# Analysis of Text Mining Clustering on Suara Surabaya Crime Report with DBSCAN Neural Network Autoencoder Algorithm

**Grace Lucyana Koesnadi[1], Muhammad Rizal Anggriawan[1], Talitha Zuleika[1], Mochamad Rasyid Aditya Putra[1], Najwa Khoir Aldawiyah[1], and M Fariz Fadillah Mardianto[1*]**

[1]Statistics Study Program, Department of Mathematics, Universitas Airlangga, Surabaya, Indonesia
*Corresponding author: m.fariz.fadillah.m@fst.unair.ac.id

**Received:** 28 June 2024
**Revised:** 3 July 2024
**Accepted:** 22 July 2024

**ABSTRACT** – **Criminality, or crime, is a behavior that violates the law or is contrary to applicable values and norms. A high number of criminal behaviors criminal behaviors in a community significantly impacts its social conditions, leading to a decrease in welfare, unrest, and material losses that pose a threat to an individual's life. This study examines text mining on crime report data from Suara Surabaya using the DBSCAN clustering method and the Neural Network Autoencoder. The neural network autoencoder algorithm effectively reduces the data dimension, with an input dimension of 300 and an encode dimension of 64. Clustering analysis using the DBSCAN method based on the silhouette coefficient value criterion resulted in three clusters, with cluster 1 dominating the report. The clustering results show essential patterns in complaint reports, and LDA analysis reveals critical topics in the report. Cluster 0 shows a diversity of reports focusing on motor loss, interaction with homes or properties, and people's entry into homes. Cluster 1 is more focused on the loss of vehicles, both cars and motorcycles, with specific details such as vehicle color, number, brand, and related transactions or social interactions. Meanwhile, cluster 2 focuses on reports related to interactions with police stations and information on the location of incidents. This text mining approach to community crime report data not only improves analysis accuracy and efficiency, but also provides essential information that can support efforts to handle and prevent crime.**

**Keywords**– **Text Mining Clustering, DBSCAN, Autoencoder Neural Network, Criminality.**

## I. INTRODUCTION

Security and order are fundamental aspects of community life. Surabaya, one of Indonesia's metropolitan cities, faces various challenges in maintaining security stability. Crime reports compiled by various media, including Suara Surabaya, are essential for providing information about security conditions to the public and authorities [1]. However, the number of crime reports continues to increase as the population grows and urbanization makes manual analysis increasingly ineffective. Therefore, efficient methods are needed to process and analyze the data to generate insights that can be used to improve crime prevention efforts.

Text mining is one technique that can overcome this problem by extracting valuable information from the text of crime reports. Using text mining, initially unstructured data can be transformed into information that is easier to analyze [2]. One of the main challenges in analyzing crime report data is grouping data based on specific patterns that may not be visible to the naked eye. For this reason, clustering techniques such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) can be used. DBSCAN can identify clusters based on data density and handle noisy data or outliers well, making it suitable for application to crime report data that is often not homogeneous. However, a more sophisticated approach is needed to improve the clustering process's accuracy and efficiency. This is where the role of neural network autoencoders becomes relevant. Autoencoder is an artificial neural network used for unsupervised learning [3]. This type can reduce the dimension of the data, reduce noise, and find a more profound representation of the features of the data. Combining neural network autoencoders and DBSCAN allows the clustering process to be carried out more effectively, resulting in more accurate and meaningful data groups.

This study aims to group textual data in the form of community complaint reports related to crime in the city of Surabaya and how to depict the security conditions of crime in the city of Surabaya. This approach can produce a model that can classify crime reports with higher accuracy to provide better insights for crime prevention. The results of this research are expected not only to make scientific contributions in data mining and artificial intelligence but also to provide practical benefits for the police and the community in improving the security of Surabaya City. This research also has direct relevance to the Sustainable Development Goals (SDGs), especially on point 16 regarding peace, justice, and resilient institutions. SDGs point 16 focuses on reducing all forms of violence and violence-related deaths everywhere, as well as ensuring access to justice for all and building effective, accountable, and inclusive institutions at all levels. By analyzing crime reports more effectively and accurately, this study's results can help devise better strategies for crime prevention and maintaining public order, ultimately contributing to achieving SDG 16 goals.

Previously, research has been conducted using a text mining approach to identify and classify cybercrime datasets in cyberbullying cases [4]. The research showed that the classification using the Support Vector Machine (VCM) method produced an accuracy value of 87.14% [4]. Another study conducted a grouping of crime-prone areas in East Java

Province in 2021 using the ward clustering method [5]. The study resulted in 3 (three) sub-district clusters with similar characteristics based on crime indicators with six variables used [5]. This research requires developing data on crime indicators that affect the level of crime in an area, specifically [5]. In addition, other studies are using the k-means and *k*-medoids clustering methods in classifying crime by province in Indonesia, where each method produces 2 (two) clusters with high crime and low crime classification [6]. Then, another study maps the level of crime in Karawang Regency using the k-means algorithm [7]. The study concluded that 3 clusters were obtained, divided into non-vulnerable areas, vulnerable areas, and very vulnerable areas with the k-means algorithm using the silhouette coefficient [7].

This research can help authorities allocate resources more efficiently and responsively to crime-prone areas by providing deeper insights into crime patterns through text mining and clustering methods. This approach not only increases the effectiveness of law enforcement but also supports efforts to create a more peaceful and secure society in line with the global agenda for sustainable development.

