75. HASIL CEK_60960140

by 60960140 Te

Submission date: 09-Aug-2022 12:34PM (UTC+0700)

Submission ID: 1880554093

File name: 75. TE-60960140-Identification of Cyber Bullying by using Clustering Methods on Social Media Twitter.pdf (1.11M)

Word count: 5247

Character count: 26728

Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Identification of Cyber Bullying by using Clustering Methods on Social Media Twitter

To cite this article: Nina Meliana *et al* 2019 *J. Phys.: Conf. Ser.* **1373** 012040

Identification of Cyber Bullying by using Clustering Methods on Social Media Twitter

Nina Meliana[1,*], Sunardi[1], Abdul Fadlil[1,*]

[1]Magister of Information Technology Program, Universitas Ahmad Dahlan, Jl.Prof. Dr. Soepomo S.H., Janturan, Yogyakarta 55164.

Corresponding Email: nmeliana@uad.ac.id, fadlil@mti.uad.ac.id

Abstract. In today's era social media is very important for some people, because it is the nature of social media that can make people addicted to using social media, this has been widely proven in terms of medical, social interaction has been reduced due to social media, things related to the physical to be reduced, for example going out, and playing outdoors for children, for adults social media can be positive and negative, positive if used to offer old friends who don't meet for long, but negative things can be used to crime or things that are not good, what else will soon be the presidential election, social media becomes a place or means of defamation of one party with another, the initial goal of social media is to build good relations Actually, there is a sentence that says "Yout thumb your tiger" with comments that are not good on social media. If we take us to the criminal section, and our words on social media can be legally justified, one example is intimidation, intimidation is one of the ITE Law articles, intimidation will be lifted from Twitter's social media, with Twitter we can see examples how much of this intimidation. In retrieving data on social media there are several methods used in retrieving data, the method used is the clustering method or data grouping method. The study used the Naive Bayes and Decision Tree J48 classification method. The results to be obtained in this study are the Naive Bayes method which obtained an average value of the overall scenario of 92% success rate and 8% not detected and the result of Decision Tree J48 method with an accuracy value of 100%. The most widely used type of cyberbullying is psychology related bullying or related to one's ignorance or nature

1. Introduction

Twitter is one of the fastest growing social networking sites nowadays where users can interact with other users from their computers or mobile devices anywhere and anytime. After its first introduction in July 2006, the number of Twitter users has increased very rapidly. As of September 2018, an estimated number of Twitter users registered around 326 million users. Twitter's own social media users consist of various groups of users who can interact with friends, family and colleagues. Twitter gives its users access to send a short message consisting of a maximum of 140 characters (called a tweet). The Tweet itself can consist of a text and photo message. It is through this tweet that Twitter users can interact more closely with other Twitter users by sending about what they are thinking, what is being done, about the events that have just happened, about the latest news and other things they have experienced. In 2018, the number of tweets posted almost reached 500 million tweets / day. In this paper, the author conducts research on data mining on social media, which aims to get the results and its correlation to cyber-bullying identification.

Published under licence by IOP Publishing Ltd

In the paper discusses the phenomena of fake news that occur in America, the spread of fake news through social media where social media has become the land to spread false news and in the discussion fake news has become an influential tool in US elections in 2016. False news is increasingly rising. Thus handling is needed in eradicating false news. One of them is using the "Fake News" detection tool where the tool can detect the truth of a news, but psychological factors are also needed so that people can be more careful in making or spreading false news. Overall the need to develop and apply methods in detecting false news is important and must be further developed so that the news becomes clear and ethical for the community [1]. Based on the research conclusions, it can be seen clearly that hoax news is a common enemy that must be anticipated. Public awareness of media literacy must be increased so that it can counteract the dangers of hoax news.

The hope of this researcher is that more research on this in more depth will have a significant influence on the wider environment or community [2]. In comparison to an automated approach to verifying information online with multimedia content can be presented with 3 different methods. The first method is the MCG-ICT method. The second method is the CERTH-UNITN method. The third method is the TweetCred method. There are several ways to overcome challenging verification, to measure verification by evaluating these methods using cross validation, and reporting scores per case. The MCG-ICT method has the highest precision score with a very low false positive rate [3]. With the support of social media, SPL can provide contributions, in the form of presenting an interface disaster report that is currently happening, and can also provide information such as evacuation points in the disaster area and the safe point of disaster victims. With social media support, SPL and describe information in real time about the location of the disaster that is happening [4].

