Incident Clustering in the Warehouse Workspaces by Using Text Mining

F A Syaani¹, Irhamah²*, A Mukarromah³, K Fithriasari⁴

¹,²,³ Department of Statistics, Institut Teknologi Sepuluh Nopember Surabaya
Jl. Raya ITS, Kampus ITS Sukolilo, Surabaya, Indonesia 60111

Corresponding Author Email: irhamah@statistika.its.ac.id²*

Abstract. Incidents can occur anytime and anywhere whether intentional or unintentional, especially in the industrial area. To prevent the incident, an analysis is needed to find out the patterns that formed based on the information from incident reports. The goals of company are to be able to identify and take action to deal quickly if some incident happens. Text Mining is done by applying the bigram technique to form the incident patterns. The analytical method used is K-Means Clustering and Hierarchical Clustering. In addition, a feature selection is also used by applying the Genetic Algorithm method to obtain optimal features. The results obtained state that the feature selection process is very influential on the formation of incident clusters. When compared, the K-Means Clustering and Hierarchical Clustering methods have different effects on each warehouse sector. The best results in the Sector of Life Style and Sector of Technology are formed by using the K-Means Clustering method while in the Sector of Consumer, Retail and SPL the best cluster results are obtained based on the Hierarchical Clustering method.

1. Introduction
Incidents can happen anytime and anywhere. As it is true with traffic incidents, industrial incidents are also relatively easy to happen. Although some incidents may arise from unexpected things, however, most incidents have a similar pattern to incidents that were happened before. Therefore, studying past incidents is very important to determine any preventive measures to be taken in the future. Based on this principle, many people, particularly those engaged in industrial sector, have started to collect data on incidents reports that have occurred in the past [1]. Individuals or companies that have studied the causes of previous occurrence of incidents are generally able to identify and take immediate measure to deal with any occurrence of incidents at any time of such occurrence [2]. In the process of identifying an accident in the past, any related information is needed. Such information can be in the form of legitimate company reports pertaining accident investigations in the work environment [3]. Usually, such reports are provided in the form of text data which are relatively large in size so that if done manually it will take a long time. One of the techniques that can be used to collect information based on text data is Text Mining [4]. Text Mining facilitates the extraction of information hidden in a set of data quickly and accurately [5]. In the process of information extraction, several techniques are needed to obtain interesting patterns by applying the Bigram technique to perform word combination.
Bigram can provide information more effectively because it is formed from words that are related to one another [6].

The analytical method applied in this study is cluster analysis by comparing between Ward's Clustering with K-Means Clustering. K-Means method can work quickly on relatively large datasets based on the Euclidean distance [7]. Meanwhile, Ward's Clustering method can effectively minimize the variance between group members based on the sum of square values [8]. In determining the optimal number of clusters, the value of the Silhouette Coefficient is used for each clustering process. Both methods then shall be subject to an evaluation of clustering by comparing the value of icd-rates in each cluster formed. In improving the accuracy of the cluster formed thereby, a feature selection method is also carried out using a Genetic Algorithm. This algorithm is used to find features that can be used in determining cluster type labels. Such process is performed by minimizing any features that are considered as uninformative features in order to obtain the main features in describing a class. Genetic Algorithm method is able to improve the quality of the results of a text summarization where the objective function (fitness) in the Genetic Algorithm is able to find words with more optimal weight when compared to ordinary method [9].

The final results of the research describe incident clusters based on their respective characteristics so that it can find out what problems are often reported based on each the acquired mapping. It is hoped that this research can provide information that can be applied in minimizing the problem of accidents in the workplace. Thus, the company can provide any appropriate preventive measures and response for the safety, health and security of their employees.

2. Literature Review

2.1. Text Mining

Text Mining is a process extracting information from data during the certain period and processed using the analytical method that correspond. Text Mining is used to extract the information from unstructured text data sources through identification and exploration of interesting patterns [10]. There are several stages before data processing in order to make the data can be processed more easily. This stage is better known as text pre-processing. Text pre-processing is an initial process before further analysis to change unstructured data to be more structured [11].

2.1.1. Data Cleaning. Data Cleaning is a process to clean the data without any noise anymore. In the text pre-processing, this process shall include changing of capital letters to lowercase (case folding), removing punctuation, removing numbers, removing irrelevant words (stop-words removal), removing double spaces (whitespace), translating of report sentences into a certain language desired, and breaking the sentences into words (tokenizing) [12].