## II. LITERATURE REVIEW

### A. Criminality

Crime is all kinds of economically and psychologically detrimental actions and acts that violate the applicable laws in the Indonesian state as well as social and religious norms [8]. The factors that cause crime are biological, sociological, and economic. Changes in these factors spur a person's personality to force themselves to commit crimes, from low crime to high crime. Criminologists assume that deviant behavior is a crime that must be explained by looking at the structural conditions in society in the context of inequality of power, authority, and prosperity and its relation to various economic and political changes that exist in society [9]. According to the Code of Criminal Procedure (KUHP) the Republic of Indonesia, the types of crimes are as follows [10].

1) Crimes against life, namely murder.
2) Crimes against the body, namely severe abuse, minor abuse, and domestic violence.
3) Crimes against morality, namely rape and molestation.
4) Crimes against human freedom, namely kidnapping and employment of minors.
5) Crimes against property/goods with the use of violence, namely theft with violence, theft with violence using firearms, and theft with violence using sharp weapons.
6) Crimes against property/goods, namely theft, theft with aggravation, theft of motor vehicles, destruction/destruction of goods, intentional arson, and seizure.
7) Narcotics-related crimes, namely the circulation and use of narcotics and psychotropics.
8) Crimes related to fraud, embezzlement, and corruption, namely fraud/fraudulent acts, embezzlement, and corruption.
9) Crimes against public order.

### B. Artificial Neural Network (ANN)

An artificial neural network is an information processing system inspired by the nervous system. This network is an information processing system with characteristics similar to those of biological neural networks, which were created as a generalization of the mathematical model of human cognition [11]. This system is made up of a large number of interconnected processing elements (neurons) working simultaneously to solve a specific problem. A neural network is a processor distributed in parallel that tends to store the knowledge it gains from experience and make it available for use [12]. The layers that make up the artificial neural network are divided into an input layer, a hidden layer, and an output layer.

The way neurons in the nerve network work is similar to the nervous system in the human brain. The information will be transmitted with a specific weight and then processed by a propagation function. The sum of these weights will then be compared to a particular threshold value (threshold) through the activation function of each neuron. The neuron will be activated if the input crosses a particular threshold value. Otherwise, it will not be activated. When a neuron is activated, it sends an output through its output weights to all the neurons associated with it, and so on. The ANN architecture is generally divided into four: Single-Layer Feedforward Networks (SLFN), Multi-Layer Feedforward Networks (MLFN), recurrent networks, and lattice structures.

### C. Text Mining

Text mining is one part of the field of data mining. Text mining is the process of extracting patterns from various data sources by identifying patterns of interest. In the case of text mining, a data source is an unstructured set of textual data on a document [13]. This unearthed information will become a new fact that can be further researched. Text mining is different from web searching. Web search aims to rediscover information written by someone or has existed before, while text mining aims to find unknown information that is not yet known and cannot be written [14]. Text mining can solve problems such as processing, organizing/grouping, and analyzing large amounts of unstructured data [15].

### D. Text Preprocessing

Text preprocessing is the process of normalizing text so that the information contained is dense and concise but still represents the information contained in it. In text preprocessing, unnecessary words are reduced, which have no meaning

in the text database or document, to make the data more structured and ready to be processed [16]. In this stage, there are several processes, namely:

1) Cleansing

Cleansing is a process to ensure that all text data used in the analysis is clean, relevant, and free of unnecessary information, commonly referred to as noise. Several stages will be carried out in the cleaning process, including:

- Parsing or decomposition breaks down a series of documents into separate components. At this stage, we determine which document unit to use according to the desired implementation needs.
- Case folding is a stage that changes all letters in a document to lowercase letters. Only the letters 'a' through 'z' are accepted. Characters other than letters are omitted and considered delimiters [17]. Changing characters to lowercase letters is done to ensure data consistency.
- Removing irrelevant characters means removing characters such as punctuation digits and irrelevant symbols. For example, using exclamation marks can show personality traits, so punctuation can be eliminated to increase statistical strength when modeling.
- Stopword removal or filtering, is a vital word selection process. Two methods are used, including stoplist/stopword, which is to prepare a set of words that are not descriptive (unimportant). Words included in the stoplist will be discarded and not used in the following process. The second method is a wordlist, which is the opposite of a stoplist. In this method, a descriptive (important) wordlist will be used in the next process, while other words will be discarded.

2) Feature selection

Feature selection is the process of selecting the most relevant or essential subset of features from the text to improve the machine learning model's performance. This helps reduce data dimensions and overfitting, improve model accuracy, and speed up training. The following are the stages in the feature selection process:

- Stemming is the stage of changing the form of words into a base word or finding the root word of each word. Reduce the size of the vocabulary by using a heuristic algorithm to remove morphological suffixes from words and leave only the base words that are not necessarily the primary form of the word [18].
- Lemmatization is the stage of reducing the size of vocabulary by using morphological information to remove inflection suffixes from words to obtain the primary form of the word or lemma [19].
- Tokenizing or lexical analysis, is the stage of cutting the input string into smaller pieces called tokens. In this process, numbers, punctuation marks, and other characters are also removed, which are considered to not influence text processing [13].

### *E. Clustering DBSCAN with Autoencorder Neural Network*

The Density-based Spatial Clustering of Application with Noise (DBSCAN) algorithm is a density-based clustering method of data observation positions with the principle of grouping relatively close data [20]. DBSCAN is often applied to data that contains much noise because DBSCAN will not enter data that is considered noisy into any cluster. DBSCAN requires two input parameters before the clustering process: epsilon (eps) and minimum points (minPts). Epsilon is the maximum distance between two data points in one cluster allowed, and minimum points are the minimum amount of data in the epsilon distance for a cluster to form. The distance method used in DBSCAN is the Euclidian distance. In addition to epsilon and minPts, there are several other terminologies in the DBSCAN method: direct density reachable, core point, border point, and noise point.