In making news or information must be sourced from various speakers or informants and also compare experiments with each other so that the information produced does not favor one experiment. In the future it is expected that there will be an explanation and detection for the hoax news language style. People or news recipients can compare the truth of the news if they get more explanation about the characteristics of hoax news [5]. Fake news affects various fields such as the economy, law, environment and also one's mentality. The impact of fake news can be more dangerous than the impact of a fraud case. Fake news can also affect a person's level of trust in others. To reduce the existence of false news, it would be better if a group of people held a direct appointment rather than having to receive information in the mass media that was still doubtful. Because it could be that what was conveyed by Party A would be changed by the party B and received by Party C. Therefore, it is better for the A B C party to hold discussions directly [6].

According to this paper the news of hoaxes is the most serious threat that must be addressed immediately, because it can divide society. To be able to overcome it, there needs to be cooperation between the government and the community in order to reduce or eliminate the hoax news. One way we can do it is not to be easy to believe in news that is not necessarily the truth. [7]. According to this paper the netizens must be smart in avoiding the hoax news so that the news is not widespread. [8]. There is also a harmonic method leading to the smallest difference in classification starting from two different basic truths. But this harmonic method has two disadvantages. First, this is based on a very specific graph algorithm that cannot be extended to additional features in the same way as other machine-learning algorithms. Second, adjusting harmonious methods to provide good results is a very time-consuming process. [9] The goal in social media marketing is sincerity as the key to being able to communicate useful.

Disbelief in social media is due to the many fake news or hoaxes that are spread on social media that are not useful, so we as social media users must use social media as wisely and well as possible. [10]. So a comparative study for an automated approach to verifying information online with multimedia content has three methods, namely: the MCG-ICT method, using the highest score with some unique cases and very low false positive rates. So the accuracy is certainThe CERTH-UNITN method, produces consistently high results, and especially on events with lots of tweets. TweetCred method, From the method above to measure the accuracy of verification using cross validation and each case report itself. [11]. So Hoax news is very dangerous and will make it complicated with the emergence of debates. And will face legal problems if the news is declared Hoax [12]. Of the 13 million tweets, 15% of the stories were wrong, in the attempt to detect the initial hoax of social media, they were: 1. Early detection within minutes of the first report or first tweet; 2. Compile benchmark datasets 3. Use the word embeddings model for the entire dataset [13]. The journal above contains the perception of hoax news in various social media in the view of law and science. Then the youth also have responsibility and have high digital skills to protect the public from hoax news on social media. [14].

2. Experimental Method

Web pages clustering methods can be categorized into four types based on features employed: link-based, content-based, URL-based and structure-based. Link-based methods state that links are considered to be topically similar if they share similar parent nodes. Content-based approaches categorize web pages according to the words and sentences contained. These methods argue that pages with similar functions share similar topic features. Vector space model is commonly used to represent documents. However, content-based approaches are limited due to vagueness of topical differences and diffculties in understanding cross-language websites. URL-based methods have an advantage of clustering pages without downloading less. However, dealing with duplication and redirection problems is a non-trivial task. Once the patterns of URLs are decided, regular expression can be used to cluster and classify web pages. Structure-based approaches are more general ways to cluster web pages. Some previous research estimates the similarity of pairwise web pages through their tree structures. However, it is computationally expensive to do pairwise page comparison and not suitable for web-scale applications.

Another idea is representing web pages with vector space models. Two pages are considered to be structurally similar if they have common links placed in the same location. Minimum Description Length Principle (MDL) is adopted in combining small clusters. Further exploration on this idea has been done by exploiting Xpaths in recording positions of specific links. The method, known as UXCluster, is claimed to be more accurate and efficient than the method produced by pairwise distance. Another alternative is to cluster web pages represented by repetitive region patterns extracted by web wrappers. A drawback of most of structure-based clustering is that it requires heuristic threshold to decide whether to form a new cluster or to merge into an existing cluster [14].

We consider two common crawling Algoritma, crawling for user created content (crawling-for-UCC) and crawling for a target page type (crawling-for-target) to show how a structure-oriented sitemap supports different social media crawling strategies.

1. Crawling for UCC. Typically when crawling social media, some types of pages are considered less interesting or useful. For example, near-duplicates and non-user-created pages are less valuable than pages shared by users.
2. Crawling for target. Often it is desirable to focus crawling on pages that match the page type of an example page provided by some other process (a target page type) [15].

The clustering algorithm method in data mining can be used to find data clusters naturally derived from data extracted or examined using formulas from data mining.

2.1. Naive Bayes Classifier

Naïve Bayes Classifier is a classification method rooted in the Bayes theorem. Naïve Bayes algorithm is a classification method using probability and statistical methods. This method uses the Bayes theorem, which was discovered by Thomas Bayes in the 18th century.

