Detection Of Cyberbullying On Social Media Using Data Mining Techniques

By Imam Riadi

WORD COUNT 4234 TIME SUBMITTED 18-JUL-2017 04:00PM
PAPER ID 30577517

International Journal of Computer Science and Information Security (IJCSIS),
Vol. 15, No. 3, March 2017

Detection Of Cyberbullying On Social Media Using Data Mining Techniques

Hariani
Department of Informatics Engineering
Islamic University of Indonesia
Yogyakarta, Indonesia
enenhazm@gmail.com

Imam Riadi
Department of Information System
Ahmad Dahlan University
Yogyakarta, Indonesia
imam.riadi@is.uad.ac.id

***Abstract*— Development and advancement of technology in addition to bring a positive impact also introduced new problems when used inappropriately. This is often referred to as cyber crime. One of cybercrime is being lively at the moment is cyberbullying. The practice of cyberbullying is not just limited to children but also in adults, it is called cyberstalking or cyberharrasment. Social media is one of the media for the development of cyberbullying, for example is Twitter. This research focus to analyze how much cyberbullying in Indonesia development on twitter and what type of cyberbullying is often used by abusers to do the bullying. the research was carried out using data mining techniques. Here there are several stages such as data collection, preprocessing, TF-IDF weighting, data validation and classification using Naive Bayes Classifier. the results showed that the content containing bullying is 86.97% and type of cyberbullying most used are related psychology that 61.63%.**

Keywords-cyberbcrime; cybebrullying; data mining; Social Media

I. INTRODUCTION

Cybercrime is as a form of tort law that made using the internet based on the sophistication of computer technology and telecommunications [1]. One of cybercrime is being lively at the moment is cyberbullying. Cyberbullying is an action or a teen intentionally intimidate, threaten or embarrass someone, or a group of other children through information technologies such as social media or mobile device [2]. Cyberbullying involves the use of information and communication technologies such as e-mail, cellular phone, chat rooms, social networking and personal is done deliberately and repeatedly, intended to hurt the other party [3]. Twitter is one of the widely used social networks, the rapid development of social networking twitter as communication tools are easy to use by anyone and can be accessed anywhere brings new trends in the community as the event to conduct an online or suppression of action known as cyberbullying. The abundance of the phenomenon of cyberbullying among communities that have a negative impact on result in either in law or in psychology makes the researchers interested in analyzing the type of cyberbullying is widely used, this research uses data mining approach which simply refers to the extraction of information or patterns of Yan's important or interesting data in database [3, 4]. Twitter log related to cyberbullying in the process of with text mining, namely the application of the data mining concepts and techniques for finding patterns in text, the analysis process of the text to summarize the useful information for a particular purpose. The methods used to process text mining algorithm is Naïve Bayes Classifier (NBC), this algorithm is simple and has a high speed in the training process and classification [5, 6].

II. BASIC THEORY

A. Cyberbullying

Cyberbullying is a term that refers to the use of information technology to bully people to send or post text that is intimidating or threatening [7]. According to a survey there is much cyberbullying development on social media [8]. the graph shown in figure 1:

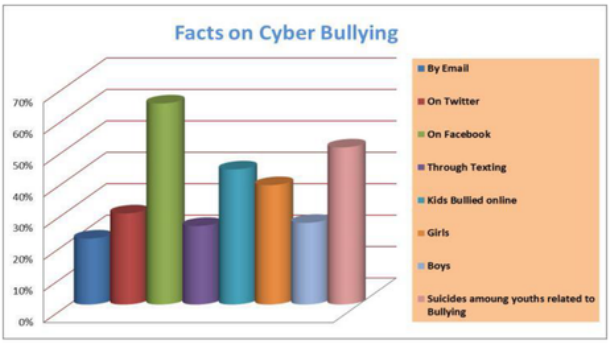

Figure 1 Graph Cyberbullying

Figure 1 shows the maximum percentage of the survey bullying done through on Facebook is 70%, on Twitter is 27% and email is 25%. About of children experienced bullying disorders is 48%, woman is 45% and children of 30%, while suicide victims around 55%.

