Crime Rate Mapping in Bandung City Area with Online Media Data by Using Naïve Bayes Method

Kautsar Aditya Rahman, Budhi Irawan, Casi Setianingsih, Ismatullah Hadi, Renaldy Eka Putra, Muhammad Hafizh Haekal, and Muhammad Idri Junando

***Abstract*— With online media, information from any part of the world can be obtained. High speed in providing information, making online media widely used by the community at this time. In that case, we can use it for certain interests such as research, search data, and collecting data. Based on the collection of data and information we can know the information we want such as crime in a city based on data that has been collected from some online media data. With the data already collected can be classified into the respective categories of news articles that have been collected using the naïve Bayes classifier method. From the research results of this research in classifying news articles about crime and the mapping of it obtained the best accuracy is 95%.**

***Index Terms*— Online Media, Applications and services, Naïve Bayes, Information technology, Classification.**

I. INTRODUCTION

In the current era of globalization, technology is developing very rapidly. This thing is closely related to the human need for information and technology. In order to complete their needs, humans use various methods and media. One of the human needs is the need for information. This information is of course obtained through mass and non-mass media. Media mass consists of television, radio, newspapers, magazines, tabloids and films. When today, the mass communication media that is developing very rapidly is online media (online newspaper). With online media, information from any part of the world can get. Its high speed in providing information, makes Online media is widely used by the public at this time. In that case that we can use for certain purposes such as research, searching for data, and collect data. Based on this collection of data and information, we can find out the information we want such as a crime somewhere city based on data that has been collected from several online media data [1].

Crime is a problem that is often presented in various media, be it electronic media to print media. This happens both in big cities and small towns, of course, from minor to serious crimes. Many ways that people try to avoid the crime such as using a man for security, but it doesn't longer consider as an effective method in this era [2]. Another way is using the GPS to get the information if there is a victim, and it can be reported to the police [3]. Also, for improving the security, Artha et.al, in their research about a new technique of room access security using brain computer interface where it is reading signal from an array of neurons and using signal processing to translate the signal into action [4]. But that research only for room and it's also important for the big city such as Bandung where it is one of the cities that has a fairly high crime rate. It is proven that on average there have been about 80 cases of conventional crime in the city of Bandung every week, one of one of which is due to the lack of awareness of the people of the city of Bandung itself [5]. The number of victims fell and only a small number of perpetrators could be arrested not only because there was an opportunity, but also because of the lack of fast action against crimes that occurred from related parties in particular. Big city resort police of Bandung. Based on this, people are now required to increase their own awareness of crimes in certain areas.

The platform used as information to the public about crimes in the Bandung area is Web-based with the consideration that there is no similar application on the platform. All information that will be provided to users or the public is obtained from social media by collecting a lot of data related to crimes that will be calculated and processed. The data is processed using Naïve Bayes to classify these crimes and provide suggestions to avoid crimes in a certain area.

Kautsar Aditya Rahman, Budhi Irawan, Casi Setianingsih, Ismatulah Hadi, Renaldy Eka Putra, Muhammad Hafizh Haekal Muhammad Hafizh Haekal, Muhammad Idri Junando is with School of Electrical Engineering Telkom University, Bandung, Indonesia (corresponding author to provide e-mail: kautsaradiya@student.telkomuniversity.ac.id).

II. RELATED WORK

Data mining is a statistical, mathematical, artificial intelligence, and machine learning technique that is useful for extracting and identifying useful information and related knowledge from large databases. Bayes' decision theorem is a statistical approach that is fundamental to pattern recognition. Naïve Bayes is based on the simplifying assumption that attribute values are mutually independent if given an output value. In other words, given the output value, the probability of observing collectively is the product of the individual probabilities [6].

The advantage of using Naïve Bayes is that this method only requires a small amount of training data to determine the parameter estimates needed in the classification process. The Naïve Bayes method distinguishes between constant string data

and continuous numeric data. This difference is seen when determining the probability value for each criterion, both criteria with string data values and criteria with numeric data values [7].

