Webpage Classification Using Naïve Bayes Classifier and Information Retrieval Method to Block the Pornography Contents

Andreas and Lusia Permata Sari Hartanti

***Abstract*— The need of information is such magnitude. It would trigger advances in information technology. Many things can be done by using the internet. The main objective that people use the Internet is to find the information. Unfortunately, not all the information available on the internet it is true and good consumed. There are plenty of information on the internet which is a hoax and contains elements that are contrary to morals and ethics such as terrorism, racism, and pornography. Thus, it required a strong faith and knowledge to be able to sort the incoming information. Due there are many children who also use the internet and conditions where the parents cannot supervise the activities of their children continuously, it would require an application that is embedded in the web browser so that bad content can be blocked automatically. This article focuses on the research of handling pornographic content on the text based webpages. It needed a smart application to be able to distinguish text that contains pornography. Thus, we are implementing artificial intelligence into our research by applying Naive Bayes and information retrieval method. As a result, the application is able to block 88.02% of the pornographic content**.

***Index Terms*—Information retrieval, pornography, smart system, web content filtering.**

I. Introduction

In early 2016 the International Telecommunication Union (ITU) noted that there are seven billion people (95% of the global population) live in an area that is covered by a mobile-cellular network [1]. Nowadays it almost reaches 4 billion people (53% of global population) use the internet for various purposes. Most people use the internet is to obtain and exchange information. The average of the growth of internet users from 2010 to 2016 is 6.7% per year. This rapid growth rate is influenced by the increasing needs of the society to information.

There is so much information available from various sources on the internet. Unfortunately, not all the information which exists on the internet it is true and good consumed. There are plenty of information on the internet which is a hoax and contains elements that are contrary to morals and ethics such as terrorism, racism, and pornography. This kind of information can have a negative effect on society, such as the declining of nation's morality, ethical violation, the rising of crime rates, security threats, character assassination, and others.

In order to avoid these negative influences, the discretion is required in selecting the information obtained from the internet. The problem faced in Indonesia is still many people who cannot distinguish between the right information and hoax, harm content or not. Moreover, if the receiver of the information is the children that in fact has a very low ability to select and sort the information. The main issue for children is pornography. Thus, this article focuses on build the embedded application that block the pornographic content in the web pages which accessed by users.

Based on the results of a survey conducted by TechAddiction, 12% of the website on the internet are pornographic, that is 24,644,172 sites. These data show out how big the interest of internet users to pornographic content.

Pornography is derived from the Greek pornographos composed of two Greek words; porne (prostitutes) and graphos (image or text). Maurice Yaffe expalined that Pornography was related to obsenity more than just eroticism. As described by Maurice Yaffe that pornography related to obscenity more than just eroticism [2]. Indonesian dictionary has two definition of pornography. First, Pornography is the depiction of erotic behavior with painting or writing to arouse lust. Second, it means reading material that is deliberately and solely designed to arouse lust in sex [3]. Pornography is materials (such as text, photos, erotic films) that describes sexual activity or erotic behavior in a manner designed to desire sexual arousal [4].

The internet can be used as the best fertile medium for the pornographic content dissemination. This is possible due the development of information technology is so rapid and the information disseminated over the Internet is "uncensored" [5]. Based on this case then many programmers build applications that used to blocking improper web contents, including pornography that called as content filtering.

The current content filtering applications work based on the domain name, the title of the article and the metadata of the web page. Based on this reason thus many contributors camouflage their article that containing pornography by using domain name, article's title and metadata that far from the pornographic attributes.

Another problem is that the current content filtering applications work based on keywords that were given to them. They examine the existence of those given keywords. If any of these keywords found in one of those three factors (domain name, article's title, and metadata), then the web page will automatically be blocked. But not all the web page containing such wods are categorized as pornography. As an example, education website with an article about sex education might

Manuscript received July 15, 2017; revised November 3, 2017.

Andreas is with Universitas Pelita Harapan Surabaya, Indonesia (e-mail: andreas.jodhinata@uph.edu).

Lusia Permata Sari Hartanti is with Universitas Pelita Harapan Surabaya, Indonesia (e-mail: lusia.hartanti@uph.edu).

doi: 10.18178/ijfcc.2017.6.4.509

contain many words related to adult words like sex and genital organs.

Based on all problems above, it is necessary to build the better application that can automatically block web pages that contain pornography. To make the application "smarter" than before, there must be other methods that implemented into it rather than only works based on the given keywords. This article modifies the Bahasa Indonesia stemmer algorithm, implements the Naive Bayes algorithm, and using the information retrieval method. The application built in this research works by reading all the content of the web pages and determines its category. If the content is categorized as pornographic then the web page will be blocked (not be displayed on the web browser).

