IMPLEMENTATION OF TEXT MINING ON KIDS INTERNET USAGE MONITORING APPLICATION "DODO KIDS BROWSER"

Firdaus Akhmad Muttaqin[1], Adam Mukaharil Bachtiar[2]

[1,2]Teknik Informatika - Universitas Komputer Indonesia
Jl. Dipati Ukur No. 112-116, Bandung 40132

E-mail: firdaus.akhmad66@gmail.com[1], adammb@outlook.com[2]

ABSTRACT

Dodo Kids Browser is a parental control software for search activities or surf the Internet by children. Supervision carried by blocking every word that has a negative context then a message appears on the mobile application belongs to the parents for give the action, however lack of information about the sentiment of the keywords being entered difficult for parents to know whether the keyword included on negative sentiment or not. It has an impact on the selection of action will be provided by parents. The application of text mining can be used as a solution.

Implementation of text mining is used for perform the classification process to search the child in obtaining information about the sentiment. Steps being taken for process the first classification is preprocessing of data. Furthermore, the results of the result data preprocessing algorithm applied to the Naïve Bayes classifier for the classification process. Classification results are displayed in the form of information about the advice in determining action by parents.

The results of text mining implementation of the system has been testing the functionality of the system, test the naïve Bayes classifier algorithm, and testing of some samples of test data. Results of these tests concluded that the system is able to provide information in the form of advice that can help parents in deciding pemberia action against her internet activity.

Key World: Text Mining, Sentiment Analysis, Naïve Bayes Classifier, Classification.

1. INTRODUCTION

Internet service as a medium of information is increasing has started to spread to all people, not just teenagers or adults but the kids were already using the internet as a media service information retrieval either for personal benefit or for education. It has a positive and negative impact, so there are several vendors that provide applications or services for monitoring and can restrict children's internet activities. Dodo Kids Browser is a software that serves as parental controlling for child's internet activity. This application can provide notification to parents when children do a search.

Based on observations made by trying the service provided on the application Dodo Kids Browser among them are content filtering on keywords entered the child when performing a search, the app will do the blocking on any keyword significantly negative so that each word has a negative meaning will always be subject to blocking even though the keyword entered has a positive meaning when it becomes a phrase or sentence. This causes problems in the availability of information that should be accessible to children but become can not be done because the entered keywords are words that are negative. For example, when a child doing a search in the English language with the keywords "how to avoid violence", the keyword being entered that contained the word "violence" which has a negative meaning for the child but if in a sentence, the keywords being entered has a meaning positive. This was due to the limited ability to generate conclusions of search keywords entered by a child. It can be difficult for parents to get a reference for determining the appropriate action to children.

Based on the outlined problem that needed a solution that can classify the keywords entered by the child when doing a search to produce a positive or negative conclusion of the keywords entered. This is possible with the use of text mining is a process that is semi-automatic classification of patterns derived from unstructured database. Results from the classification can be used as a medium to provide advice to parents in determining action against child when doing a search on the internet.

In doing classification there are many algorithms that can be used to classify the search keywords into the classroom negative or positive one that is naïve Bayes. Based on several studies regarding the naïve Bayes algorithm performance comparison with other algorithms concluded naïve Bayes has a 87.88% accuracy rate for categorical data were better than the accuracy of decision tree algorithm which has 84.85% [1]. Beside that, there is research on the application of naïve Bayes on spam classification of training data to 80 sms has an accuracy rate of 85.11%[2]. Based on that allows the naïve Bayes algorithm to be applied in classifying the search

keywords. Besides naïve Bayes algorithm is a conventional and simple therefore naïve Bayes suitable to be implemented in the child's internet usage monitoring application "Dodo Kids Browser".

1.1 Text Mining

Text Mining is a measure of text analysis is done automatically by the computer system to generate new information that has not been known previously taken from a series of texts which are summarized in a document [3]. Text Mining is a multi-disciplinary field involving information retrieval, text analysis, information extraction, clustering, Categorization, visualization, machine learning and other techniques [4]. Text mining using data mining application to convert unstructured data into structured data through the stages, namely [4]:

1. Text preprocess is solving a set of characters into words.
2. Feature Generation / Text Transformation is changing the words into a basic shape while reducing the number of words.
3. Feature Selection is the selection of features to reduce the dimensions of a collection of texts.
4. Text Mining / Pattern Discovery that can be unsupervised learning (clustering) or supervised learning (classification).
5. Interpretation / Evaluation that measurement to evaluate the effectiveness of methods applied using precision parameter.

1.2 Sentiment Analysis

Sentiment analysis or can be called opinion mining is the process of understanding, extracting and processing the textual otamatis text data to obtain information sentiment contained in an opinion sentence [5]. Sentiment analysis aims to determine the contents of a dataset shaped tesktual or sentence whether positive or negative sentiment worth [6]. Opinion mining can be considered also as a combination of text mining and natural language processing. Classification method is a method that can be used to solve problems on text mining. One of them is by using an algorithm Naïve Bayes Classifier (NBC). Natural language processing whereas befungsi to provide word class (tag) to each word in a sentence.