Autoencoder is a particular class of algorithms that can learn the efficient representation of input data without labels. It is a neural network designed for unsupervised learning. An essential principle of an automated decoder is effectively learning to compress and represent input data without unique labels. This is achieved using a two-fold structure consisting of an encoder and a decoder. Encoders aim to encode or compress the input data into more miniature-size representations and at the same time store as much important information as possible. From that representation, the decoder reconstructs the initial input. For a network to derive meaningful data patterns, the process of coding and decoding facilitates the definition of critical features.

There are two main steps in deep autoencorder based clustering: training and testing clustering. At the training step, an in-depth autoencoder with an encoder and decoder is trained using a training set. The flattened input vector is inserted into the encoder in a multilayer with representative low-dimensional learning. This learned representation is further fed into a set-top box that tries to recover an output the same size as the input. This autoencoder training process tries to reconstruct as much input as possible. In the following grouping step, we apply the autoencoder to the test set. The output from the encoder will be fed to the classic DBSCAN algorithm for clustering.

### *F. Topic Modelling*

Topic modeling is analogous to factor analysis, it is a probabilistic model that assesses the frequency of a term, assigns n-gram loading to a topic, and assigns a document loading to the extent that the topic contains each topic [21]. Topic modeling is one of the approaches to text mining that is quite reliable in discovering hidden text data and finding relationships between texts from one text to another from a corpus [22]. Topic modeling includes unsupervised learning because the data used does not have labels. The data is grouped by topic by paying attention to its similarity. Topic modeling involves counting words and grouping similar word patterns to infer topics from unstructured data. By

detecting patterns such as word frequency and spacing between words, the topic model groups similar feedback and the most frequently occurring words and expressions. This information makes it possible to quickly deduce what each series of texts is talking about.

Latent Dirichlet Allocation (LDA) is one of the topic modeling methods used to find hidden topics in a collection of text documents. LDA is a generative model that assumes that a document is a mixture of several topics and that each is a mixture of words. LDA is a powerful tool for finding topic structure in a collection of documents. By understanding the distribution of topics in a document and the words in a topic, we can gain deeper insights into the content of the analyzed text. LDA is often used in various applications such as text analysis, information search, and document grouping. In the data input process, LDA accepts text documents as input. Each document is considered a mixture of several topics, each containing a word distribution. After that, LDA starts by initializing a random topic for each word in the document. LDA uses the Gibbs Sampling algorithm to update the topic assignment for each word. Gibbs Sampling is an iterative Monte Carlo Markov Chain (MCMC) method. This iterative process continues until the model reaches convergence when the change in the distribution of topics and words becomes very small from one iteration to the next. After convergence, LDA generates a word probability distribution for each topic.

## III. METHODOLOGY

### *A. Data and Data Sources*

The data used in this study is secondary data from January 1, 2024, to May 24, 2024, obtained from the Suara Surabaya radio station. The observation unit used was 1709 reports related to traffic in Surabaya. The data used in this study consists of 14 attributes with the specifications of the dataset used in this study as follows.

**Table 1** Research Data Attributes

| Attribute | Information |
|---|---|
| Date | Detailed descriptions of dates consisting of dates, months, and years |
| Date-only | Numbers showing the date in the month |
| Months-only | Numbers that indicate the month of the year |
| Years-only | Numbers that indicate a specific year |
| Days-only | Names that indicate the day of the week |
| Hours-detail | Detailed descriptions of time durations consisting of hours, minutes, and seconds |
| Hours-only | Units of time to determine the time of day |
| Minutes-only | Units of time to set the duration or interval of time |
| Seconds-only | Units of time to determine time that requires high precision |
| Criterion 1 | Types of reports submitted by the complainant |
| Criterion 2 | Specification of the type of report submitted by the complainant |
| Call Status | Status of incoming and outgoing calls (call in and call out) |
| Onair Status | Broadcast live information delivery |
| Report | Information submitted by the complainant |

### *B. Data Analysis Steps*

The data analysis in this study was carried out using the Neural Network and Clustering autoencoder method using the DBSCAN algorithm. The reduction of the dataset dimensions uses an autoencoder trained to copy the input and output so that the dataset is lighter in application to the model. DBSCAN is a clustering algorithm in data mining and data analysis used to group objects based on spatial proximity and density of surrounding objects. The primary purpose of DBSCAN is to identify clusters in irregular data or data that do not have a precise geometric shape. The data analysis steps carried out are as follows:

1. Data collection by licensing the Suara Surabaya radio station to collect data on crime reports in Surabaya.
2. Preprocessing data. The data taken is processed through several stages: data cleaning and normalization, abbreviations, irrelevant characters, and stopwords removal or filtering.
3. Feature selection. This stage reduces the dimensions of text data so that the results are better quality. The process carried out is stemming, spell correction, and tokenization.
4. Feature extraction. Word Embedding with FastText algorithm and neural network autoencoder is used at this stage.
5. Determining the best number of clusters is carried out based on the results of the silhouette coefficient calculation.
6. Data clustering process with k-means clustering method.
7. Analyze the types of crime reports in each cluster

The above analysis steps can be visualized through the flowchart as follows.