II. Stemmer Modification for Bahasa Indonesia

In linguistic morphology, stemming is a process which turns affixed words to their base form. The stemming process itself is widely used as a preliminary process in many information retrieval processes, especially in text mining and Natural Language Processing (NLP), because of the assumption that the words that have the same stem usually also have the same meaning [6].

The stemming process works verbatim. In reality, the stemming process is applied to the sentence. There are several steps before doing the stemming process: stage tokenization and stop words removal. Tokenization is a process that splitting a sentence into several words based on spacing characters and the punctuations. The simple process of tokenization is called segmentation.

There are several issues to note in tokenization phase, especially if the process is carried out on the words that are quite complex. A new problem arises at tokenization phase which is how to decide the proper stem from a sentence.

TABLE I. Example of the Stop Word List

Category	Examples
Conjunctions	and, as, because, but, for, just as, or, neither, nor, not only, so, yet, etc.
Pronouns	he, she, him, her, they, it, we, that, this, who, which, everything, etc.
Prepositions	about, after, among, at, below, above, besides, for, from, in, into, on, etc.
Adverbs	almost, very, rather, really, here, everywhere, today, yesterday, etc.
Numbers	a, one, two, three, many, much, few, a lot, etc.

Stop word removal phase is a process of removing the words that commonly appear but have no meaning in that context. Table I shows the example of the stop word list.

Based on the technique used, the stemming algorithm can be divided into three types. First, the stemming algorithm that works by removing the affixes according to the given reference table, Second, the stemming algorithm that used the dictionary of root words (stems) as a reference to turn words into their stem, Third, the stemming algorithm that works based on the result of the training process. There are many documents needed in the training process. The collections of the documents that are involved in training process were called as the corpus.

The stemming process in Bahasa Indonesia is more complex than in English since there are three types of affixes in the morphology of Bahasa Indonesia, namely: awalan (prefix): affixes which are in the front of the word. akhiran (suffix): affixes which are in the end of the word and sisipan (infix): affixes which are in the middle of the word.

The morphology of Bahasa Indonesia also divides affixes based on their composition into three types, which are:

1) Derivational Prefix (DP) is a group of prefixes that could be added into pure root word or to root word which already added up to two prefixes. Derivational prefixes including:
a. Prefix that can morphologies (“me–”, “be–”, “pe–”, and “te–”)
b. Prefix that cannot morphologies (“di–”, “ke–” and “se–”)
2) Inflection suffix that is a group of suffix that does not alter the stem (root word). This group can be divided further into two types:
a. Particle (P): “–lah”, “–kah”, “–tah”, and “–pun”.
b. Possessive Pronoun (PP): “–ku”, “–mu”, and “–nya”.
3) Derivational Suffix (DS) that is a group of original suffixes in Bahasa Indonesia which added directly to the root word like “–i”, “–kan”, and “–an”.

According to the morphology of Bahasa Indonesia, a root word can only have up to three derivational prefixes, one derivational suffix, one possessive pronoun, and one particle, thus the form of affixed word in Bahasa Indonesia can be modeled as:

$$\left[DP + \left[DP + [DP +]\right]\right] root\_word\left[[+DS][+PP][+P]\right] \quad (1)$$

III. Naïve Bayes Classifier And Information Retrieval Method

Naive Bayes classifier is one of the classifiers that widely used especially in the text classifier and works based on the popular Bayes' theorem and combined with the Naïve algorithm [7], [8]

Posterior probability is a probability of a particular object that belongs to a class given its observed feature values. The form of Bayes' theorem that is used as a text classifier can be formulated as:

$$p(c_j|d) = \frac{p(c_j)p(d|c_j)}{p(d)} \quad (2)$$

d is the text document that represented as a set of words $(w_1, w_2, \dots, w_n)$ where w_1 is the first word and w_n is the last word in the document, c_j is the text category in document d that will be classified, and $p(c_j)$ is probability of text category c_j [9].