In their research, all educational institutions, especially schools, offer many scholarships to students regardless of whether they can achieve good results. Scholarships are designed to help reduce the cost burden for students who receive them. Not all students can apply for scholarships because of the large number of students who apply for scholarships and the many evaluation criteria. SMK Pasim Plus Sukabumi does not yet have a system to help identify grantees more effectively and efficiently. The Naïve Bayes algorithm method is expected to help determine potential recipients. The Naïve Bayes algorithm is one of the top 10 data mining methods in the most popular data mining classification among other algorithms. The Naïve Bayesian method also evaluates the possibility of classifying documents more than other classification methods in terms of accuracy and computational efficiency [8].

III. Research Method

A. Online Media

The general understanding of online media is any type of media format that can only be accessed via the internet which contains text, photos, video, and sound. In this sense, online media can also be interpreted as a means of online communication. A specific definition of online media is media that presents news journalistic works, and articles online. Technically, online media are telecommunications and multimedia-based media that are connected to the internet. Some categories of online media are portals, websites, blogs, online radio, online TV, and email [5].

B. Classification

The definition of news according to the Big Indonesian Language Dictionary (KBBI) is a story or information about events or events that can be conveyed in the form of articles. The news article is one type of text data that can be used for research in text classification. Text classification is a process of grouping documents into a certain class that has been previously defined so that it can be used to predict the class of a document that is not known to the previous class. There is a major problem in text classification, namely how to build a system that can determine the actual class in a document only by utilizing the information contained in the document. Text classification is the process of classifying text data into a predefined class or category automatically. The general function of text classification is as follows [7]:

$$\gamma=X\rightarrow C \quad (1)$$

Where γ is a collection of documents and X is a class or category. There are two types of text classification, namely supervised and unsupervised text classifications [9]. At the classification stage, classification will be carried out on the testing data based on the classification model that has been built previously. In this study, the type of text classification carried out by the author is included in the type of supervised text classification, because it applies the learning process to training documents that have class labels.

C. Data Mining

Data mining is a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify various useful related information from large databases. The term data mining is a discipline whose main goal is to find knowledge from the data or information that we have. Data mining, is also known as Knowledge Discovery in Database (KDD). KDD is an activity that includes the collection, and use of historical data to find regularities, patterns or relationships in large data sets. Data mining is divided into 2 techniques, namely unsupervised learning and supervised learning [10].

D. Text Mining

Text mining is mining carried out by computers to get something new, something that was not known before or to rediscover implicitly implied information to be used as new learning, which comes from information extracted automatically from various text data sources. Text mining is a technique used to deal with classification, clustering, information extraction and information retrieval problems. The work process of text mining adopts a lot of data mining research, but the difference is that the patterns used by text mining are taken from a set of unstructured natural languages, while in data mining the patterns are taken from a structured database. The stages of text mining in general are text preprocessing and feature selection [10].

E. Mapping

Mapping is a grouping of a collection of areas related to several geographic locations of the area which includes highlands, mountains, resources, and a potential population that have an impact on socio-cultural characteristics that have special characteristics in using the right scale. Another definition of mapping is a step that must be done in making map. the first step that must be done is to create data followed by data processing and presentation in the form of maps. So, from the two definitions above and adapted to this research, mapping is a data collection process that is used as the first step in making a map, as well as describing the distribution of certain natural conditions spatially, transferring the actual situation into a large map expressed on a map scale [11].

F. Text File

Text files are files that contain information that is made in text form. The data contained in the document can be in the form of information on words, sentences, numbers, names, and others that are used as input in the text data.

G. Naïve Bayes

In the data classification process, the Naïve Bayes Classifier (NBC) algorithm is used. A Naïve Bayes Classifier is a statistical classification algorithm based on the Bayes theorem.

This algorithm shows high predictive performance and obtains comparable results with other classification techniques, such as decision trees and artificial neural networks. The Naïve Bayes Classifier assumes that the existence of a feature in a class has nothing to do with the existence of other features. Suppose something is considered an apple if it is yellow, round, and about 3 inches in diameter. Although these features are interrelated with each other, NBC still considers these features to be independent.