The stages of the Naive Bayes algorithm process are:

a. Calculate the number of classes / labels.
b. Count the number of cases per class.

c. Multiply all class variables.
d. Compare results per class

Bayes Theorem Equation

$$\boldsymbol{P(C|X)} = \frac{\boldsymbol{P(X|C)P(c)}}{\boldsymbol{P(x)}} \tag{1}$$

where,
x = Data that is unknown to the class
c = The data hypothesis is a specific class
P (c | x) = Probability of a hypothesis based on conditions (posteriori probability)
P (c) = Probability of hypothesis (prior probability)
P (x | c) = Probability based on conditions in the hypothesis
P (x) = Probability c

The above formula explains that the chance of the entry of certain characteristic samples in class C (Posterior) is the chance of the emergence of class C (before the entry of the sample, often called prior), multiplied by the probability of the emergence of sample characteristic characteristics in class C (also called likelihood), divided by opportunities appearance of sample characteristics globally (also called evidence). Therefore, the formula above can also be written as follows:

$$posterior = \frac{prior\ x\ likelihood}{evidence} \tag{2}$$

Evidence values are always fixed for each class in one sample. The value of the posterior will be compared with the posterior value of the other classes to determine to what class a sample will be classified. Further elaboration of the Bayes formula is done by describing (c | x1, ..., xn) using the multiplication rule as follows:

$$\begin{aligned} &P(C|X1,\ldots,Xn)=P(C)P(X1,\ldots,Xn|C) \\ &= P(C)P(X1|c)(X2,\ldots,Xn|C,X1) \\ &= P(C)P(X1|c)P(X2|C,X1)(X3,\ldots,Xn|C,X1,X2) \\ &= P(C)P(X1|c)\ P(X2|C,X1)P(X3|C,X1,X2)\ \ldots\ P(Xn|C,X1,X2,\ldots,Xn\text{-}1) \end{aligned} \tag{3}$$

It can be seen that the results of the translation cause more and more complex factor factors that affect the probability value, which is almost impossible to analyze one by one. As a result, the calculation becomes difficult to do. This is where the assumption of very high independence (naive) is used, that each of the instructions is free from each other. With these assumptions, a similarity applies as follows:

$$P(c|X1,\ldots,X_n) = P(C)\prod_{i=1}^{n} P(Xi|C)$$
$$P(c|X) = P(x_1|c)P(x_2|c)\ldots P(x_n|c)P(c) \tag{4}$$

The above equation is a model of the Naive Bayes Theorem which will then be used in the classification process. For classification with continuous data, the Gauss Density formula is used:

$$P = (X_i = x_i|Y_i = y_i) = \frac{1}{\sqrt{2\pi\sigma ij}}e^{-\frac{(x_i-u_{ij})2}{2\sigma^2 ij}} \tag{5}$$

where,

P : Opportunity
Xi : Attribute to i
xi : Value of attribute to i
Y : Class sought
yj : Y sub-class searched
u : Mean, declares the average of all attributes

o : Standard deviation, expresses the variance of all attributes

$$\mu = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{x_i} \qquad \sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^n\right]0{,}5$$

Mean Formula Standard Deviation

2.2. Decision Tree J48 Classifier

This test uses classification using the Decision Tree J48 method. Decision Tree J48 is an implementation of the C4.5 algorithm that produces Decision Tree used in machine learning as a standard algorithm. Decision Tree is one of the classification algorithms in data mining data.

From the results of testing using the Decision Tree J48 algorithm that has been done through WEKA on the dataset, the results of the evaluation of the accuracy of classification of the Decision Tree J48 algorithm have an accuracy of 93.60% and error rate of 6.40%, as shown in the Table 1.

Table 1. Learning Dataset Table

No	Word bully	Type of bullying	cyberbuly	ID
1	anjing;	animals;	negatif;	non-bully;
2	babi;	animals;	negatif;	non-bully;
3	cebong;	animals;	positif;	bully;
4	cebong;	animals;	positif;	bully;
5	buta;	Someone's inability;	negatif;	non-bully;
6	buta;	Someone's inability;	negatif;	non-bully;
....	....	....	....	....
34	....	....	....	....

The table shows training data on category attributes to determine the word including bullying or non bullying.

$$\boldsymbol{Entropy(S) = \sum_{s-1}^{n} - pi * log2\ pi} \tag{6}$$

S : Set of cases
k : Number of partitions S
P_i : The probability obtained from the number (Yes / No) is divided by the total cases

Information Gain calculations for each formula attribute are:

$$Gain\ (A) = Entropi\ (S) - \sum_{i=1}^{k}\frac{|S_i|}{|S|} \times Entropi(S_i) \tag{7}$$

2.3. Stage of Data Mining

The stages carried out in the data mining process starts from the selection of data from the source data to the target data, the preprocessing stage to improve data quality, transformation, data mining and the stages of interpretation and evaluation that produce output in the form of new knowledge that is expected to contribute better. To choose an attribute as root, it is based on the highest Gain value of the existing attributes, as shown in the Figure 1.