The approach of the Bayes algorithm is by selecting text category that has the highest probability c_{MAP}, where MAP is stand for Maximum a Posteriori, and can be written as:

$$c_{MAP} = argmax \frac{p(c_j)p(d|c_j)}{p(d)} \quad (3)$$

Since $p(d)$ is the probability of the document and its value is constant for each c_j, thus it can be ignored, so Eq. 3 can be simplified as:

$$c_{MAP} = argmax\, p(c_j)p(d|c_j) \quad (4)$$

The probability of $p(c_j)$ can be specified by counting the number of training documents in each text category c_j. The calculation of the distribution of $p(d|c_j)$ might be very

difficult to do, especially when involving a large number of text documents in the classifying process due to its value can be very large. The value of $p(d|c_j)$ is equal to the sum of all word position combinations multiplied by the number of categories to be classified.

$$p(d|c_j) = \prod_i \; p(w_i|c_j) \tag{5}$$

Thus:

$$c_{MAP} = argmax\; p(c_j) \prod_i \; p(w_i|d_j) \tag{6}$$

The value of $p(c_j)$ and $p(w_i|c_j)$ are calculated during training process using the following formulas:

$$p(c_j) = \frac{n(c_j)}{n(\text{sample})} \tag{7}$$

$$p(w_i|c_j) = \frac{1+p(w_i,c_j)}{|V|+\text{count}\,(c)} \tag{8}$$

$n(c_j)$ is the number of documents in j category and $n(sampel)$ is the number of sample documents that used in training process, while $p(w_i, c_j)$ is an occurrence number of the word w_i in c_j category, $|V|$ is the number of all words in c_j category and $count(c)$ is the number of unique words in all training data.

IV. Discussion and Results

Application built in this article is embedded into web browser, in this case is Google Chrome, do several sub processes which are cleaning, stemming, weighting and classifying the content of the web pages. Fig. 1 shows the complete process of this study.

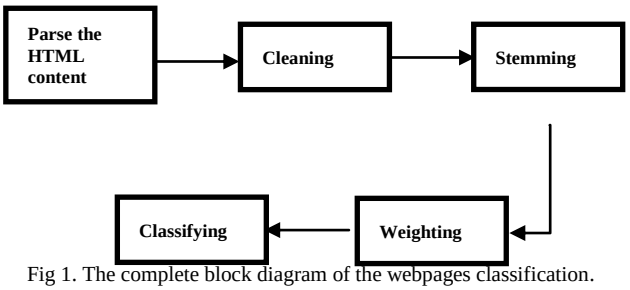

Fig 1. The complete block diagram of the webpages classification.

The first step of this research is to get the web page content that saved in the HTML document. It has known that a HTML was built as a structure with many tags in it. Once the HTML document is grabbed, then all the tags in it will be parsed.

The second step is the cleaning step which consists of two steps: tag cleaning and character cleaning. In the tag cleaning process, all tags that are not the main content will be removed. Removing the unwanted tag is recursive, which means that if a tag is removed due it is categorized as an unwanted tag then all the tag which nested in it will also be removed. As the last step in the cleaning process is to convert all the remaining text into lowercase.

The third step is stemming. At tokenization phase, all the remaining text will be split with whitespace as the delimiter, and then save them into the proper array. The next phase in the stemming process is stop words removal. All the words which not considered as keywords will be removed at this phase. This phase needs a list of common words as a reference to do the removal. There are two lists that used in this article, i.e. the common words list from Indonesian Dictionary website (indodic.com) and another list as an additional that contains words from the testing result. The stop word list from IndoDic.com containing 208 words and the additional stop words relating to pornography in Bahasa Indonesia are 1,316 words. Next, all words that composed of three or less character will also be deleted.