H. Naïve Bayes

Accuracy is the level of comparison between the prediction and the actual value, the extent to which the value is close. While precision is the level of response between the information inputted by the user and the answers given by the system, and the last is recall, the level of success of a system in executing the given command.

I. TFIDF

The combination of the Term frequency method with the Inverse Document Frequency method gave birth to a combined method known as TF·IDF. This combined method is the result of multiplying the TF method with the IDF method. Salton proposed the combination of these methods with the aim of getting better performance. In this method, a high weight value will be given to terms that often appear in a document, but rarely appear in a collection of documents. When written in an equation, it can be written as follows:

$$TF*IDF(d,t)=TF(d,t)*IDF(t) \quad (2)$$

It can be seen from the above formula, that it can be concluded that the greater the similarity, the better if the TFIDF value is equal to zero, the term will not be taken.

IV. SYSTEM DESIGN & OVERVIEW

A. System Overview

In this study, a data mining classification technique will be applied using the Naïve Bayes algorithm. This algorithm works to create a pattern that will represent the classification. The case study that will be discussed in this research is to identify a collection of data about news articles which will be grouped into several categories such as murder, robbery theft, and hate speech. The input given is a text containing articles about crimes and the output is a mapping of crime rates from the results of the classification system. The overview system can be expressed in Fig. 1.

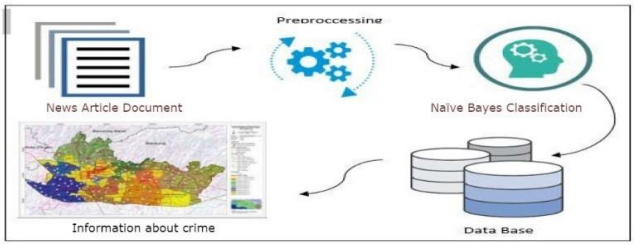

Fig. 1. System Overview

B. Preprocessing Flow Chart

Fig. 2 is a flow chart of the preprocessing processing flow in obtaining training data and test data.

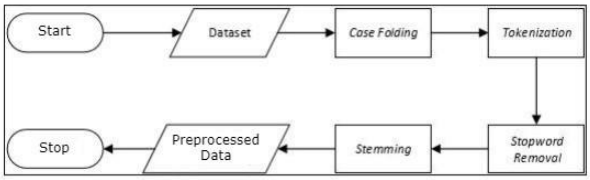

Fig. 2. Preprocessing Flow Chart

Case folding is changing all letters in the document to lowercase, only letters 'a' to 'z' are accepted. Characters other than letters are omitted and are considered delimiters. Tokenizer accepts input string and sorts it into tokens (smallest units) as document identifiers with the following rules: -Tokens are separated by whitespace characters (spaces) - T a b c (such as “!“, ?“, .“, ,“) is omitted - A token starts with a letter or number The output of tokenization is a token as well as additional information such as word frequency, word position in the document. While the filtering stage is the stage of taking important words from the term results. Can use a stoplist algorithm (remove words that are less important) or wordlist (save important words). Stoplists/stopwords are non-descriptive words that can be discarded in the bag-of-word approach. The stemming stage is the stage of finding the root word of each filtered word. At this stage, the process of taking various word formations into the same 10 representations is carried out.

C. TFIDF Flow Chart

Fig. 3 is a flow chart of the weighting flow of each document.

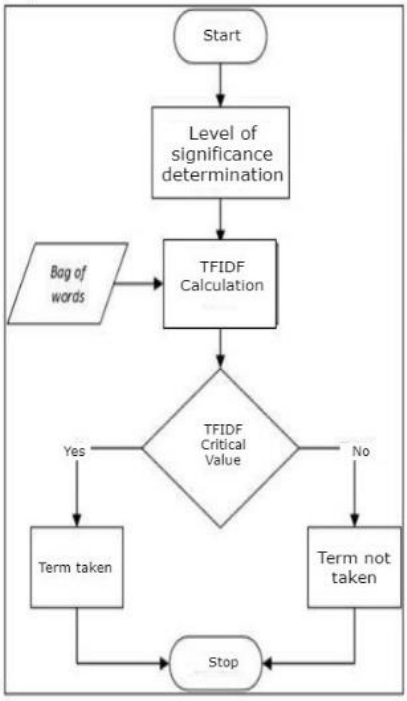

Fig. 3. TFIDF Flow Chart

D. *Naïve Bayes Flow Chart*

Fig. 4 is a flow chart of the Naïve Bayes classification process.