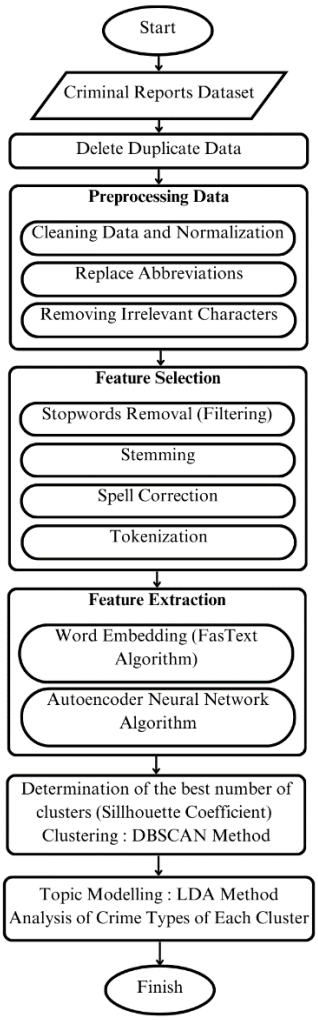


**Figure 1** Flowchart of Research Steps

## IV. RESULTS AND DISCUSSIONS

### *A. Preprocessing Data*

This process is a data processing stage that changes the original and unstructured format to a structured format to be processed for the next stage. The stages carried out are as follows.

1. Vehicle Plate Normalization

   Normalization of reported vehicle plates to identify vehicles that are the object of crime cases and vehicle administration areas. In this study, the normalization of motor vehicle plates is only focused on the letters before the police number that identifies the administrative area of the vehicle concerned. The results of normalization on motor vehicle plates are presented in Table 2.

**Table 2** Vehicle Plate Normalization

| The Report | Plate Number | Administrative Region |
|---|---|---|
| 1st | None | Unknown |
| 2nd | L 6301 PD | Surabaya City |
| 3rd | W 4348 NCL | Sidoarjo Regency |
| 4th | L 4785 AAN | Surabaya City |

2. Case Folding

   Case folding is the stage of changing all uppercase or capital letters in the report data to lowercase using the function 'df["Laporan"] = df["Laporan"].str.lower()'.

3. Replace Abbreviations

   Replace abbreviations will replace words containing abbreviations in their original form so they can be appropriately identified and reduce the data dimensions so they are not too large. The author also compiles the collection of words containing abbreviations and their original forms according to the needs of the research data.

4. Normalization of Occurrence Time

   Normalizing the time of reported events helps identify the period of crime cases that often occur. This study's normalization of the time of events is more focused on the period of events, such as dawn, morning, afternoon, evening, and night. The normalization results at the time of the event are presented in Table 3.

**Table 3** Event Time Normalization

| The Report | Report | Time Normalization |
|---|---|---|
| 2nd | reg//<br /> kehilangan motor sekitar jam 1 pagi tadi / hilang... | 01:00 at dawn |
| 4th | kehilangan sepeda motor// hilang baru saja/ sekitar jam 16.00 sore// hilang di daerah pandugo... | 16:00 in the afternoon |
| 94th | ... di parkiran salon haini// jam 22.15 ini tadi... | 22:15 at night |
| 96th | …hilang 21 desember, jam 9 malam// hilang di parkiran… | 09:00 at night |

5. Remove Punctuation dan single letter
   Remove punctuation, which removes punctuation that exists on the report data. Since this study only classifies text data, it will be removed from the report data in addition to alphabet characters. In addition, single letter characters or those that only consist of one letter are also removed.
6. Remove Digit
   All numbers in the text document will be removed by using the regex function 're.sub(r'\d+', '', text)'.

The stages of data preprocessing from stage (1) to stage (6) can be seen in Table 4.

**Table 4** Results of Data Preprocessing Process

| The Report | Before Data Preprocessing | After Data Preprocessing |
|---|---|---|
| 1st | REG//<br />SAYA HABIS KECOPETAN DI IR SUKARNO PAS SETELAH GALAXY MALL// SAYA NAIK MOTOR SUPRA BERBONCENGAN DENGAN ISTRI DAN TAS ISTRI SAYA AMBIL/ PELAKU NAIK MOTOR SPORT TIDAK ADA NOPOL… | saya habis kecopetan di insinyur sukarno pas setelah galaxy mall saya naik motor supra berboncengan dengan istri dan tas istri saya ambil pelaku naik motor sport tidak ada nomor polisi… |
| 2nd | REG//<br />KEHILANGAN MOTOR SEKITAR JAM 1 PAGI TADI / HILANG DI JL PETEMON BARAT 28/ MOTOR HONDA SUPRA X HITAM/ 2014/ L 6301 PD/ SELEBOR BELAKANG PECAH/ SPION STANDART/ DI DPN TOTOK ADA STIKER PASAR SEMEMI KUNING// MAU LAPOR KE POLSEK BENOWOEK… | kehilangan motor sekitar subuh tadi hilang di jalan petemon barat motor honda supra hitam dengan asal dari surabaya selebor belakang pecah spion standart di depan totok ada stiker pasar sememi kuning mau lapor ke polsek benowoek… |
| 3rd | KEHILANGAN MOTOR//<br /> HONDA PCX TH 2022/ WARNA MERAH/ NOPOL W 4348 NCL// CIRI2/ MASIH STANDAR//<br /> HILANG DI DLM RUMAH// RMH DALAM KONDISI KOSONG// KAYAKNYANYA MALINGNYA TAHU TEMPAT P[ENYIMPANAN KUNCI KALAU PAS BEPERGIAN/… | kehilangan motor honda pcx tahun warna merah nomor polisi dengan asal dari sidoarjo ciri masih standar hilang di dalam rumah rumah dalam kondisi kosong kayaknyanya malingnya tahu tempat penyimpanan kunci kalau pas bepergian… |