TABLE II. Rules of Affixes Separation

No	Word's Form	Separation
1	berV…	ber-V… | be-rV…
2	berCAP…	ber-CAP… ; C≠'r', P≠'e'
3	berCAerV…	ber-CAerV…; C≠'r'
4	belajar	bel-ajar
5	beC$_1$erC$_2$…	be-C$_1$erC$_2$…; C$_1$≠{'r', 'l'}
6	terV…	ter-V… | te-rV…
7	terCerV…	ter-CerV…; C≠'r'
8	terCP…	ter-CP…; C≠'r', P≠'er'
9	teC$_1$erC$_2$…	te-C$_1$erC$_2$…; C$_1$≠'r'
10	me{l|r|w|y}V…	me-{ l|r|w|y}V…
11	mem{b|f|v}…	mem-{b|f|v}…
12	mempe…	mem-pe…
13	mem{rV|V}…	me-m{rV|V}… | me-p{rV|V}…
14	men{c|d|j|s|z}…	men-{c|d|j|s|z}…
15	menV…	me-nV… | me-tV…
16	meng{g|h|k|q}…	meng-{g|h|k|q}…
17	mengV…	meng-V… | meng-kV… | (meng-V… jika V='e')}
18	menyV…	meny-sV…
19	mempA…	mem-pA… ; A≠'e'
20	pe{w|y}V…	pe-{w|y}V…
21	perV…	per-V… | pe-rV…
22	perCAP…	perCAP…; C≠'r', P≠'er'
23	perCAerV…	per-CAerV…; C≠'r'
24	pem{b|f|V}…	pem-{b|f|V}…
25	pem{rV|V}…	pe-m{rV|V}… | pe-p{rV|V}…
26	pen{c|d|j|z}…	pen-{c|d|j|z}…
27	penV…	pe-nV… | pe-tV…
28	peng{g|h|q}…	peng-{g|h|q}…
29	pengC…	peng-C…
30	pengV…	peng-V… | peng-kV… | (peng-V… jika V='e')}
31	penyV…	peny-sV…
32	pelV…	pe-lV… (except 'pelajar' → 'ajar')
33	peCerV…	per-erV…; C≠{l|m|n|r|q|y}
34	peCP…	peCP…; C≠{l|m|n|r|w|y}, P≠'er'
35	terC$_1$erC$_2$…	ter-C$_1$erC$_2$…; C$_1$≠'r'
36	peC$_1$erC$_2$…	pe-C$_1$erC$_2$…; C$_1$≠{l|m|n|r|w|y}

Description:
C : consonant
V : vowel
A : consonant or vowel
P : particle or fragment of word

After removing all the common words, the remaining words would be passed into the stemming process as an input. Broadly speaking, the stemming process is the process of removing the affixes from input word. In order to accelerate the stemming process, the algorithm modified adjusted to the morphology of Bahasa Indonesia. This modification cause the dictionary reading time can be reduced. Here is a list of things that are modified:

1. The addition of suffix –isasi, –wan, –wati, and –wi
2. The addition of prefix ku–, de–, and re–
3. The addition of invalid rules for affixes combination:
   a. be– and –i
   b. se– and –i
   c. di– and –an
   d. se– and –kan
   e. ke– and –i
   f. te– and –an
   g. ke– and –kan
   h. de– and –i
   i. me– and –an
   j. re– and –i

In the affixes removal phase, the suffixes will be deleted first then followed by prefixes as shown below:

1. Remove suffix –lah, –kah, –tah, and –pun
2. Remove suffix –u, –mu, and –nya
3. Remove suffix –i, –kan, and –an
4. Remove prefix di–, ke–, and se–

After that, regular expression process will be executed based on Table II.

After the stemming process is complete, the next step is doing a the weighting of the words present in the array where words in heading level 1 have the highest weight and words in heading level 7 have the lowest one followed by other tags. Table III shows the weight of each array.

TABLE III. THE GIVING WEIGHT TABLE

Tag name (array)	Weight per word	Tag name (array)	Weight per word
H1	8	H5	4
H2	7	H6	3
H3	6	H7	2
H4	5	Other tags	1

The last step in this study is classifying the web page based on the output of the stemming process using Naïve Bayes classifier. Before the Naïve Bayes algorithm ready to use, it must have passed the training and testing stage. There are 480 documents that used in training and testing stage, which 70% of these documents are used in training stage and 30% in testing stage, 288 documents are taken from "white websites" and 192 documents from "black websites". The black website is a website that reported contain pornographic, while the white website is the opposite. The documents from the white websites that used in this article are selected only the documents which contain one or more words that exist in the list of adult words. Selection of these documents is to be used for training and testing process to find out if the built application can distinguish those adult words are pornographic or not. Documents that been used in these stages can be seen in Table IV.

TABLE IV. DOCUMENTS THAT BEEN USED IN TRAINNING AND TESTING STAGE

Set	Stage	Number of documents
I	Training	Document number 1-336
	Testing	Document number 337-480
II	Training	Document number 145-480
	Testing	Document number 1-144
III	Training	Document number 1-168 and 313-480
	Testing	Document number 169-312
IV	Training	Document number 1-100 and 245-480
	Testing	Document number 101-244

As seen in Table IV, this article conducts the training and testing phase four times with the different set of documents in each phase. The probability of words and classes in training dataset is calculated in training phase. The training phase consists of several steps. First, calculate the power of $p(w|c)$ of every word in each training documents. Second, if the value is zero, then that word will be excluded from class calculation process. Thrid, ccalculate the value of $p(c)$ of each class. Fourth, ddetermine the class of each document.