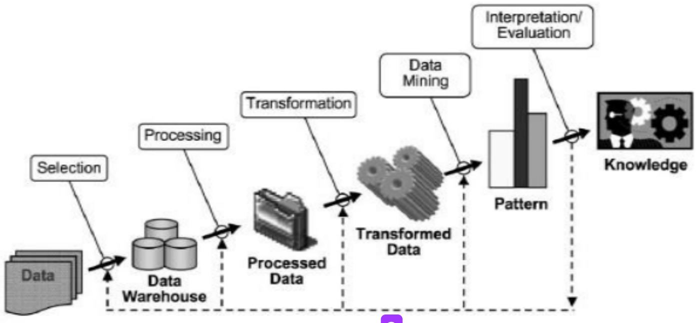


Figure 1. Stage Of Data Mining

1. Data selection

Selection of data from a set of operational data needs to be carried out before the information excavation stage in KDD starts. The selection data used for the data mining process is stored in a file, separate from the operational database.

2. Pre-processing / cleaning

Before the data mining process can be carried out, it is necessary to do a cleaning process on the data that is the focus of KDD. The cleaning process includes, among other things, removing duplicate data, checking inconsistent data, and correcting errors in data.

3. Transformation

Coding is a transformation process on selected data, so that the data is suitable for the data mining process. The coding process in KDD is a creative process and is very dependent on the type or pattern of information to be searched in the database.

4. Data mining

Data mining is the process of finding patterns or interesting information in selected data using a particular technique or method. Techniques, methods, or algorithms in data mining vary greatly. The choice of the right method or algorithm depends on the overall purpose and process of the KDD.

5. Interpretation / evalution

The pattern of information generated from the data mining process needs to be displayed in a form that is easily understood by interested parties. This stage is part of the KDD process called interpretation. This stage includes examining whether the pattern or information found is contrary to the facts or hypotheses that existed before.

3. Results And Discussion

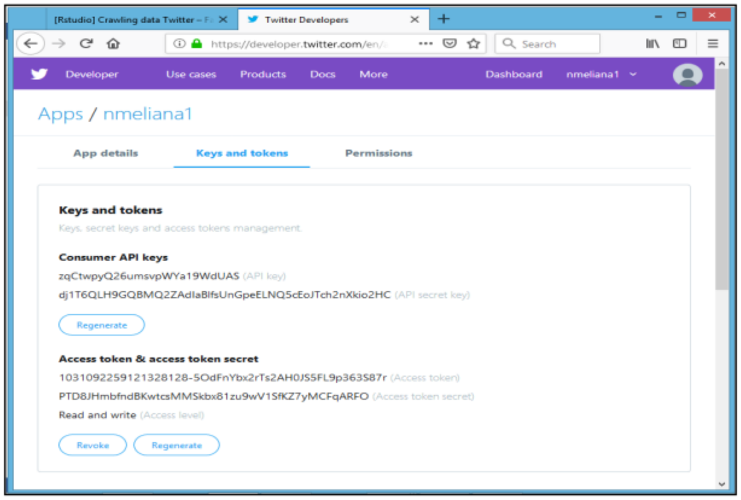
3.1. R Sistem Using Data Mining To Get Data From Twitter

R is a programming language and software system that is specifically and specifically designed to do everything related to computational statistical data that is so complicated. This programming language was first developed in 1993 by two statisticians who wanted to simplify statistical calculations, namely Ross Ihaka and Robert Gentleman at Auckland University, New Zealand. R programming language continues to grow rapidly along with the increasingly popular terminology "Big Data" and the increasing need for a company to be a data scientist to process and analyze data in the company as a basis for policy making and automation of business processes into data-driven. Programming languages like Python and R have become the main choice for researchers and users as well as practitioners in the field of data science to process and analyze data both for research and business purposes. Therefore, for someone just starting out in the field of data science, R is a programming language that is highly recommended for mastering.

3.2. Investigation Simulation

3.2.1. Collection

The collection phase of data is the stage of taking data (Tweet) from Twitter using the Rstudio program by utilizing the Twitter API (Application Programing Interface) as shown in the Figure 2.



Keys and tokens

Keys, secret keys and access tokens management.