According to a global survey conducted by Latitude News, Indonesia is a country with the world's second-highest bullying cases after Japan, and beat the United States in the bullying case occupies third place. Most bullying done through social networking. As the country with the fourth largest population in the world, Indonesia has the third largest number of Facebook users in the world and contributing 15 percent of daily tweets for Twitter. Based on studies of 91% of the respondents origin Indonesia claimed to have seen a case of Cyberbullying, and most often occurs through social media. In Indonesia, 74% of

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

respondents pointed to Facebook as being both the Cyberbullying, and 44% mention the other media website [9].

B. Data Mining

Data Mining is the application of algorithms for extracting the patterns from data and to provide useful knowledge for decision making. Data Mining has several applications in Digital Forensics. Data Mining involves identifying correlations in forensic data (association), discovering and sorting forensic data into groups based on similarity (classification), locating groups of latent facts (clustering), and discovering patterns in data that may lead to useful predictions (forecasting) [10, 11, 12]. Generally the process of data mining can be seen in the Figure 2:

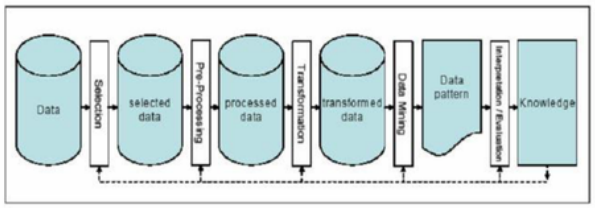

Figure 2 Data Mining Process

1. Data Selection
   The selection (selection) data from the operational database needs to be done before the excavation phase information begins. Data selection results will be used to process data mining, is stored in a file, separate from the operational data base.
2. Pre-processing/Cleaning
   Before the process of data mining can be carried out, the process of cleaning needs to be done on the data into focus. The process of cleaning include, among others, throw out the duplication of data, examine the data that were strongly inconsistent, and correct errors in the data, such as printing errors (typographical).
3. Transformation
   Coding is the process of transformation on the data that has been selected, so that data according to the process of data mining. The process of coding in KDD is a creative process and depends greatly on the type of information or patterns that will be searched in the database.
4. Data mining
   Data mining is the process of looking for patterns or interesting information in the selected data with the use of a particular method or technique. Technique, method, or algorithm in data mining is very varied. The selection of the right method or algorithm is heavily dependent on the purpose of the study.
5. Interpretation/Evaluation
   Pattern information generated from the process data mining needs to be displayed in a form that is easily understood by the parties concerned. This stage includes examination of whether patterns or information found at odds with facts or hypothesis that existed previously.

C. API Twitter

API stands for Application Programming Interface, is a supporting application of social networking, one of its functions, namely to extract data in JSON form. The API is a way for a program to complete the task, usually in a take and modify the data. Twitter provides a API at almost any feature that can be used to create applications, websites, widgets, and other projects that interact with Twitter. Communication between applications that are created with the Twitter API made through Hypertext Transfer Protocol (HTTP). Twitter API has several components, all of which are free. However, Twitter gives a number of Tweets to be had in an hour. Some parts of the Twitter API not require basic authentication use the form user name and password, and some authentication using OAuth. Most of parts the API on Twitter use the model of the REST, except for the streaming API, which always require a connection to the Internet. Data obtained from Twitter return into JSON and XML formats. But as time passes, returns the data in the form of XML is removed [13].

Twitter Search API is API that is used to run the search index real time from the last tweet. There are some limitations with the Twitter Search API, among other things:

a. Search API is not complete index all tweets but 1500 index tweets.
b. Search API cannot be used to find the tweet was more than a week.
c. Search API returns a maximum of 100 tweets was more than a week.
d. Query which is complex may not be successful Search does not support authentication which means all Queries were published anonymously.

D. Text Mining

Expressions of text mining is generally used to indicate the system analyze of the large data in the form of a natural language text and detect usage patterns or linguistic lexical in an attempt to extract useful information [14].