To accelerate the calculation process in the training phase, then only words that do not exist in the training documents will be calculated its $p(w|c)$ value due the classification process of testing documents is similar with training documents. After that, the value of $p(w|c)$ of each word will be exponent to the weight (amount) of the word, then its results will be added into the calculation process of its class. The value of $p(w|c)$ has so many decimal digits, thus the class calculating process can be simplified using the logarithmic function as shown in Eq. 9.

$$c_{MAP} = argmax\left[\log p(c_j)\prod_i \log p(w_i|c_j)\right] \quad (9)$$

The calculation results will be negative, thus the final results are the classes that have calculation results closest to zero. The treatment of the real documents is the same as the testing documents.

TABLE V. CONFUSION MATRIX SET I

		Actual Class	
		Blocked	Not Blocked
Expected Class	Blocked	81	0
	Not Blocked	21	42

TABLE VI. CONFUSION MATRIX SET II

		Actual Class	
		Blocked	Not Blocked
Expected Class	Blocked	85	0
	Not Blocked	16	43

TABLE VII. CONFUSION MATRIX SET III

		Actual Class	
		Blocked	Not Blocked
Expected Class	Blocked	89	0
	Not Blocked	13	42

TABLE VIII. CONFUSION MATRIX SET IV

		Actual Class	
		Blocked	Not Blocked
Expected Class	Blocked	86	0
	Not Blocked	19	39

TABLE IX. OVERALL EXPERIMENT RESULTS

Dataset	Accuracy Level
I	85.41667%
II	88.88889%
III	90.97222%
IV	86.80556%
Average	88.02084%

V. CONCLUSION

The average of accuracy level of this study is below 90%, but it is quite high due to over 80%. There are several things that caused the error in the classification, among others is still many terms of foreign languages that used in the article in Bahasa Indonesia. To overcome this problem, it can be done by using an additional dictionary or stop words list that includes the terms of foreign languages, but this will cause the calculation process is slower due to the extra time for reading the dictionary (list).

Naive Bayes classification process is highly dependent on the existence of a dictionary and a list of stop words. The more complete the list of stop words that is used, the higher the level of accuracy gained. The problem is very difficult to make a complete list of stop words since the words used in "black website" are very diverse and irregular.

The stemming process is also influenced by the dictionary that is used. The stemming process can be accelerated by adding rules which are accordance with the morphology of Bahasa Indonesia to reduce the reading time of dictionary.

ACKNOWLEDGMENT

The authors thank Ministry of Research, Technology and Higher Education of Republic of Indonesia through Penelitian Produk Terapan program 2016 that funded and supported this research. The authors also thank Universitas Pelita Harapan Surabaya who provided the proper laboratory and library to do this research.

REFERENCES

[1] International Telecommunication Union, Measuring the Information Society report 2014, Switzerland: ITU, 2014.
[2] M. Yaffe and E. C. Nelson, *The Influence of Pornography on Behaviour*, London: Academic Press, 1982.
[3] Kamus Besar Bahasa Indonesia (KBBI), Pornografi. (2016) [Online]. Available: http://kbbi.web.id/pornografi
[4] A. Chazawi, *Tindak Pidana: Pornografi*, Surabaya: ITS Pres, 2009.
[5] R. Potter, Obscene modernism: Literary censorship and experiment 1900-1940, Oxford: Oxford University Press, August 29, 2013.
[6] J. Asian, *Effective Techniques for Indonesian Text Retrieval*, PhD Thesis, RMIT University Australia, 2007.
[7] S. Raschka, *Naïve Bayes and Text Classification: Introduction and Theory*, 2014.
[8] I. Rish, *An Empirical Study of the Naïve Bayes Classifier*, New York: IBM Research Division, 2001.
[9] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, London: Cambridge University Press, 2009.

Andreas was born in Malang, East Java, Indonesia in 1968. He received the undergraduate degree in Informatics from Sekolah Tinggi Teknik Surabaya in 1993, master degree in Information Technology from the same institution in 2007, and currently pursuing the Ph.D. degree in Institut Teknologi Sepuluh Nopember.

He is a lecturer of Information System Department (faculty of Computer Science) of Universitas Pelita Harapan Surabaya. His research interest includes machine learning, machinima, and information technology.

Lusia Permata Sari Hartanti was born in Pekalongan, Centra Java, Indonesia in 1984. She received the S.T degree in industrial engineering from Universitas Atma Jaya Yogyakarta, master degree in Mechanical Engineering in 2012 from Universitas Gadjah Mada, Yogyakarta.

She is a lecturer of Industrial Engineering, Universitas Atma Jaya Yogyakarta. Her research interest's includes industrial management and system, production planning and system and product design.