### *B. Feature Selection*

The feature selection process is fundamental in text mining analysis because this stage is carried out to reduce the dimensions of textual data by removing irrelevant words so that the grouping process is more effective and accurate. The feature selection process in this study begins with removing stopwords based on the Indonesian corpus. Remove stopwords are used to remove words in a corpus that appear and are considered not to figure out the content of a sentence. Meaningful word selection by eliminating less essential words in building the model can improve the accuracy of the classification system. Next, to reduce the data dimension so that it is not too large, a stemming process is carried out with the corpus provided in Sastrawi with StemmerFactory(). This process is essential in reducing the number of different word indexes of a document by eliminating existing word suffixes. After the stemming process, the spellchecker process is continued to correct each word from the stemming process according to the Indonesian corpus that the researcher has compiled. The corpus in the spellchecker process is adjusted to each word variable in the research dataset. Researchers added much vocabulary related to the names of streets and areas in Surabaya, types and brands of motor vehicles, slang words often used, and English absorption words. The last stage in feature selection carried out in this study is the tokenization process. The tokenizing process is carried out to cut the text of each word based on spaces. The results of the feature selection process, from the removal of stopwords to spell correction stages, are presented in Table 5.

**Table 5** Feature Selection Process

| The Report | Stopword Results | Stemming Results | Spell Correction Results |
|---|---|---|---|
| 1st | habis kecopetan insinyur sukarno pas galaxy mall motor supra berboncengan istri tas istri ambil pelaku motor sport nomor polisi kejar unair kampus… | habis copet insinyur sukarno pas galaxy mall motor supra bonceng istri tas istri ambil laku motor sport nomor polisi kejar unair kampus… | habis copet insinyur sukarno pas galaxy mali motor supra bonceng istri tas istri ambil laku motor sport nomor polisi kejar unair kampus… |
| 2nd | kehilangan motor subuh hilang jalan petemon barat motor honda supra hitam surabaya selebor pecah spion standart totok stiker pasar sememi kuning lapor polsek benowoek… | hilang motor subuh hilang jalan petemon barat motor honda supra hitam surabaya selebor pecah spion standart totok stiker pasar | hilang motor subuh hilang jalan petemon barat motor honda supra hitam surabaya selekor pecah spion standar totok stiker pasar |

| The Report | Stopword Results | Stemming Results | Spell Correction Results |
|---|---|---|---|
| | | sememi kuning lapor polsek benowoek… | sememi kuning lapor polsek benowo... |
| 3rd | kehilangan motor honda pcx warna merah nomor polisi sidoarjo ciri standar hilang rumah rumah kondisi kosong kayaknyanya malingnya penyimpanan kunci pas bepergian… | hilang motor honda pcx warna merah nomor polisi sidoarjo ciri standar hilang rumah rumah kondisi kosong kayaknyanya maling simpan kunci pas pergi… | hilang motor honda pcx warna merah nomor polisi sidoarjo ciri standar hilang rumah rumah kondisi kosong kayaknyanya maling simpan kunci pas pergi… |

### C. Feature Extraction

At this stage, each word will be represented as a numerical vector representing a word's popularity or a nearing word. The word weighting used in this study is Word Embedding with the FastText algorithm. FastText is a method that is an extension of Word2Vec that not only represents words in vectors that can carry the semantic meaning of words but also takes into account the intensity or frequency with which the word appears. FastText embeddings also have advantages that other word embeddings do not have, namely detecting words that are not in the previous dictionary or commonly called Out of Vocabulary (OOV). Figure 2 presents the results of word embedding vectorization with the FastText algorithm.

```
[[ 0.12427455  0.14955792  0.13155879 ...  0.21026844 -0.01394737
  -0.1993193 ]
 [ 0.122805    0.15755904  0.13867915 ...  0.21800579 -0.01706691
  -0.20458488]
 [ 0.11832102  0.15974644  0.13023043 ...  0.21309866 -0.01384855
  -0.18272534]
 ...
 [ 0.12612858  0.16239387  0.12648982 ...  0.21395713 -0.01942041
  -0.18323877]
 [ 0.12354628  0.16500916  0.15367524 ...  0.22339255 -0.02208017
  -0.20393634]
 [ 0.12096962  0.16296433  0.12947948 ...  0.19357947 -0.002875
  -0.20179372]]
```

**Figure 2** Data Vector with FastText

The next feature extraction stage uses the neural network's autoencoder algorithm to identify hidden patterns in the report data represented as vectors using the previous FastText model. In the input layer, dimensions correspond to the length of the text vector (300 dimensions) according to the FastText model used. The activation functions used are Relu on the hidden layer and Sigmoid on the output layer to reconstruct the original input data. The autoencoder algorithm uses Mean Squared Error (MSE) loss and Adam optimizer with a learning rate 0.001. The training is performed for 50 epochs to ensure the model can extract good features from the report data. The parameters used are an input dimension of 300 (according to the text vector dimension of the FastText model) and an encoding dimension of 64 (a hidden representation dimension selected to reduce the dimension but retain relevant information).

### D. Text Clustering

The best number of clusters in this study was determined using the DBSCAN method based on the silhouette coefficient values evaluation criteria. The optimal parameters based on the silhouette coefficient value criteria are through a hyperparameter tuning process with a combination of epsilon values between 0.1 to 1.1 with multiples of 0.1 and minimum samples between 2 to 20 with multiples of 2. The best combination of parameters obtained is the criteria for epsilon value one and minimum samples 4 with a silhouette coefficient value of 0.4752 so that the number of clusters formed is 3 clusters. Figure 3 presents a visualization of the clustering results with DBSCAN clustering with PCA.