Consumer API keys

zqCtwpyQ26umsvpWYa19WdUAS (API key)

dj1T6QLH9GQBMQ2ZAdlaBlfsUnGpeELNQ5cEoJTch2nXkio2HC (API secret key)

Access token & access token secret

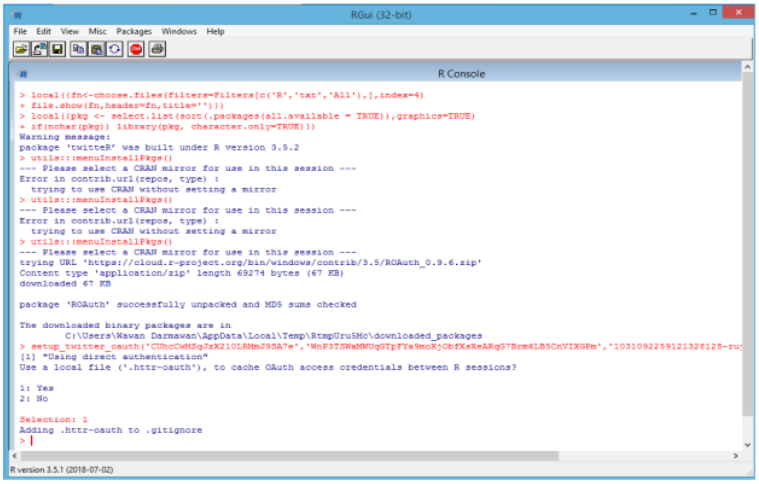
1031092259121328128-5OdFnYbx2rTs2AH0JS5FL9p363S87r (Access token)

PTD8JHmbfndBKwtcsMMSkbx81zu9wV1SfKZ7yMCFqARFO (Access token secret)

Read and write (Access level)

Figure 2. Twitter Developer API page

After getting the Twitter API key the next step is crawling or collection of data using Rstudio software. Previously installed necessary libraries such as library twitteR, ROAuth library, wordcloud library, and tm library, as shown in the Figure 3.


RGui (32-bit)
File Edit View Misc Packages Windows Help
R Console
R version 3.5.1 (2018-07-02)

```
> local({fn<-choose.files(filters=Filters[c('R','txt','All'),],index=4)
+ file.show(fn,header=fn,title='')})
> local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)
+ if(nchar(pkg)) library(pkg, character.only=TRUE)})
Warning message:
package 'twitteR' was built under R version 3.5.2
> utils:::menuInstallPkgs()
--- Please select a CRAN mirror for use in this session ---
Error in contrib.url(repos, type) :
  trying to use CRAN without setting a mirror
> utils:::menuInstallPkgs()
--- Please select a CRAN mirror for use in this session ---
Error in contrib.url(repos, type) :
  trying to use CRAN without setting a mirror
> utils:::menuInstallPkgs()
--- Please select a CRAN mirror for use in this session ---
trying URL 'https://cloud.r-project.org/bin/windows/contrib/3.5/ROAuth_0.9.6.zip'
Content type 'application/zip' length 69274 bytes (67 KB)
downloaded 67 KB

package 'ROAuth' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Wawan Darmawan\AppData\Local\Temp\RtmpUru5Mo\downloaded_packages
> setup_twitter_oauth('CUhcCwMSqJzX210LRMmJ9SA7e','WnP3T5WxMWUqGTpFYa9moXjObfXsXeARqG7Brm6LB5CnVIXGPm','1031092259121328128-ru
[1] "Using direct authentication"
Use a local file ('.httr-oauth'), to cache OAuth access credentials between R sessions?

1: Yes
2: No

Selection: 1
Adding .httr-oauth to .gitignore
> 
```

Figure 3. The console page on Rstudio

3.3. Examination

The next stage of the digital forensic investigation process is Examination, at this stage the data sources taken in the collection phase are processed from various data into more regular data because not all words reflect the meaning or content contained in a document. For this reason, an initial process is called preprocessing text.

Data obtained were 250 data, then cleansing data from irrelevant tweets to 167 relevant data. The data is then done manually labeling (tagging) based on predetermined categories. Based on these data can be calculated the number of types of categories. The dataset obtained will be divided into two parts, namely training data by 70% and testing data by 30%, as shown in Table 2.

