blogspot.co.id
32. http://thoriquna.wordpress.com
33. http://tauhiddjihat.blogspot.co.id



```
   : 332px;" src="./I&#39;dad Jundi_ Militery of Study (Asas Ilmu Perang)_files/taliban3.jpg" alt="
" border="0"></a><br><br><div style="text-align: center;">Militery of Study<br>Support for
Mujahidin Fi Sabilillah<br>Pelajaran Dasar Militer<br>1432 H<br>Di susun oleh Jaka Lelana Perindu
Surga<br>Militery of Study 1432 H Created by Jaka Lelana Perindu Surga (support for Mujahidin Fi
Sabilillah) 2 1432 H<br><br></div><br>MILITERY OF STUDY<br><br>Militery of Study (Pendidikan
Militer) dibagi dua, yaitu:<br>1. Weapon (Senjata)<br>2. Technical of Militery (Taktik Militer):<br>
• Personal (Perorangan)<br>• Regu<br><br>I. WEAPON (SENJATA)<br><br> Terdiri dari:<br><br>1.
Cartride Bullet (Peluru) 4<br><br>Bagian-bagian dari peluru:<br>1) Murmy (Proyektil)<br>2) Black
Powder (Serbuk Mesiu)<br>3) Primer/Fuse<br>4) Cartridecase (Selongsong)<br><br>Proses meledaknya
peluru<br> Ketika black powder meledak maka peluru akan keluar hal ini terjadi karena pukulan dari
Firingpine (Palu Pemukul).<br><br><br>2. Type of Cartride (Jenis-jenis Peluru):<br><br>1) Red/
Orange Colour: untuk membakar, sekali tembak peluru terbakar;<br>2) Green Colour: untuk sinyal/tanda
/SOS;<br>3) Black Colour: untuk menembus besi;<br>4) Gray Colour: untuk menembus besi atau membakar;
<br>5) Blue Colour: untuk menembus besi atau sinyal;<br>6) Alumunium: untuk menembus alumunium;<br>
7) Ball Carpet/Ball Cannon: untuk membunuh (peluru tajam);<br>8) Double Action: setelah peluru
meledak mengenai sasaran, maka murmynya meledak lagi;<br>9) Granadly Louncher/Granat Lontar: gas
air mata, waktu letunya kurang lebih 4 detik.<br><br>
<iframe allowtransparency="true" frameborder="0" scrolling="no" src="./I&#39;dad Jundi_ Militery of
```

Figure 2. Web-scrapped news content.

3. RESULT AND DISCUSSION

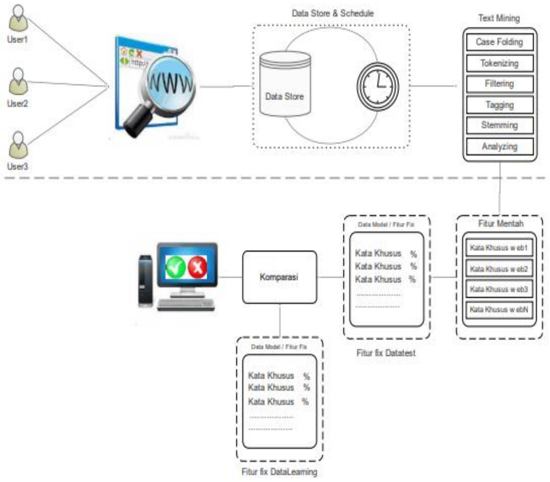


Figure 3. Design trial system

We implement the system in the form of a content search engine application using the PHP programming language. It is intended that the application or system built can be accessed by any operating system and PHP is open-source, meaning that anyone can develop or build applications using the programming language. This stage begins with building a search engine form to make it easier for users to search for news-related content.

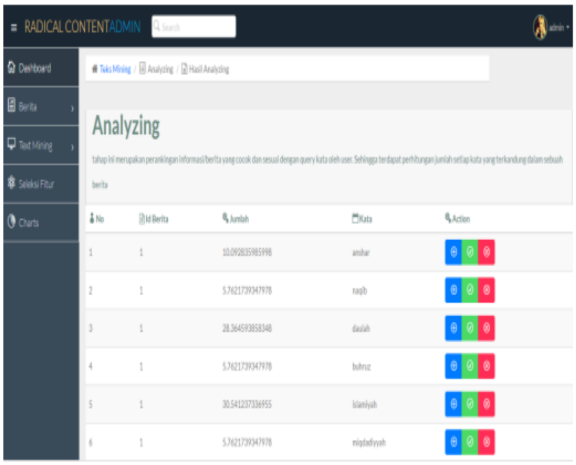
Figure 4. shows the process of testing the dataset that has been formed. First, the user will perform a query in the form of news or in the form of a sentence, which will be queried later. The results of the query will be text mining, then compared with the words (keywords) stored in the database, then the system will check how close the word being queried is to the data in the database as a keyword. Finally, the system will issue news labels that the user searches for.