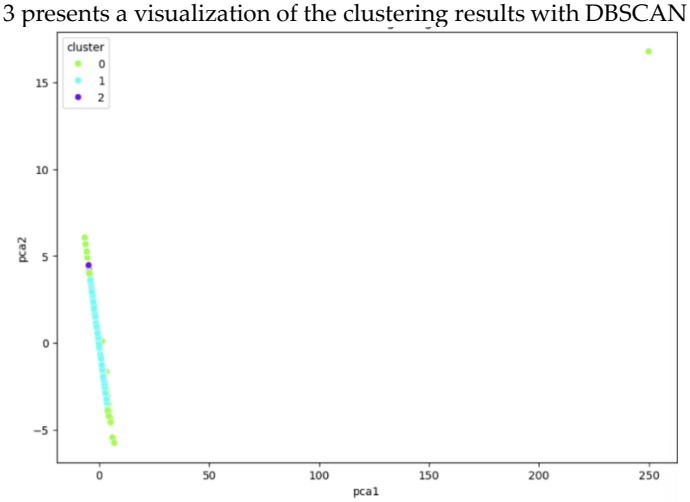


**Figure 3** Clustering Results with DBSCAN

The specific number of reports distributed in each *cluster* is presented in Figure 4 as follows.

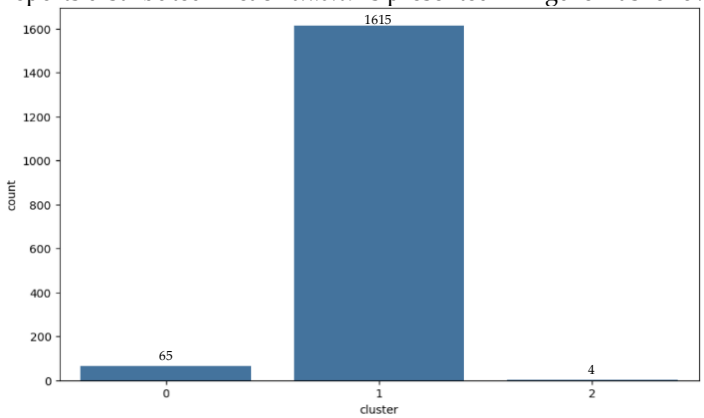


**Figure 4** Distribution of Clustering Results

Figure 3 shows that the cluster with the highest number of reports is cluster 1 with 1615 reports, then cluster 0 with 65 reports, and cluster 2 with 4 reports. Determining the type of content of each report cluster can be done by analyzing the words that appear most frequently in each cluster. These words will be taken from the top 10 words that appear most frequently to figure out the type of crime report in each formed cluster, as presented in Table 6.

**Table 6** Text Clustering Results

| Cluster | Most Frequently Occurring Words | Report Contents |
|---|---|---|
| 0 | motor, hilang, anak, rumah, ga, masuk, warna, kemarin, sepeda, saudara | Reports of a broader range of crime incidents, such as losing motorcycles and entering strangers into homes. The inclusion of the words "anak" and "saudara" indicates a report involving family members or children. |
| 1 | hilang, motor, warna, nomor, surabaya, hitam, polisi, honda, lapor, mobil | The report of a missing motorcycle in Surabaya City includes details of the black color, Honda brand, police number, and interaction with the police. |
| 2 | motor, hilang, polsek, lapor, vario, warna, beat, september, gudang, sidoarjo | Reports of missing motorcycles involving interactions with local police or police stations led to crimes involving details of the Vario and Beat motorcycle brands and Sidorajo locations in September. |

Next, Figure 5-7 is presented which shows a visualization of the types of community reports related to crime in cluster 0, cluster 1, and cluster 2.


**Figure 5** Word Cloud Cluster 0


**Figure 6** Word Cloud Cluster 1


**Figure 7** Word Cloud Cluster 2

### *E. Topic Modelling Analysis on Data Clustering*

Through topic modeling analysis with Latent Dirichlet Allocation (LDA), it can be seen that dominant words provide additional information to understand the various aspects analyzed. In topic modeling using LDA, each word is assigned a weight value that reflects its contribution to a particular topic. This process involves statistical analysis of a collection of documents or texts to determine the pattern of occurrence of certain words in a specific context. This weight indicates how significant a word is in representing the essence of a topic. The weight of each word can be seen in Table 7 below.