The process of text mining requires several stages, remembering the text data has characteristics that are more complex than the usual data. Based on the presentation [15] which States that generally a document has the following characteristics:

1. Text on Database has a large size (Large textual database)
2. Have a high dimension, one word one dimension (high dimension)
3. Contain the phrase and among other phrases with one phrase can have different meanings and interdependent of each other (dependency)
4. Many words/sentences contain ambiguous (Ambiguity)
5. Contain data noise, such as an abbreviation. Terminology and spelling mistake.
6. Contains structures that are not raw suppose abbreviations on words.

Process of extracting the information from a set of text documents such as web pages, twitter, and other documents need several interrelated processes. Processing unstructured

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

documents from being more structured by applying some of the techniques of extraction and filtering on words in the document at once by weighting the importance of words with weighting method

Results weighting data are then processed by using techniques of data mining in accordance with the purposes of the data processing. This section will explain the basics of the theory used in the extraction process and document with the weighting method of TF-IDF weighting.

E. Naïve Bayes Classifier

Naïve Bayes statistics analysis is an algorithm, which performs data processing numeric data using Bayesian probability. Classification of Bayes statistics classification which is able to predict the probability of a class member [16].

Broadly speaking, the workings of this method can be explained as follows:

1. Take the probability of positive and negative words.
2. Calculate the average probability of both
3. Specify the classification based on the value of the above probability.

To get the probability of each word, passing the learning each word and the probability. In this learning process, needed a training set, which is a set of sentences is positive and negative have been classified. Naïve Bayes classification technique is a simple and fast [17]. This technique works well with representation statistics. Unlike the method of rule-based, Naïve Bayesian can learn incrementally. But the shortcomings of the Naïve Bayesian vector is the size of the resulting feature quite large and need a technique to minimize the size of the vector [14].

For possible categories for a given document, here is an explanation of the Naïve Bayes [18]:

1. Any data represented as an –n dimensional vector i.e. $X = (x_1, x_2, x_3 \dots x_n)$ is a picture of the size that was created in a test of the n attribute i.e. $A_1, A_2, A_3 \dots A_m$ where m is a collection of categories $C_1, C_2, C \dots C_m$. Given given the data test X unknown category, then the classifier will predict that X belongs to the category with the highest posterior probability based on the condition of x. Accordingly, Naïve Bayes Classifier marks that the test X unknown earlier to the category C_1 if and only if it is seen in the Parallels:

$$P(C_1|X) > P(C_j|X) \text{ for } l \leq j \leq m, j \neq i \quad (1)$$

Then we need to maximize P (C_i | X) based on the equation

$$P(C_1|X) = \frac{P(X|C_1) \cdot P(C_1)}{P(X)} \quad (2)$$

2. Px is constant for all categories, only $P(X|C_1).P(C_1)$ needs to be maximized. If the prior probabilities may be estimated by calculation $P(C_i) = \frac{Si}{S}$ where Si is the amount of training data from the C_i category, and S is the total number of training data.
3. Given data with many attributes, this will be a complex computation for computated P (X |C_i). To reduce the computation at the time of evaluating the P (X |C_i), then it can be calculated using the equation:

$$P(X|C_i) = \sum_{k=1}^{n} P(x^k|C_i) \quad (3)$$

where X is the attribute values in the sample X and probability

P (X_1 | C_i), P (X_2 | C_i),……., P (X_n | C_i) can be estimated from the data of the training.

III. RESEARCH METHOD

This research consists of some stages. First stage is a collecting Log Data from twitter, and then do preprocessing or data cleansing that has been data in the crawl to make the acquired data are structured, next do TF-IDF weighting and validation data and the last do classification using Naïve Bayes Classification uses Machine Learning WEKA. As for the plot of this study shown in Figure 3:

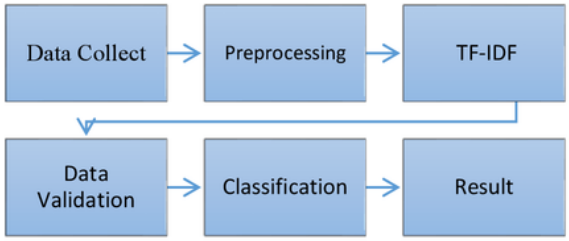

Figure 3 Research Flow

A. Log Data Collection

Data collection is done on twitter database. the data collected as much 1245 data, after a preprocessing amount of data into a 583 data. Crawling data on period November-December 2016. the technique of data collection can be seen in the figure 4.