1.3 *Preprocessing*

A preprocessing stage before the classification process is necessary for cleaning, removing, changing the data source, whether it be a non-alphabetic characters and words are not needed. It is intended that the data used is optimal when used in the classification process. Preprocessing stages each case can vary. Here's a preprocessing stage and the explanation used in this study.

1. *Cleansing*

   Cleansing is the process of cleaning the data to be used from the characters and even the words are not needed. It aims to reduce the noise that can lead to the calculation process in the classification is not optimal.
2. *Case Folding*

   Case Folding is the process of converting data into the appropriate format. It aims to reduce redundancy of data that will be used in the classification process so that the calculation process becomes optimal. For example change the format of the data into lowercase or uppercase according to the needs required in the process of classification.
3. *Tokenizing*

   Tokenizing is a separation process or cut the data in the form of phrases, clauses, or sentences being said perkata based delimiters were used that space.

1.4 Naïve Bayes Algorithm

Naïve Naïve Bayes classifier is a classifier method which refers to the Bayes theorem is a theorem which refers to the concept of conditional probability. In this method required a combination of previous knowledge to new knowledge [7]. In carrying out the necessary classification training set as training data. At each sample from the training data has a class of its own label. The following is a mathematical model that is naïve Bayes classifier:

$$p(H|X) = \frac{p(H)p(X|H)}{p(X)} \text{ ... [2]}$$

Where:
X = Data with unknown class
H = hypothesis of data X is a specific class
p (H | X) = probability of the hypothesis H is based on the condition X (posterior probability)
p (H) = The probability of the hypothesis H (prior probabilty)

2. RESEARCH CONTENTS

Fill this study aims to describe a study conducted of the analysis process hinga implementation into the system. The following discussion of this study.

2.1 Analysis of The Problem

The problems that occurred in this study is the parents as users need to determine the appropriate action to searches conducted by children whether positive or negative, so that the necessary information classification search results in the form of suggestions for determining the action to be awarded.

2.2 Data Source

Source of data used in the form of a URL keyword searches a search engine. In conducting the search request, a search engine will do the data request using the GET method with an example by sending a parameter containing the keywords entered. Here is an example of a data source is presented in Tabel 1.

Tabel 1 Data Source

Key words	URL Format	Parameter name
Good people example	http://www.bing.com/search?q=good+people+example&go=Submit&qs=n&form=QBLH&pq=good+people+example&sc=0-0&sp=-1&sk=&cvid=f8323014b9c64795b68123461eb2a982	q
Kind of cas	https://www.google.com/search?q=Search+something&gws_rd=ssl	q
How to avoid violence	https://www.google.com/search?q=Search+something&gws_rd=ssl#q=what+do+i+search	q
How to bully people	https://www.google.com/search?q=How+to+bully+people&gws_rd=ssl	q
Example of violence	https://www.google.com/search?q=How+to+bully+a+people&gws_rd=ssl	q
Good violence	http://www.bing.com/search?q=Good+violence&go=Submit&qs=n&form=QBRE&pq=Good+violence&sc=8-10&sp=-1&sk=&cvid=166e131d89424cefa4e2aec4be4891fd	q

2.3 Preprocessing Implementation

Preprocessing process is done in order to transform the source data into the appropriate format and easy to do the classification process so that the classification process can be optimized. Preprocessing stages conducted in this study is from the cleansing process, case folding, and the last is tokenizing process. These stages seen on Gambar 1.



Gambar 1 Preprocessing Steps

Here is an explanation of the implementation of these stages.

1. Cleansing

At this stage cleaning of symbols and letters are not necessary. Additionally done also change certain symbols associated with search keywords that space format, which in this case space will be transformed into the + (plus), so do the conversion back into space. Step-by-step cleansing process is presented in the form of a flowchart which can be seen in Gambar 2.



Gambar 2 Cleansing Flowchart

Data input in this cleansing process in the form of a URL is generated when performing a search on a web browser. Here is an example of the application of the cleansing process is presented in Tabel 2

Tabel 2 Cleansing Proses

Input	Cleansing Result
http://www.bing.com/search?q=good+people+example&go=Submit&qs=n&form=QBLH&pq=good+people+example&sc=0-0&sp=-1&sk=&cvid=f8323014b9c64795b68123461eb2a982	*Good people example*
https://www.google.com/search?q=Search+something&gws_rd=ssl	*Kind of cas*
https://www.google.com/search?q=Search+something&gws_rd=ssl#q=what+do+i+search	*How to avoid violence*

Input	Cleansing Result
https://www.google.com/search?q=How+to+bully+people&gws_rd=ssl	*How to bully people*
https://www.google.com/search?q=How+to+bully+a+people&gws_rd=ssl	*Example of violence*
http://www.bing.com/search?q=Good+violence&go=Submit&qs=n&form=QBRE&pq=Good+violence&sc=8-10&sp=-1&sk=&cvid=166e131d89424cefa4e2aec4be4891fd	*Good violence*

2. Case Folding

At this stage of folding case made of converting data into a cleansing process results into the same shape. In this case the conversion is done be in lower case format. Here is a case folding process steps shown in flowchart form in FIG.