**Table 7** Results of Weight Values per Word with LDA

| Cluster | Topic | Word Weight |
|---|---|---|
| 0 | 1 | 0.026*"motor" + 0.018*"ga" + 0.016*"anak" + 0.016*"hilang" + 0.012*"pagi" + 0.010*"sepeda" + 0.010*"warna" + 0.010*"rumah" + 0.008*"cctv" + 0.008*"dupak" |
| | 2 | 0.015*"rumah" + 0.011*"share" + 0.010*"cek" + 0.010*"ngecek" + 0.008*"orang" + 0.008*"polisi" + 0.008*"surabaya" + 0.008*"bpjs" + 0.008*"liss" + 0.008*"pake" |
| | 3 | 0.032*"motor" + 0.020*"masuk" + 0.015*"hilang" + 0.013*"saudara" + 0.013*"anak" + 0.011*"rumah" + 0.010*"laku" + 0.010*"siang" + 0.010*"beat" + 0.008*"kemarin" |
| 1 | 1 | 0.024*"motor" + 0.019*"warna" + 0.015*"mobil" + 0.014*"polisi" + 0.014*"orang" + 0.012*"hilang"+0.011*"nomor" + 0.010*"bawa" + 0.010*"hitam" + 0.009*"anak" |
| | 2 | 0.034*"mobil" + 0.012*"orang" + 0.011*"teman" + 0.011*"tipu" + 0.009*"beli" + 0.009*"bawa" + 0.008*"surabaya" + 0.007*"pinjam" + 0.007*"laku" + 0.006*"juta" |
| | 3 | 0.074*"hilang" + 0.064*"motor" + 0.033*"warna" + 0.026*"surabaya" + 0.026*"nomor" + 0.024*"honda" + 0.023*"hitam" + 0.021*"polisi" + 0.018*"parkir" + 0.017*"beat" |
| 2 | 1 | Topik 1: 0.024*"motor" + 0.024*"hilang" + 0.024*"polsek" + 0.024*"vario" + 0.024*"lapor" + 0.024*"scoopy" + 0.024*"september" + 0.024*"beat" + 0.024*"surabaya" + 0.024*"warna" |
| | 2 | Topik 2: 0.087*"polsek" + 0.067*"hilang" + 0.067*"vario" + 0.067*"motor" + 0.047*"warna" + 0.047*"beat" + 0.027*"wonocoload" + 0.027*"kos" + 0.027*"hitam" + 0.027*"pagi" |
| | 3 | Topik 3: 0.108*"motor" + 0.059*"lapor" + 0.058*"hilang" + 0.033*"krianek" + 0.033*"teman" + 0.033*"coklat" + 0.033*"jg" + 0.033*"nfl" + 0.033*"sekolah" + 0.033*"mi" |

After completing the topic modeling process using Latent Dirichlet Allocation (LDA) in community complaint reports related to crime, the topic results and explanations from each cluster can be seen in Table 8.

**Table 8** Results of LDA Topic Analysis

| Cluster | Topic | Topic Analysis Results |
|---|---|---|
| 0 | 1 | This topic includes reports of motorcycle losses, incidents around the house, and possible references to CCTV surveillance. |
| | 2 | This topic may include a variety of crime reports, including incidents around the home, interactions with authorities such as the police, and possibly references to specific services or activities, such as the Social Security Administration Agency (BPJS). |
| | 3 | This topic includes reports on people entering homes, losing motorcycles, and cases surrounding lost items with details of the time of events such as "siang" or "kemarin". |
| 1 | 1 | This topic focuses on reports of missing motorcycles and cars accompanied by attempts to report them to the police. The report submitted includes detailed information about the vehicle's color and the police number. |
| | 2 | This topic may include reports of car-related fraud or transactions involving vehicles with the mode of loan or car purchase transactions. |
| | 3 | This topic focuses on reports about the rampant cases of missing Honda brand motorcycles in Surabaya, which often occur at parking locations. |
| 2 | 1 | This topic focuses on reports about the loss of motorcycles, especially Vario and Beat, with the interaction of reporting to the Sector Police in Surabaya. |
| | 2 | This topic includes reports that focus on missing motorcycles with details of location (e.g., Wonocolo), place of residence (e.g., boarding house), and time (e.g., morning). |
| | 3 | This topic includes reports about motorcycle losses with specific details such as the location of the incident (e.g., Krian), interactions with friends, and also the scene of the incident (e.g., school). |

In analyzing public complaint reports regarding crime in Surabaya, cluster 0 looks more diverse, including incidents around the house, loss of goods, and interactions with certain services. Cluster 1 focuses on reports of vehicle losses with details such as interactions with police and related transactions. Cluster 2 has a similar focus to cluster 1, but with the addition of a focus on interactions with police stations and more specific crime scenes. By paying attention to these aspects in topic modeling through community complaint reports, researchers can better understand the various dimensions that affect public opinion and views on crime and provide valuable information for efforts to counter and prevent crime.

## V. CONCLUSIONS AND SUGGESTIONS

This study's text mining approach requires a reasonably complex data normalization and preprocessing process, depending on the data structure level. Data vectorization with FastText is also crucial in determining accurate and informative vector representations to improve the quality of clustering results. Next, an autoencoder neural network algorithm is used to effectively reduce the data dimension with an input dimension of 300 and an encode dimension of

64. Clustering analysis using the DBSCAN method based on the criteria of silhouette coefficient value obtained three clusters formed with the dominant report in cluster 1. The clustering results show essential patterns in complaint reports, and LDA analysis reveals critical topics in the report. Cluster 0 shows a diversity of reports focusing on motor loss, interaction with homes or properties, and people's entry into homes. Cluster 1 is more focused on the loss of vehicles, both cars and motorcycles, with specific details such as vehicle color, number, brand, and related transactions or social interactions. Meanwhile, cluster 2 focuses on reports related to interactions with police stations and other specific details, such as information on the location of the incident.

This text mining approach to community crime report data improves the accuracy and efficiency of analysis and provides essential information supporting efforts to handle and prevent crime. The recommendation that should be given is that the data used in the analysis only focus on one agency, namely Suara Surabaya. Thus, it can be used as a further evaluation to expand the scope of the data to get more accurate results. In increasing awareness of the importance of maintaining the safety of the surrounding environment, it is hoped that the community, government, and related agencies can collaborate to support a safe and comfortable Surabaya City.