Table 2. Classification Table

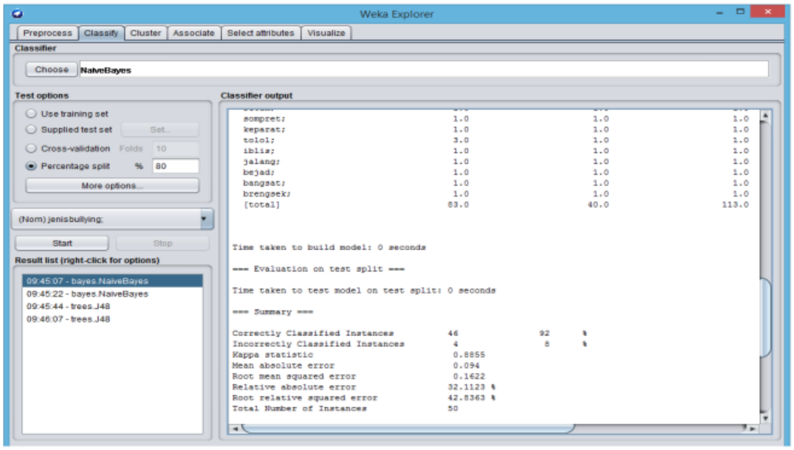
No.	Type of Cyberbullying	Bullying	Non Bullying
1	Psychology	52	3
2	Animal	43	42
3	General	8	4
4	Behavior	16	9
5	Someone's inability	48	25
	Sum	167	83
	Percentage	70%	30%

3.4. Analysis

The data obtained will then be implemented using the Naive Bayes classification method and Decision Tree J48. The classification process using Naive Bayes is done by preparing data that has been labeled as a reference or reference in the classification process.

3.4.1. Naive Bayes Mehtods

In the WEKA application, 80% percentage splits are filled. This means that 80% of the data becomes training data, the remaining 20% becomes testing data. The data validation uses 10 validations. The results get a success rate of 92% and an error rate of 8%, as shown in Figure 4.



NaiveBayes

○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 80

(Nom) jenisbullying;

09:45:07 - bayes.NaiveBayes
09:45:22 - bayes.NaiveBayes
09:45:44 - trees.J48
09:46:07 - trees.J48

```
  sompret;                        1.0                 1.0                 1.0
  keparat;                        1.0                 1.0                 1.0
  tolol;                          3.0                 1.0                 1.0
  iblis;                          1.0                 1.0                 1.0
  jalang;                         1.0                 1.0                 1.0
  bejad;                          1.0                 1.0                 1.0
  bangsat;                        1.0                 1.0                 1.0
  brengsek;                       1.0                 1.0                 1.0
  [total]                        83.0                40.0               113.0


Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          46               92      %
Incorrectly Classified Instances         4                8      %
Kappa statistic                          0.8855
Mean absolute error                      0.094
Root mean squared error                  0.1622
Relative absolute error                 32.1123 %
Root relative squared error             42.8363 %
Total Number of Instances               50
```

Figure 4. Testing the Naïve Bayes Method with a Percentage of 80%

The opportunity value obtained is:
P(Bully Type = P | wordbully= goblok) = 22 / 167 = 0.13
P(Bully Type = P | wordbully= gila) = 20 / 167 = 0.11
P(Bully Type = P | wordbully= idiot) = 6 / 167 = 0.03
P(Bully Type = P | wordbully= tolol) = 2 / 167 = 0.011
P(Bully Type = P | wordbully= saraf) = 2 / 167 = 0.011
P(Bully Type = P | wordbully= cebong) = 14 / 167 = 0.08
P(Bully Type = P | wordbully= kampret) = 8 / 167 = 0.04
P(Bully Type = P | wordbully= monyet) = 8 / 167 = 0.04
P(Bully Type = P | wordbully= babi) = 6 / 167 = 0.035
P(Bully Type = P | wordbully= anjing) = 5 / 167 = 0.029
P(Bully Type = P | wordbully= kunyuk) = 2 / 167 = 0.011
P(Bully Type = P | wordbully= sialan) = 3 / 167 = 0.017
P(Bully Type = P | wordbully= udik) = 2 / 167 = 0.011
P(Bully Type = P | wordbully= jancuk) = 2 / 167 = 0.011
P(Bully Type = P | wordbully= jelek) = 1 / 167 = 0.005
P(Bully Type = P | wordbully= item) = 0 / 167 = 0
P(Bully Type = P | wordbully= kampungan) = 0 / 167 = 0
P(Bully Type = P | wordbully= tai) = 0 / 167 = 0
P(Bully Type = P | wordbully= bangsat) = 4 / 167 = 0.02
P(Bully Type = P | wordbully= bajingan) = 2 / 167 = 0.011
P(Bully Type = P | wordbully= setan) = 2 / 167 = 0.011
P(Bully Type = P | wordbully= keparat) = 2 / 167 = 0.011
P(Bully Type = P | wordbully= jalang) = 2 / 167 = 0.011
P(Bully Type = P | wordbully= bejad) = 1 / 167 = 0.005
P(Bully Type = P | wordbully= iblis) = 1 / 167 = 0.005
P(Bully Type = P | wordbully= berengsek) = 1 / 167 = 0.005
P(Bully Type = P | wordbully= sompret) = 1 / 167 = 0.005
P(Bully Type = P | wordbully= lgbt) = 0 / 167 = 0
P(Bully Type = P | wordbully= banci) = 0 / 167 = 0
P(Bully Type = P | wordbully= jahanam) = 0 / 167 = 0
P(Bully Type = P | wordbully= biadab) = 0 / 167 = 0
P(Bully Type = P | wordbully= gembel) = 22 / 167 = 0.13