Figure 4 Data Collection

First, login to twitter, do the registration API to get the access token, create a script and input access token has been obtained beforehand, do the search data with Boolean searching technique that is use the operators "AND", OR "," NOT" depending on the needs the data [16]. Keywords in the search data using a word that is often used to make bullying for example said, "bangsat (scoundrel)", "monyet (monkey)" and others. This method follows the previous researchers on research analyzing Gender Bullying so just based on keywords LGBT such as "gay" and "bitch" [19]. then lastly, data will be collected in the JSON file format.

B. Preprocessing

On the data that has been collected, change the JSON file to CSV file and next do preprocessing or data cleansing. at this stage do preprocessing in two stages, the first is manually and secondly, using machine learning WEKA. Processed manually

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

such as remove duplicate ID, remove special character URL, RT, has tag, picture, tokenizing. Make a dictionary slangwords, stemming, make a lexicon dictionary and change CSV file into the ARFF. Processed using WEKA as TF-IDF weighting, Stopwords, Casefolding and N-Grams. The process of preprocessing can be seen in the figure 5:

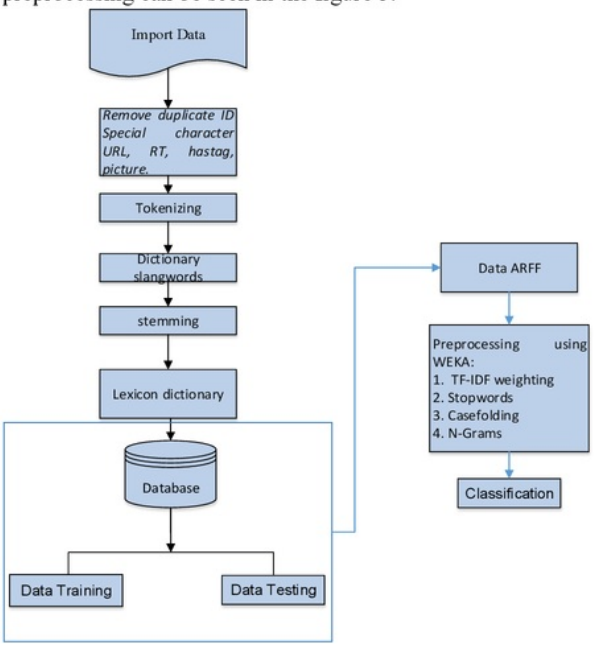

Figure 5 Preprocessing

1. Import data
   Import data from a database. the data is already CCV format.
2. Remove Duplicate ID and Special Character
   Remove duplicate ID from data has been crawl, and remove special character (.,/?[}-))13234786..etc.) and then remove a URL, Hastag, RT, and Picture.
3. Tokenizing
   Change the text in the form of a token, so easy to do normalization.
4. Dictionary Slangword
   Create a dictionary that contains the slangword and change slangwords become a formal word.
5. Stemming
   Process of stemming needed to take basic words from a word prefix.
6. Lexicon
   Create dictionary lexicon to identify containing words bullying.
7. Database
   Save data on database and divided become two stage, its data training and data testing.
8. Data ARFF
   Data training and data testing has modified into ARFF to be read by machine learning WEKA.
9. Preprocessing using WEKA
   This part, do automatic preprocessing, such us TF-IDF weighting, Stopwords, Casefolding and N-Gram.
10. Classification
    Classification by WEKA to see data results.