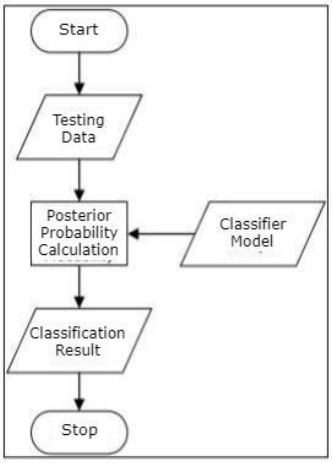

Fig. 4. Naïve Bayes Flow Chart

In this process, Naïve Bayes will classify documents by counting the number of occurrences of words from each document that will be compared in each category.

V. THE RESULT

A. *Implementation*

In the implementation section of this system, we will discuss the design of the application display for the user so that the user can check the crime rate in the Bandung city area and find information related to crimes such as prevention etc. In Fig. 5, an image of a map with location conditions of different colors indicating a level of crime in a certain area is shown.

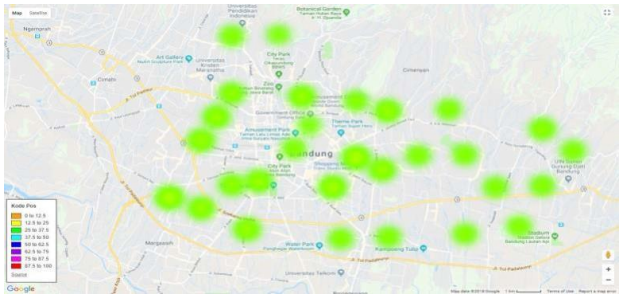

Fig. 5. Application Interface

B. *Testing*

This test serves to determine whether the performance of the system can function properly, by comparing manual and automatic calculations. The process that will be carried out first is to read the data set in Table 1 and then process it using TFIDF with the calculations in Table 2 after doing the calculations, a new data set is obtained that will be used for the classification process.

TABLE I
SUMMARY RESULT OF SIMUALTION

No	D1	D2	D3
1	*Aan*	*Abu*	*Abang*
2	*Abc*	*Acak*	*Aceh*
3	*Abdullah*	*ad*	*Adalah*
4	*Ada*	*ada*	*Ada*
⋮	⋮	⋮	⋮
1273	*Tkw*	*sangka*	*Zahabi*
1373	*Zikir*	*simpan*	*Null (blank)*
1654	*Null (blank)*	*Zona*	*Null (blank)*

In Table 1 is a display of some of the data sets that will be used where D1 is the contents of the data set for the murder category, D2 for the theft category and D3 for the drug category. The first document has 1373 words, the second document is 1654 and the last document is 1273 which is the identity of each category. In the next process, the initial data set will be processed by TFIDF before entering the classification process which will be processed in Table 2.

TABLE II
TFIDF PROCESS

Term	TF			DF	IDF	W=TF*IDF		
	D1	*D2*	*D3*			*D1*	*D2*	*D3*
aan	1	0	0	1	0.4771	0.4771	0	0
abang	0	1	0	1	0.4771	0	0.4771	0
abc	1	0	0	1	0	0	0	0
abdullah	1	0	0	1	0	0	0	0
abg	1	0	0	1	0.4771	0.4771	0	0
abu	1	1	0	2	0.4771	0.4771	0.4771	0
ada	1	1	1	3	0	0	0	0
. . . .	. .	. .	. .	. .	. . .	. . .	. . .	. .
zafenya	0	1	0	1	0.4771	0	0.4771	0
ziarah	0	1	0	1	0.4771	0	0.4771	0
zikir	1	0	0	1	0.4771	0.4771	0	0
zona	0	1	0	1	0.4771	0	0.4771	0

Table 2 is an example of some calculations in the TFIDF process, after reading the entire data set, in this process the word identity for each category will be taken if the weight or term value is not equal to zero. For example, the word aan has weight or has a value, the identity of the word aan will be owned by document 1 because it appears only in document one while the word exists appears in every document, the value of the weight is equal to zero, the word there will not be used or will be lost. After carrying out the TFIDF process, a new data set will be obtained which will be used for the classification process. The classification process can be calculated manually in the manner in Table 3.