In this classification process, documents that have been processed until the text frequency calculation stage will be classified based on the class determined (Red, Yellow, Green, White). Class label determination is done manually with Human Brain. The process is carried out by reading all the learning content obtained from the Ministry of Communication and Informatics' positive trust data scrapping. The news obtained is as much as 116 news, then sorted according to the label specified. Then do text mining, convert it into a vector document and aggregate it.

The last stage in the system is to recognize the content that the user has entered into the system. This recognition process is performed by classifying the testing data against the data model. Many methods have developed in the classification process, such as Artificial Neural Networks, Support Vector Machine (SVM), Naive Bayes, k-Nearest Neighbor, etc. Whether or not the success rate in this process is influenced by the classification method chosen. The number of data samples in this study was classified as small, as many as 116 content. In addition, the dataset is numerical, meaning that the features in the dataset are numerical, and there are no nominal values known as categorical. Based on these characteristics.

Then the suitable method to use is k-Nearest Neighbor (kNN)[11] [12] [13]. The k-NN method is an algorithm that is very simple to understand and can work well. This Algorithm is also known as non-parametric lazy learning. It means that this Algorithm does not need to assume based on the data distribution. Other than that. This method is also known as a lazy algorithm. It means that this method does not require learning to generalize data because all data is used as training. The decision made by this method is that the classification process is a consideration of the test data value for all training data. Calculating the decision or classification in this method is to identify the value of the proximity or distance of the test data to the overall training data. It was carried out by calculating the value of the euclidean distance, then this method voting for the distance values that have been sorted from the smallest to the largest value. The number of votes is a characteristic of the KNN denoted by the variable k., because of the nature of this Algorithm based on the majority of votes. Then the k values are generally odd, like 1,3,5,7, etc. From the voting results, the classification of the KNN method will be obtained on the test data against the training data. Each test data will get a label that KNN has classified at this stage.

The user will perform a query in the form of news or in the form of a sentence, which will be queried later. The results of the query will be text mining, then compared with the words (keywords) stored in the database, then the system will check how close the word being queried is to the data in the database as a keyword. Finally, the system will issue news labels that the user searches for. In this classification process, documents that have been processed until the text frequency calculation stage will be classified based on the class determined (Red, Yellow, Green, White). Class label determination is done manually with Human Brain. The process is carried out by reading all the learning content obtained from the Ministry of Communication and Informatics' positive trust data



Analyzing

tahap ini merupakan perankingan informasi/berita yang cocok dan sesuai dengan query kata oleh user. Sehingga terdapat perhitungan jumlah setiap kata yang terkandung dalam sebuah berita

No	Id Berita	Jumlah	Kata	Action
1	1	10.092835985998	anshar	⊕ ⊘ ⊗
2	1	5.7621739347978	naqib	⊕ ⊘ ⊗
3	1	28.364593858348	daulah	⊕ ⊘ ⊗
4	1	5.7621739347978	buhruz	⊕ ⊘ ⊗
5	1	30.541237336955	islamiyah	⊕ ⊘ ⊗
6	1	5.7621739347978	miqdadiyyah	⊕ ⊘ ⊗

Figure 4. Query result.

scrapping. The last stage in the system is to recognize the content that the user has entered into the system. This recognition process is performed by classifying the testing data against the data model.

4. CONCLUSION

After several experiments, conclusions can be drawn from this research, as follows: *First,* 3Classification of radical content using the K-Nearest Neigbord algorithm carried out in this study can work well. Most of the content data is classified according to manual classification with an accuracy value of 66.37%. *Second*, the percentage of accuracy is obtained at the optimal k value = 7 by calculating the accuracy value using cross-validation. The value is 66.37%, with an error rate of 33.63% of the total content of 116 content. With the comparison table (P) Precision (R) Recall, and (F) F-Measure, the higher data class value is obtained in the "white" class. *Third,* the system that was built requires a processing time of 0.704 seconds, and the largest memory usage is 884656 bytes. This indicates that the k-NN classification algorithm in memory usage is relatively high in each classification processing.