P(Bully Type = P | wordbully= budek) = 20 / 167 = 0.11
P(Bully Type = P | wordbully= kampungan) = 6 / 167 = 0.03

P(psychology) = 0.292
P(animals) = 0.230
P(general) = 0.044
P(behavior) = 0.084
P(Someone's inability) = 0.27

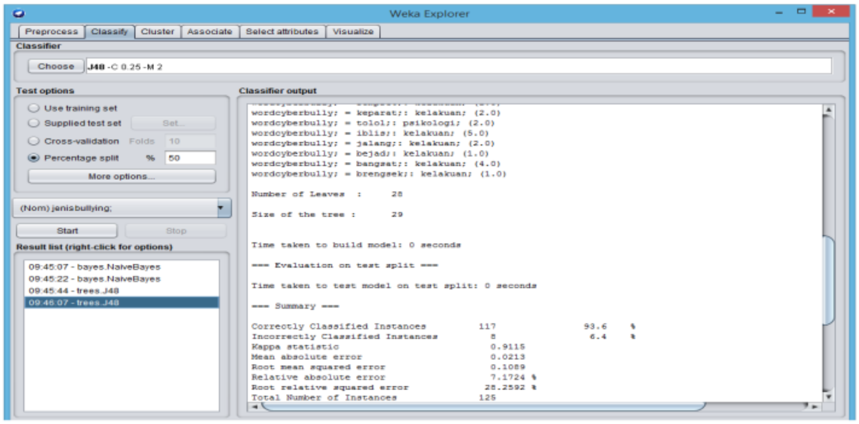
Cyberbullying(Positive) = P(psychology) × P(animals) × P(general) × P(behavior) × P(Someone's inability)
Cyberbullying(Positive) = (0.292) × (0.230) × (0.044) × (0.084) × (0.27) = **0.925**
Cyberbullying(Negative) = 1 – 0.925= **0.075**

Table 3. Probability Table (manual calculation)

Cyberbullying	Probability	Percentage
Positive	0.929263476	92 %
Negative	0.070736524	8 %

3.4.2. Decision Tree J48 Mehtods

From the test results using the Decision Tree J48 algorithm that has been done through WEKA on the dataset, the results of the evaluation of the accuracy of the classification of the Decision Tree J48 algorithm have an accuracy of 93.60% and an error rate of 6.40%. In the WEKA application, a percentage of 50% was filled. This means that 50% of the data becomes training data, the remaining 50% becomes testing data. Then for data validation use 10 times validation. The results get a success rate of 93.60% and an error rate of 6.40% and can be explained in the following Figure 5.



J48 -C 0.25 -M 2

● Percentage split % 50

10

(Nom) jenisbullying;

09:45:07 - bayes.NaiveBayes
09:45:22 - bayes.NaiveBayes
09:45:44 - trees.J48
09:46:07 - trees.J48

```
wordcyberbully; = keparat;: kelakuan; (2.0)
wordcyberbully; = tolol;: psikologi; (2.0)
wordcyberbully; = iblis;: kelakuan; (5.0)
wordcyberbully; = jalang;: kelakuan; (2.0)
wordcyberbully; = bejad;: kelakuan; (1.0)
wordcyberbully; = bangsat;: kelakuan; (4.0)
wordcyberbully; = brengsek;: kelakuan; (1.0)

Number of Leaves  :      28

Size of the tree :       29


Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         117               93.6    %
Incorrectly Classified Instances         8                6.4    %
Kappa statistic                          0.9115
Mean absolute error                      0.0213
Root mean squared error                  0.1089
Relative absolute error                  7.1724 %
Root relative squared error             28.2592 %
Total Number of Instances              125
```

Figure 5. Testing the J48 Decision Tree Method with a Percentage of 50%

In the second test, a percentage of 80% was filled. This means that 80% of the data becomes training data, the remaining 20% becomes testing data. Then for data validation use 10 times validation. The results get a success rate of 100% and an error rate of 0%.