TABLE III
FILL IN THE DATA SET AND THE DATA TO BE TESTED FOR CLASSIFICATION

	D1	D2	D3
Train	*Aan, abc, abdulah.....zikir*	*Abu, acak, ad,.....zefanya*	*Abang, adil, agk.....zahabi*
	Total words d1 = 926	*Total words d2 = 1205*	*Total words d3 = 824*
Test	*Bandung warga bandung harap hati-hati..... rabu*		

In Table 4 there is data to be tested which will be classified in certain categories. In accordance with the following stages:

- Calculating class priors

$$P(c) = \frac{Nc}{N} \quad (3)$$

where P(murder), P(theft), and P(drugs) are 1/3, respectively.

- Calculating conditional probabilities for the classification process

$$P(W|C) = \frac{Count(W,C) + 1}{Count(C) + |V|}$$

Where Count(W,C) is The probability of each attribute of W and C, Count(C) is The total number of probabilities of W and C; |V| is Sum of possible values of W,C. therefore, if |V| is 2955; Count (Murder) is 926; and Count (bandung, murder) is 0, then P (bandung | murder) is 0.000240.

After doing all the words on the test data set as above, conclusions will be obtained by multiplying the probability value of each word in its category and multiplying by the prior value of each class as shown in the Table 4.

TABLE IV
OVERALL CALCULATION RESULT

Class	Formula	Result
P (murder)	P(action|murder) *……. *P(inhabitant|murder)*P(murder)	2.9845×10^{-161}
P (theft)	P(action|theft) *……. *P(inhabitant|theft)*P(theft)	1.8238×10^{-154}
P (drugs)	P(action|drugs) *……... *P(inhabitant|drugs)*P(drugs)	3.0952×10^{-162}

In drawing conclusions from the calculation, that is by looking at the results of the greatest probability value. It can be concluded that the news articles tested through manual calculations can be classified into the Theft category. After performing manual calculations, it will be compared with the results of the classification system which will be shown in Fig. 6.

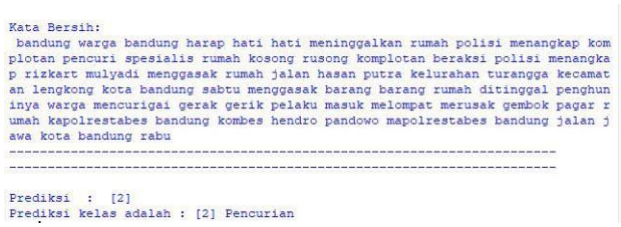

Fig. 6. Screenshot Results of Article Classification Testing Automatically

This system uses 90 articles of online media data taken from detik.com which has 3 categories, namely, murder, theft, and drugs. In each category, there are 30 news articles. After all news data passed the data pre-processing stage, all data were combined and used as a dataset.

The system performance has been tested based on the data partition size for further testing. There have 5 tests that have been divided proportionally which are 50/50, 60/40, 70/30, 80/20, and 90/10 for the train data and test data from dataset. The results of the test is how much the precision, recall, and accuracy of a percentage. The summary of each 5 tests show in Table 5.