We recommended adding the Indonesian synonym identification function to enhance the system development. If the entered keywords have synonyms with the word in the database, it is not considered a new word, so that it helps in the classification process. The radical classification of themes for each label class is determined by experts or people who are experts and responsible for determining the radical content. The subsequent research can add or be combined with algorithms such as Naïve Bayes Classification (NBC), Support Vector Machine (SVM), Decision Tree (Decision Tree), Neural Network (NN) to improve accuracy results. Apart from that, from the future system, this system can be developed further towards the Proxy, Mobile platform, etc.

AUTHORS' CONTRIBUTIONS

The contributions of this research are to develop the analysis of content website in Indonesian da'wah to classify the radical content through the HTTP protocol for user's access in the website using their end devices such laptop, computer, and smartphone.

ACKNOWLEDGMENTS

This research was funded by Kementerian Agama Republik Indonesia on Litapdimas in 2019.

REFERENCES

[1] Y. Elovici *et al.*, "Content-based detection of terrorists browsing the web using an Advanced Terror Detection System (ATDS)," *Lect. Notes Comput. Sci.*, vol. 3495, pp. 244–255, 2005, DOI: 10.1007/11427995_20.

[2] O. Margareta, S. Narti, and S. Sari, "(Framing Analysis Of The News Of The Legal Verdict Of Inmate Terrorism Abu Afif (Analisis Framing Pemberitaan Narapidana Teroris Abu Afif Pada Media Online," pp. 98–109, 2020.

[3] A. C. Pretus, N. Hamid, and H. Sheikh, "Scientific report Impact of legal punishment of nationalist political leaders on social polarization," pp. 1–18, 2019.

[4] D. Correa and A. Sureka, "Solutions to Detect and Analyze Online Radicalization: A Survey," vol. V, no. January, pp. 1–30, 2013, [Online]. Available: http://arxiv.org/abs/1301.4916.

[5] N. Chaurasia and A. Tiwari, "Efficient Algorithm for Destabilization of Terrorist Networks," *Int. J. Inf. Technol. Comput. Sci.*, vol. 5, no. 12, pp. 21–30, 2013, DOI: 10.5815/ijitcs.2013.12.03.

[6] W. A. Setianto, "Literasi Konten Radikal di Media Online," *J. Ilmu Komun.*, vol. 16, no. 1, p. 75, 2019, doi: 10.31315/jik.v16i1.2684.

[7] M. R. M. Ahyad, "Analisa Penyebaran Berita Hoax Di Indonesia," *Jurnal*, p. 16, 2017, [Online]. Available: file:///C:/Users/USER~1.LAB/AppData/Local/Temp/ ANALISIS PENYEBARAN BERITA HOAX DI INDONESIA.pdf.

[8] P. Soepomo, “Penerapan Text Mining Pada Sistem Klasifikasi Email Spam Menggunakan Naive Bayes,” *Penerapan Text Min. Pada Sist. Klasifikasi Email Spam Menggunakan Naive Bayes*, vol. 2, no. 3, pp. 73–83, 2014, doi: 10.12928/jstie.v2i3.2877.

[9] Y. Patil and S. Patil, “Review of Web Crawlers with Specification and Working,” *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 5, no. 1, pp. 220–223, 2016, doi: 10.17148/IJARCCE.2016.5152.

[10] N. Suryandari and V. Lutviana, "Impression management of buzzer in social media Twitter," *Jurnal Studi Komunikasi (Indonesian Journal of Communications Studies)*, vol. 4, no. 3. p. 614, 2020, DOI: 10.25139/jsk.v4i3.2665.

[11] A. Valdivia, M. V. Luzón, E. Cambria, and F. Herrera, “Consensus vote models for detecting and filtering neutrality in sentiment analysis,” *Inf. Fusion*, vol. 44, pp. 126–135, 2018, DOI: 10.1016/j.inffus.2018.03.007.

[12] A. Valdivia, M. V. Luzíón, and F. Herrera, "Neutrality in the sentiment analysis problem based on fuzzy majority," In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1-6. 2017, DOI: 10.1109/FUZZ-IEEE.2017.8015751.

[13] M. López, A. Valdivia, E. Martínez-Cámara, M. V. Luzón, and F. Herrera, "E2SAM: Evolutionary ensemble of sentiment analysis methods for domain adaptation," *Information Sciences"*, vol. 480, pp. 273-286, 2019, doi: 10.1016/j.ins.2018.12.038.