Gambar 3 Flowchart Cleansing Process

BBased on Gambar 3, The following is an example of the application process are presented in folding case Tabel 3.

Tabel 3 Implementation of Case Folding Process

Input	*Case Folding* Result
Good people example	*good people example*
Kind of cats	*kind of cats*
How to avoid violence	*how to avoid violence*
How to bully people	*how to bully people*
Example of violence	*example of violence*
Good violence	*good violence*

3. Tokenizing

Tokenizing is a stage split a combination of two words or more, or may be called as a phrase or sentence so that it becomes one by one. In this case the separation is done based on space as a delimiter. The following is a tokenizing process steps presented in the form of a flowchart in Gambar 4.



Gambar 4 Flowchart Tokenizing Process

Based on Gambar 4, The following is an example of the application of the tokenizing process served on Tabel 4.

Tabel 4 Implementation of Tokenizing Process

Input	*Tokenizing* Result
good people example	*good* *people* *example*
kind of cat	*kind* *of* *cat*
how to avoid violence	*how* *to* *avoid* *violence*
how to bully people	*Howe* *to* *bully* *people*
example of violence	*example* *of* *violence*
good violence	*Good* *violence*

Based on the table, the input data is the data results of the process of folding the case then

tokenizing process is carried out so that the resulting separation of each word.

2.4 Implementation of Naïve Bayes Algorithm

Stage In this stage, Naïve Bayes algorithm analysis process which is important in the classification of the sources of data on its sentiment is positive or negative. In this phase there are two main processes to do the classification is the process of learning and classification process. The following is an explanation of each process.

1. Learing Process

   In this process naïve Bayes classifier needs to be given prior knowledge to be used as a reference in order to perform the classification of the textual data based on sentiments. In the process of teaching or learning, there are three main steps. Here are the three main steps including its explanation.

   a. Determination of Data Class Practice

      At this stage, the determination of the class of data. Determination of the class is determined with the help of users by providing an opinion on whether the search keywords included in the positive class or negative class. Here is an example of the determination of class training data are presented on Tabel 5.

Tabel 5 Determining The Data Class

Data	Word	Sentiment Class
D1	food people example	Positif
D2	kind of cat	Positif
D3	how to avoid violence	Positif
D4	how to bully people	Negatif
D5	example of violence	Negatif
D6	good violence	Negatif

   b. Probability

      At this stage, probability calculations on the data that has been determined class. Tabel 6 the calculation of the probability of each class.

Tabel 6 Probability Accounting

Sentiment class	Count glasses						Probability
	D1	D2	D3	D4	D5	D6	
Positif	3	3	4	0	0	0	10/19
Negatif	0	0	0	4	3	2	9/19
Total	**3**	**3**	**4**	**4**	**3**	**2**	**1**

   c. Determining The Probability of a Item

      Once the probability of each class is calculated, then calculated the probability of each item. Here is the formula to calculate the probability per-item.

$$p(i) = \frac{f(i)}{f(c)}$$

      p (i) = Probability item
      f (i) = Frequency item
      f (c) = The total number of items based on class sentiments.
      The following is a calculation of the probability of each item presented on Tabel 7.

Tabel 7 Count Item robability

Data	Sentiment Class	
	Positive	Negative
good	$\frac{1}{10}$	$\frac{1}{9}$
people	$\frac{1}{10}$	$\frac{1}{9}$
example	$\frac{1}{10}$	-
kind	$\frac{1}{10}$	-
of	$\frac{1}{10}$	$\frac{1}{9}$
cats	$\frac{1}{10}$	-
how	$\frac{1}{10}$	$\frac{1}{9}$
do	$\frac{1}{10}$	$\frac{1}{9}$
avoid	$\frac{1}{10}$	-
violence	$\frac{1}{10}$	$\frac{2}{9}$
bully	-	$\frac{1}{9}$
people	-	$\frac{1}{9}$

2. Classification Process

   In this phase will be the classification of the new data, namely as test data using naïve Bayes classifier. Here is a plot of the classification process which can be seen in Gambar 5.



Gambar 5 clasikfikactnu Flowchart

Bsaed on Gambar 5 The initial stage is to input test data. Test data used comes from a data source that has been done preprocessing. Here is an example of test data presented in Tabel 8.