## ACKNOWLEDGEMENT

The author would like to thank the Directorate of Learning and Student Affairs (Belmawa), the Directorate General of Higher Education, and the Ministry of Education, Culture, Research, and Technology for providing support in this research through the Student Creativity Program (PKM), especially in the field of Social Humanities Research (RSH). In addition, the author expressed his gratitude to Suara Surabaya for providing the research data. The award was also given to the Statistics Study Program, Faculty of Science and Technology, Universitas Airlangga, and all parties supporting this research and publication.

## REFERENCES

[1] M. R. P. Musa, A. B. Lesmana, R. N. Arthamevia, P. A. Pratama and N. Savitri, "Human Rights and Pancasila: A Case of Tionghoa Ethnic Discrimination in Indonesia," *Indonesian Journal of Pancasila and Global Constitutionalism,* vol. 1, no. 1, pp. 119-170, 2022.

[2] E. Kahya-Özyirmidokuz, "Analyzing unstructured Facebook social network data through web text mining: A study of online shopping firms in Turkey," *Information Development,* vol. 32, no. 1, pp. 70-80, 2016.

[3] H. Choi, M. Kim, G. Lee and W. Kim, "Unsupervised learning approach for network intrusion detection system using autoencoders," *The Journal of Supercomputing,* vol. 75, no. 9, pp. 5597-5621, 2019.

[4] S. Andleeb, R. Ahmed, Z. Ahmed and M. Kanwal, "Identification and classification of cybercrimes using text mining technique," *In 2019 International Conference on Frontiers of Information Technology (FIT),* pp. 227-232, 2019.

[5] I. Insiyah, M. Khasanah and T. P. Hendarsyah, "Penerapan Metode Ward Clustering Untuk Pengelompokkan Daerah Rawan Kriminalitas Di Jawa Timur Tahun 2021," *Jurnal Statistika dan Komputasi,* vol. 2, no. 1, pp. 44-54, 2023.

[6] H. D. Tampubolon, S. Suhada, M. Safii, S. Solikhun and D. Suhendro, "Penerapan Algoritma K-Means dan K-Medoids Clustering untuk Mengelompokkan Tindak Kriminalitas Berdasarkan Provinsi," *Jurnal Ilmu Komputer dan Teknologi,* vol. 2, no. 2, pp. 6-12, 2021.

[7] R. N. Fahmi, M. Jajuli and N. Sulistiyowati, "Analisis Pemetaan Tingkat Kriminalitas di Kabupaten Karawang Menggunakan Algoritma K-Means," *INTECOMS: Journal of Information Technology and Computer Science,* vol. 4, no. 1, pp. 67-79, 2021.

[8] A. D. Putra, G. S. Martha, M. Fikram and R. J. Yuhan, "Faktor-Faktor yang Memengaruhi Tingkat Kriminalitas di Indonesia Tahun 2018," *Indonesian Journal of Applied Statistics,* vol. 3, no. 2, pp. 123-131, 2021.

[9] A. O. Edwart and Z. Azhar, "Pengaruh Tingkat Pendidikan, Kepadatan Penduduk dan Ketimpangan Pendapatan Terhadap Kriminalitas di Indonesia.," *Jurnal Kajian Ekonomi Dan Pembangunan,* vol. 1, no. 3, pp. 759-768, 2019.

[10] R. Soesilo, KUHP Kitab Undang Undang Hukum Pidana Lengkap serta Komentarnya, Bogor: Politea, 1976.

[11] M. D. Wuryandari and I. Afrianto, "Perbandingan Metode Jaringan Syaraf Tiruan Backpropagation Dan Learning Vector Quantization Pada Pengenalan Wajah," *Jurnal Komputer dan Informatika (Komputa),* vol. 1, no. 1, pp. 45-51, 2012.

[12] S. Haykin, Neural Networks: A Comprehensive Foundation, New York: Macmillan, 1994.

[13] R. Feldman and J. Sanger, The text mining handbook: advanced approaches in analyzing unstructured data, Cambridge: Cambridge University Press, 2006.

[14] V. Gupta and G. S. Lehal, "A Survey of Text Mining Techniques and Applications," *Journal of Emerging Technologies in Web Intelligence,* vol. 1, p. 60–76, 2009.

[15] F. Nurhuda, S. W. Sihwi and A. Doewes, "Analisis sentimen masyarakat terhadap calon Presiden Indonesia 2014 berdasarkan opini dari Twitter menggunakan metode Naive Bayes Classifier," *ITSmart: Jurnal Teknologi dan Informasi,* vol. 2, no. 2, pp. 35-42, 2016.

[16] A. T. J. Harjanta, "Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining," *Jurnal Informatika UPGRIS,* vol. 1, 2015.

[17] C. Triawati, M. A. Bijaksana, N. Indrawati and W. A. Saputro, "Pemodelan Berbasis Konsep untuk Kategorisasi Artikel Berita Berbahasa Indonesia," *In Seminar Nasional Aplikasi Teknologi Informasi (SNATI),* 2009.

[18] S. Bird, E. Klein and E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, O'Reilly Media, Inc, 2009.

[19] C. D. Manning, P. Raghavan and H. Schütze, Introduction to Information Retrieval, Cambridge: Cambridge University Press, 2008.

[20] H. Jiawei, K. Micheline and P. Jian, "Data Mining: Concepts and Techniques The Morgan Kaufmann Series in Data Management Systems," *Elsevier*, 2011.

[21] T. Schmiedel, O. Müller and J. Vom Brocke, "Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture," *Organizational Research Methods*, vol. 22, no. 4, pp. 941-968, 2019.

[22] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia tools and applications*, vol. 78, pp. 15169-15211, 2019.