3.5. Reporting

At this stage is the reporting phase of data collection until testing the data using the Naïve Bayes classification method and Decision Tree J48. From the classification that has been done, Naive Bayes has an accuracy rate of 92%, while the Decision Tree is 100%. Both methods have different levels of accuracy, but are still relatively good because they are still above 80%. If calculated based on this level of accuracy, both methods are quite appropriate for the classification of existing datasets, as shown in Table 4.

Table 4. Classification (Reporting)

Classification	Data Training		Data Prediction			
			Naïve Bayes		Decision Tree	
	Positive	Negative	Positive	Negative	Positive	Negative
Cyberbullying	63%	37%	92%	8%	100%	0%
Psychology	31%	2%	86%	14%	93.60%	6.40%
Animal	26%	25%				
General	5%	2%				
Behavior	10%	5%				
Someone's inability	29%	15%				

4. Conclusions

From this work, it can be concluded that (1) Stage of Detecting cyberbullying on Twitter can be done with several techniques. Starting from the Twitter API registration, lowering data with input access tokens from the Twitter API, storing data in the database, doing preprocessing analysis and data cleaning, classifying using WEKA and doing weighting and validation and classification with Naïve Bayes method and Decision Tree J48 Method. (2) Based on the accuracy obtained by each method in the detection of Cyberbullying on Twitter for each scenario, the method that produces the best accuracy is obtained by the Decision Tree J48 method with an accuracy value of 100% success rate, and for the Naive Bayes method the average value of the overall scenario is 92% success rate and 8% not detected. (3) The results of the classification can be seen that the highest type of cyberbullying on Twitter social media is the type of psychology or bullying related to ignorance / trait that is equal to 31%.

References

[1] J. Meinert, M. Mirbabaie, S. Dungs, A. Aker "Is It Really Fake? – Towards An Understanding Of Fake News In Social Media Communication", University Of Duisburg-Essen, Duisburg, Germany 2018.

[2] D. S. Adhiarso, P. Utari, S. Hastjarjo, "The Influence Of News Construction And Netizen Response To The Hoax News In Online Media", Universitas Sebelas Maret Surakarta 2018.

[3] C Boididou, S. E Middleton, Z. J. S. Papadopoulos, D. D. Nguyen, G. Boato, Y. Komppatsiaris "Verifying Information With Multimedia On Twitter" 2018.

[4] C Slamet "Social Media-Based Identifier For Natural Disaster", Et Al Iop Conf. Ser Mater. Sci. Eng. 288 012039, 2018.

[5] O. Ray, " A Survey On Natural Language Processing For Fake News Detection", Natural Sciences I College Of Arts And Sciences The University Of Tokyo 2018.

[6] A. Yap, Ph, " The Information War In The Digital Society: A Conceptual Framework For A Comprehensive Solution To Fake News", Academy Of Social Science 2018.

[7] M. K. Gunawan, A Wijaya, Salma, A. H. Idrus "Handling Of Hoax Messages From The Legal Perspective: A Comparative Study Between Indonesia And Singapore" 2018.

[8] D. S. Adhiarso, P. Utari, S. Hastjarjo "The Influence Of News Construction And Netizen Response To The Hoax News In Online Media" 2018.

[9] L. D. Alfaro, “Reputation Systems For News On Twitter: A Large-Scale Study” 6 Maret 2018 – 8 Januari 2018.

[10] T. Pesonen. “The Effects Of Fake News On Consumer Trust In Social Media Marketing” April 2018.

[11] C. Boididou, S. E. Middleton, “Verifying Information With Multimedia Content On Twitter”, 2018.

[12] A. Yap, Ph.D,”The Information War In The Digital Society: A Conceptual Framework For A Comprehensive Solution To Fake News”, 2018.

[13] A.Zubiaga. "Learning Class-Specific Word Representations For Early Detection Of Hoaxes In Social Media", 5-6, 2018.

[14] A. Shensa, MA, J. E. Sidani, MPH, PhD, M. A. Dew, PhD, C. G. E. Viera, MD, PhD, B. A. Primack, MD, PhD. " Social Media Use and Depression and Anxiety Symptoms: A Cluster Analysis", 5.101.220.175 on: Mon, 30 Jul 2018.

[15] K. Xu, K. Yingkai Gao, J. Callan. " A Structure-Oriented Unsupervised Crawling Strategy for Social Media Sites", arXiv:1804.02734v1, 8 Apr 2018.

75. HASIL CEK_60960140

ORIGINALITY REPORT

19%	18%	6%	16%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Politeknik Kesehatan Kemenkes Surabaya Student Paper	7%
2	download.garuda.kemdikbud.go.id Internet Source	6%
3	www.blogforlearning.com Internet Source	6%

Exclude quotes	On	Exclude matches	< 5%
Exclude bibliography	On		