Tabel 8 Testing Data

Data Testing	
Before *preprocessing*	**After *preprocessing***
https://www.google.com/search?q=Violence+in+classroom &gws_rd=sal	*violence in classroom*

The next stage is a very important process that calculates the probability of each class, both negative and positive. The process will be explained as follows.

a. Count Probability
At this stage the probability calculation process by using naïve Bayes classifier. The test data of each class will be determined based on the probability value of the learning process. The following is the calculation process.

1). Probability calculations for positive

Ppositif = P_ (c +) * P ("violence" │positif) * P ("in" │positif) * P (classroom | positive)
Ppositif = 0.5263 * 1/10 * 1/10 * 1/10
Ppositif = 0.000005263

2). Calculations for the probability of negatif
Pnegatif = P_ (c -) * P ("violence" │negatif) * P ("in" │negatif) * P ("classroom" │negatif)
Pnegatif = 0.4737 * 2/9 * 1/9 * 1/9
Pnegatif = 0.001299588

b. Determing maximum sentimen.
From the results of previous calculations compared anatara Ppositif and Pnegatif value, obtained the highest score is Pnegatif so that it can be concluded that the search conducted classified into negative sentiment.

2.5 System Implementation

Implementation stage is the stage of implementation of the elements that have been performed on the stage of the analysis and design of systems to be implemented into a system. This phase includes the implementation environment, the implementation of the data, and interface implementation.Lingkungan Implementasi

a. Sistem Operasi Windows 8.1 Pro
b. WeBuilder 2014
c. MySQL DBMS
d. Visual Studio 2013
e. MySQL Workbench 6.3
f. StarUML 5.0.2.1570

Hardware spesification .

a. Processor Core i3 M380 @2.53GHz
b. RAM 6 GB
c. HDD 256 GB
d. Monitor LED
e. Keyboard dan Mouse

1. Data Implementations
The data involved in a system built that is derived from a series of text searches on a web browser by using extension. Here is the structure of the implementation of the data used in a system built served on Tabel 9.

Tabel 9 Data Usage

No	Collection	Data sent
1	Data Testing	a. uid: integer b. post_id : string c. kontent: string d. sugest: string e. id_notif: integer f. uri
2	Data Training	a. id_post: integer b. word: string c. status: string

2. Interface Implementation
Implementation of the interface contains the user interface of a system built results of the implementation of the previous design. The interface

is built to interface implemented on the application Dodo: Kids Browser Windows Phone mobile version in which the visualization of the results of the classification of the information that has been done in the form of advice for parents. Here is a view that implementation of the system built interfaces can be seen in Gambar 6.



Gambar 6 Interface of Application

3. CLOSING

Based on the results of research that has been done that the application of text mining in surveillance applications internet use child "dodo kids browser" by providing solutions in the form of a provision of information in the form of suggestions classification results from keywords used by children in a search on the web browser to help parents in giving action against the child when the child is indicated using keywords with a negative context has been implemented in accordance with the previous design and analysis. It can be concluded that the results of this classification is able to assist parents in obtaining information in the form of advice in determining the appropriate action for the child indicated to search when surfing the word implies bad.

BIBLIOGRAPHY

[1] D. Oktafia and D. C. Pardede, "Perbadingan Kinerja Algoritma Decision Tree dan Naive Bayes dalam Prediksi Kebangkrutan," UG Repository, Jakarta, 2014.

[2] E. A.W, M. and T. , "Penerapan Naive Bayes Untuk Sistem Klasifikasi SMS Pada Smartphone Android," EPrints 3 , Palembang, 2013.

[3] I. F. Rozi, S. H. Pramono and E. A. Dahlan, "Implementasi Opinion Mining (Analisis Sentimen) Untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi," *Jurnal EECCIS*, vol. 6, pp. 37-43, 2012.

[4] J. Ling, I. P. E. N. Kencana and T. B. Oka, "Analisis Sentimen Menggunakan Metode Naive Bayes Classifier Dengan Seleksi Fitur Chi Square," *E-Jurnal Matematika*, vol. 3, pp. 92-99, 2014.

[5] S. Andini, "Klasifikasi Dokument Teks Menggunakan Algoritma Naïve Bayes Dengan Bahasa Pemograman Java," *Jurnal Teknologi Informasi & Pendidikan*, vol. 6, pp. 140-147, 2013.

[6] A. Nurani, B. Susanto and U. Proboyekti, "Implementasi Naive Bayes Classifier Pada Program Bantu Penentuan Buku Referensi Matakuliah," *Jurnal Informatika*, vol. 3, pp. 32-36, 2007.

[7] S. F. Rodiyansyah and E. Winarko, "Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification," *IJCCS*, vol. 6, pp. 91-100, 2012.