TABLE V
SUMMARY OF THE RESULTS BASED ON DATA PARTITION

No. test	Precision	Recall	Accuracy
1	96%	96%	95%
2	95%	94%	94%
3	93%	93%	92%
4	95%	94%	94%
5	92%	89%	88%

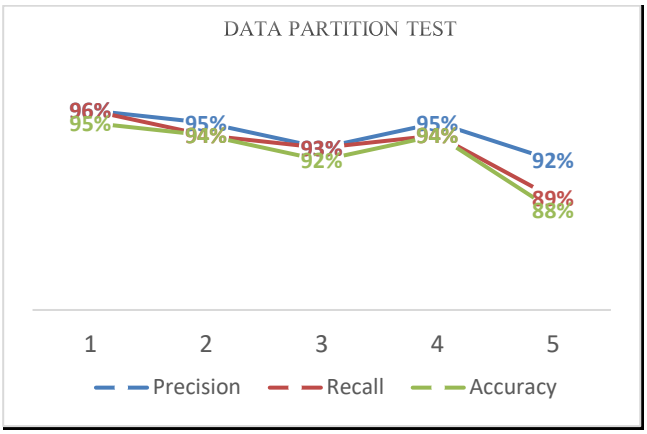

Fig. 7. Bar graph of test results based on the data partition

From the graph results on Fig. 7 above, we can conclude why the accuracy value decreases is that the training data used has ambiguity in each document because every news article has almost a lot of similarities in each class that has been determined.

VI. Conclusion

Based on the results of the tests and analyses that have been carried out, the conclusions obtained are as follows: (i) The Naïve Bayes method can be implemented and realized in classifying news articles; (ii) Test result and system accuracy are matched with the output of Bandung City Resort Police, (iii) The results of the Naïve Bayes test on the system with manual calculations show the appropriate results.

References

[1] E. Draught, F. Fang, P. Sistla, C. Yu, and W. Meng, "Stop word and related problems in web interface integration," *Proc. VLBD Endow*, vol. 2, no. 1, pp. 349-360, 2009.
[2] P. Sihombing, J. T. Tarigan, B. Ginting, and D. Sitompul, "Security System Based on Vibration and Infra-Red Sensors Using Raspberry," *Internetworking Indonesia Journal*, vol. 11, no. 1, pp. 11–16, 2019.
[3] S. Liawatimena and J. Linggarjativ, "Vehicle Tracker with a GPS and Accelerometer Sensor System in Jakarta," *Internetworking Indonesia Journal*, vol. 9, no. 2, pp. 9–15, 2017.
[4] A. I. Simbolon, M. F. Amri, M. A. Suhendra, and A. Turnip, "A New Technique of Room Access Security based Brain Computer Interface," *Internetworking Indonesia Journal*, vol. 11, no. 1, pp. 63–67, 2019.
[5] M. Dede, I. Setiawan, and A. Mulyadi, "Application GIS to analyse crime risk in Bandung," *In Proceeding of The 2nd International Conference in Sociology Education* (ICSE), October, 2017.
[6] J. C. Wyatt and P. Taylor, "Decision Support Systems and Clinical Innovation," *Getting Research Findings Into Practice*, pp. 123–137, 2018, doi: 10.1177/014107680009301206.
[7] Z. M. Ali, N. H. Hassoon, W. S. Ahmed, and H. N. Abed, "The Application of Data Mining for Predicting Academic Performance Using K-means Clustering and Naïve Bayes Classification," *International Journal of Psychosocial Rehabilitation*, vol. 24,no. 3, 2020.
[8] S. N. Rahman, A. I. Jamhur, Y. Elva, and E. Rianti, "Comparison of the Effectiveness of C. 45 Algorithm with Naive Bayes Algorithm in Determining Scholarship Recipients," *International Conference on Computer Science and Engineering* (IC2SE) IEEE. Vol. 1, pp. 1-5, November, 2021.
[9] D. L. Olson, and D. Delen, "Advanced data mining techniques," Springer Science & Business Media, 2008.

[10] A. D. Hartanto, E. Utami, S. Adi, and H. S. Hudnanto, ”Job seeker profile classification of twitter data using the naïve bayes classifier algorithm based on the DISC method,” *4th International Conference on Information Technology, Information Systems and Electrical Engineering* (ICITISEE) IEEE, pp. 533-536, November, 2019.

[11] S. Hajrahnur, M. Nasrun, C. Setianingsih, and M. A. Murti, ”Classification of posts Twitter traffic jam the city of Jakarta using algorithm C4. 5,” *International Conference on Signals and Systems* (ICSigSys) IEEE. pp. 294-300. May, 2018.