C. Data Validation

Data validation using 10 Fold Cross Validation from machine learning WEKA. The process can be seen on figure 7:

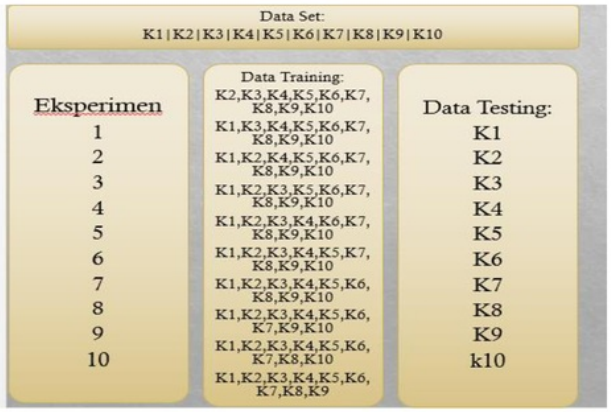

Figure 7 10 Fold Cross Validation Process

Data is shared with part 9 became 10/10 part is used for the process of training and 1/10 used for part of the process of testing. The iteration lasts 10 times with variations of data training and testing using a combination of 10 sections of data [20].

D. Classification

This stage is done the process of classification using machine learning WEKA, as for the classification process can be seen in the figure 6:

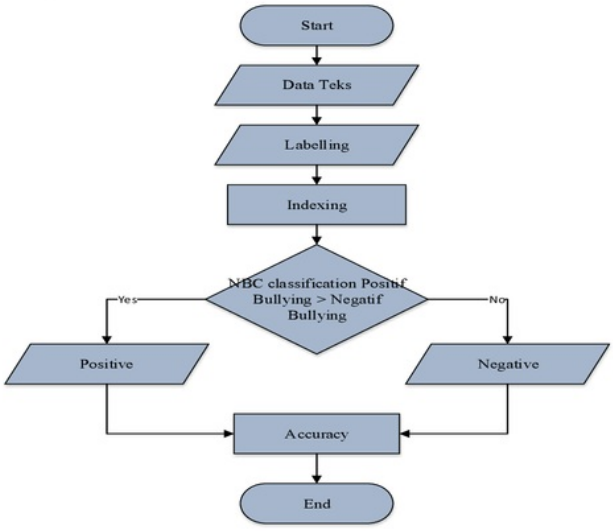

Figure 6 NBC Process

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Data text used is clean data that have gone through preprocessing, labeled training data do manually, change File CSV into ARFF, then do TF-IDF weighting and validation data using 10 fold cross validation and Then do classification using Naïve Bayes on WEKA, to the positive content of bullying going on Group on class result bullying, and the negative content to group on class result negative. So for this type content of cyberbullying, such as cyberbullying which related psychology will be in going on group a class related psychology and so on.

IV. RESULT

The results of the evaluation of the data base on the Division of the data set. The amount of data that is used is 583 data tweet. The data is divided into two parts, data training and data testing. the data are divide in a balanced, 50% for data training and 50% for data testing [20].

A. Cyberbullying and Non-Bullying

The process of classification specified in the pattern applied to the training data [21]. the pattern applied to detection of cyberbullying can be seen in the table I:

TABEL I. PATTERN OF CYBERBULLYING

BadWord!	Pronoun	…	…
You	BadWord	…	…
…	BadWord	Pronoun	…
Pronoun	BadWord	….	….
Pronoun	….	BadWord	…

The sentence using negation word that followed badword become negative bullying. for example: "your (pronoun) not (negation) rascal (badword) person" this sentence become negative bullying. "your (badword) rascal (badword) person" this sentence become positive bullying without negation . next, on the data training do manually labeling on a text containing cyberbullying and non-bullying. from the pattern above result classification on data training can be seen in the figure 8:

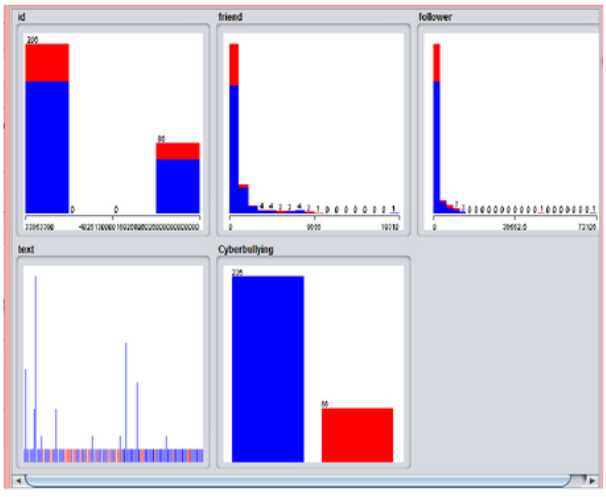

Figure 8 Classification Result on Data Training

The data used has some attributes that is ID, Friend, Follower, Text and cyberbullying, for attribute cyberbullying do it manually. on the graph can be seen that blue colors is content containing bullying and red colors is negative bullying. classification result can be seen on figure 9:

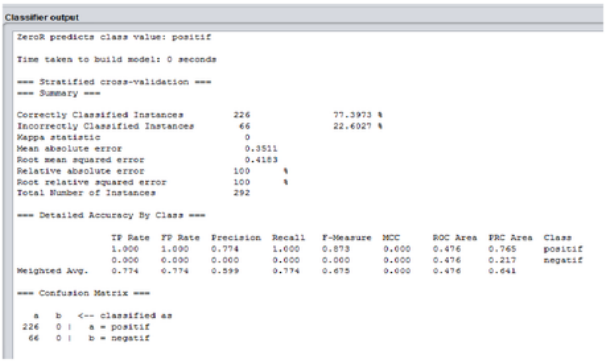

Figure 9 Classifier Output

There is content containing bullying us much 77.40% and content containing negative bullying is 22.60%.

Based on the pattern of training data results then obtained for data testing results as in the following figure 10:

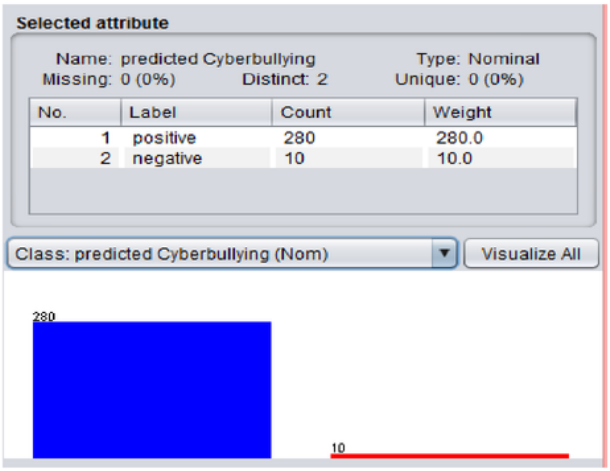

Figure 10 Prediction of cyberbullying using data testing

Prediction on data testing can be seen in the figure that the content containing of bullying is 96.56% and content containing negative bullying is 3.44%.

B. Types of Cyberbullying

Types of cyberbullying is bullying content types that are commonly used in social media. There are several kinds of cyberbullying to be on detection such us bullied related psychology (like word: "stupid","idiot" etc.), bullied related

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

animals (example in Indonesian word: "anjing (dog)","babi (pig), etc.), general bullying (like word: "damned","scoundrel", "devil" etc.), and bullied related sexuality (like word: "bitch" etc.) [22]. the process of classification same as in the above, different is on labeling in data training. For type related psychology grouping in the class related psychology and so on. The result classification can be seen in figure 11:

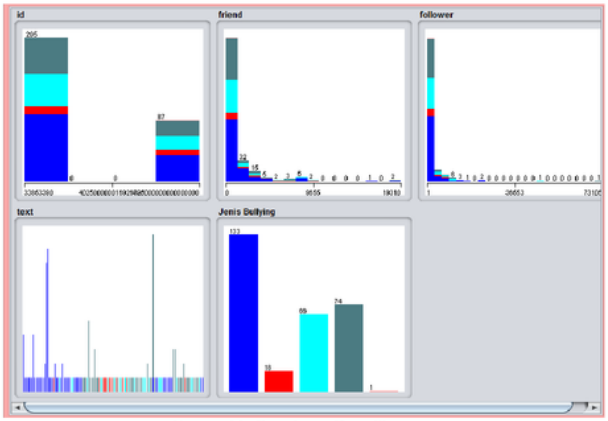

Figure 11 Result on Data Training

The results of the classification on data training it can be seen that type of cyberbullying is the most widely used bullying is related psychology (blue color) is 45.56%, and then general bullying (grey colors) is 25.34%, bullied related animals (red colors) is 6.16% and sexuality (pitch colors) is 0.34%. for prediction from data testing can be seen in the figure 12:

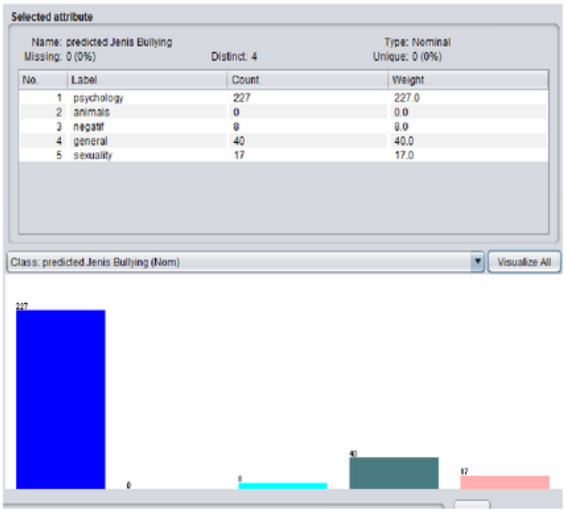

Fig 12. Result prediction types of cyberbullying

Figure 12 indicate that result prediction for related psychology is 77.73%, general bullying is 13.69%, sexuality is 5.82% and then related animals is 0.00%.

Exposure the result classification of cyberbullying and classification type of cyberbullying shown in table II:

TABLE II PRESENTATION OF CLASSIFICATION

Classification	Data Training NBC		Data prediction NBC		Presentation	
	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.
Cyberbullying	77.40%	22.60%	96.55%	3.44%	86.97%	13.02%
Psychology	45.56%		77.73%		61.63%	
Animals	6.16%		0.00%		3.08%	
General	25.34%	22.60%	13.69%	2.75%	19.51%	12.67%
Sexuality	0.34%		5.82%		3.08%	

On the Table II shows that cyberbullying on social media twitter on period November-December 2016 is very high with type of cyberbullying used much is related psychology.

V. CONCLUSIONS

Detection of cyberbullying on twitter can be done with some technique. First, Data collection is done by a few steps, do login twitter, do registering on API twitter to get the access token, and then create scripts for crawling data and input access token that has been obtained in API twitter, then save log data in the database in the form of JSON files. Second, doing the analysis preprocessing and data cleansing with the method described previously to get structured data. Third, classification using WEKA, classification is performed on the data that has been clean, TF-IDF weighting and validation data using 10 fold cross validation and then do classification. Result of classification can be see that cyberbullying on social media twitter on period November-December 2016 is very high with type of cyberbullying used much is related psychology.

VI. FUTURE WORK

It is recommended to perform the analysis of cyberbullying on different social media such as Facebook, Instagram, Youtube and others using different algorithms of data mining such us SVM, K-Means, C45 and others.

REFERENCES

[1] Hamzah, Andi, Marsita and Boedi, "Criminal Aspects in the Field of Computers," Sinar Grafika, 2012.

[2] M.Firman and A. N. Ngazis, "Cyberbullying Threat to Children on the Internet," January 2012. [online: access october 2016].

[3] Yumalita, "Cyberbully[illegible] on Social Networking Twitter (Trending Topic Semiotics Analysis)," Faculty of social and political sciences, University of Syiah Kuala, Banda Aceh Darussalam, 2016.

[4] Y. G. Sucahyo, "Data Mining: Dig Out Hidden Information," wsilfi.staff.gunadarma.ac.id, 2013.

[5] S. Susanto and D. Suryadi, "Introduction to Data Mining, Digging Up Data Chunks Of Knowledge," C.V. Andi Offset, Yogyakarta, 2010.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[6] N. W. S. Saraswati. Tagawa. "Text Mining with Naïve Bayes Method and Support Vector Machines for Sentiment Analysis" Thesis, Udayana Univesity Denpasar, 2011.

[7] R. Machsun. "The Phenomenon Of Cyberbullying In Adolescence," Journal Of The Science Of Library, Information, Archival Khizanah Al-Hikmah, 35-44, 2016.

[8] P. Singhal and A. Bansal. "*Improved Textual Cyberbullying Detection Using Data Mining,*" International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 3, Number 6, pp. 569-576, 2013.

[9] Satalina and Dina. "Cyberbullying Behavior Tendency Of Extrovert And Introvert Personality Type". Ejournal UMM, Vol. 02, No. 02, ISSN: 2301-8267, 2014.

[10] V. P. Kayarkar, R. P. Nirt and A. Motwani. "Mining Frequent Sequences for Email in Cyber Forensics Investigation," International Journal of Computer Application (0975-8887), Volume 85-No 17, 2014.

[11] M. Zulfadhilah, Y. Prayudi, I. Riadi. "Cyber Profiling Using Log Analysis and K-Means Clustering, International Journal of Advanced Computer Science and Applications Vol. 7 No.7, 2016.

[12] I. Riadi, J. E. Istiyanto, A. Ashari, Subanar. Internet Forensics Framework Based-on Clustering, International Journal of Advanced Computer Science and Applications Vol. 4 No. 12. 2013

[13] P. Aliandu. "Indonesian Language Tweet Sentiment Analysis On Twitter," thesis, Graduate Courses in Computer Science, Faculty of Mathematics and Natural Sciences, University of Gajah Mada. Yogyakarta, 2012.

[14] F. R. Sandi. "Classification Of Posts Twitter Traffic Jams Of Bandung Using Naïve Bayesian Classification" Thesis, Master of Study Program Computer Science, University Of Gajah Mada, 2012.

[15] Auvil, Loretta, Searsmith, and Duane. "IJsing Text Mining for Spam Filtering," Automated Learned Group National Center for Supercomputing Applications University of Illinois, 2010.

[16] N. Saputra. "Sentiment analysis-based Lexicon and emoticons," Thesis. Graduate School Of The Faculty Of Engineering University Of Gajah Mada, Yogyakarta, 2015.

[17] C. Darujati and B.A Gumelar. "Utilization Technique Of Supervised Classification For Indonesian Language Text," Jurnal Link Vol. 16 No.1. 2012.

[18] I. Nuraini, B. Susanto and U. Proboyekti. "Implementation of Naïve Bayes Classifier On The Aids Program is The Determination Of The Reference Book Subjects" access ti.ukdw.ac.id, 2011.

[19] H. Sanchesz and S. Kumar. "Twitter Bullying Detection," Dept of Computer Science UC Santa Cruz, 2011.

[20] A. G. Buntoro. "Sentiment analysis Hatespeech On Twitter using Naïve Bayes Classifier Method and Support Vector Machine Dynamics," Journal of Informatics, volume 5, No 2, 2016.

[21] D. Yin, Z. Xue and L. Hong. "Detection of Harassment on Web 2.0," Departement of Computer Science and Engineering Lehigh University, 2009.

[22] H. Margono, X. Yi and G. K. Raikundalia. "Mining Indonesia Cyber Bullying Patterns in Social Media," Proceedings of the Thirty-Seventh Australasian Computer Science Conference (ACSC), Auckland, New Zealand, 2014.

Detection Of Cyberbullying On Social Media Using Data Mining Techniques

ORIGINALITY REPORT

2%

SIMILARITY INDEX

PRIMARY SOURCES

1	"A New Index for Evaluating Academic Performance: Hos - index", International Journal of Computer Science Issues, 2017 Crossref	36 words — 1%
2	www.overcomebullying.org Internet	14 words — < 1%
3	Baojiang Cui, Shanshan He. "Anomaly Detection Model Based on Hadoop Platform and Weka Interface", 2016 10th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2016 Crossref	13 words — < 1%
4	acehresearch.org Internet	12 words — < 1%

EXCLUDE QUOTES OFF
EXCLUDE BIBLIOGRAPHY OFF
EXCLUDE MATCHES OFF