2.1.2. Stemming. Stemming is a process of mapping and decomposition from various forms of words into their root words [12]. A set of rules in stemming in Indonesian shall be applied to remove affixes contained in a words, for example, "membetulkan", shall be changed into "betul" and "membaca" shall be changed into "baca". The Library which is used in the Python in the stemming process for Indonesian documents is Sastrawi (literal). Before carrying out a process of weighting onto each word in the document, it is necessary to calculate the frequency of each word therein or sentence and form a matrix.

2.1.3. n-Grams. n-Grams is a technique which is used to combine closely related n words together [13]. n-Grams with n = 2 is called bigram. Examples of the use of Bigram are such as "atap bocor", “lantai lubang”, “sampah berserakan”, etc. In general, n-Grams (n > 1) will provide information more effectively because they are made up of words that are closely related to one another.
2.1.4. Term Weightening. The term weightening is a method used to obtain the frequency value of each word in a document so that we can get a comparison of one certain word’s frequency to that of another due to difference of level of importance of the words respectively [14]. One method which is frequently used to calculate the weight of a term or word in a document is the Term Frequency-Inverse Document Frequency (TF-IDF). Term Frequency-Inverse Document Frequency (TF-IDF) is a measurement used to measure how important a word is to a document. Term Frequency (TF) is used to summarize the appearance of a word in a document. Meanwhile, the Inverse Document Frequency (IDF) is used to calculate the frequency of appearance of a word in a set of documents entirely [11]. There are 7 stages in weighting process with TF-IDF method that shall include Document Indexing, Term Frequency Weighting, Log Term Frequency Weighting, Document Frequency, Inverse Document Frequency, Term Frequency-Inverse Document Frequency (TF-IDF) and the final value of TF-IDF.

2.1.5. Feature Selection. Feature selection is one of the stages in data pre-processing which purpose to select influential features and remove features that do not have any role in data modelling. One of the methods that can be used is Genetic Algorithm. Genetic Algorithm is a meta-heuristic method in the field of Evolutionary Algorithm which is inspired by Darwin's theory of evolution [15]. The process of determining the best feature using the Genetic Algorithm method is as follows [16]:

- Population initialization, which is a stage of randomizing chromosome strings for a large number of predetermined populations;
- There are two main genetic operators used to maintain the diversity of the population in the Genetic Algorithm, include:
  - Crossover, the technique used in one-cut-point in which it is selected two parent chromosomes in a population and then the value of the contents of the chromosome subset is exchanged starting from a certain index point. This process is repeated until it is obtained the offspring number that meets the required amount from the crossover rate;
  - Mutation, which is a process carried out to prevent the algorithm from being trapped in the local optimum solution by adding a very small random value with a relatively low chance for hereditary variable to emerge.
- Selection, which is a process of evaluating fitness of reproduction results. This process purpose to determine the individual who will be selected. A fitness value is needed in the evaluation process to see the quality of the chromosome.

2.2. Text Clustering
Text clustering is an unsupervised process used to group several datasets in the form of text by applying different clustering algorithms.

2.2.1. Ward’s Clustering. Ward's Clustering is a clustering method that works based on the principle of the Agglomerative Hierarchical Clustering algorithm. Ward's method is the only method out of other several methods in Agglomerative Clustering that has a calculation of the number of clusters based on the sum of square criteria [8].

2.2.2. K-Means Clustering. K-Means clustering is a clustering technique that works based on Partitioned Clustering. The principle of Partitioned Clustering is to group documents randomly because they are influenced by centroids. Furthermore, in each iteration of the grouping based on Partitioned Clustering allows combined selection of documents to happen more than once [17]. The K-Means method has the ability to classify data with a fairly large amount. The time required to perform this clustering technique is also relatively fast and efficient [10].
2.3. Evaluation of Clusters
Evaluation of clustering results is carried out using two approaches, namely internal evaluation which is based on the Silhouette Coefficient value and external evaluation with the ic-drate value. The Silhouette Coefficient is used to measure the equations that occur in an object by comparing certain clusters with other clusters. Matters that need to be known when determining the silhouette value is the result of partitioning or clustering results and grouping all proximity between objects [11]. The formula used is as follows:

\[ s(g) = \frac{b(g) - a(g)}{\max\{a(g), b(g)\}}, \text{if} |C_g| > 1 \]  

(1)

With:
- \( a(g) \) = the smallest value of the average distance between documents \( g \) and all objects in one cluster;
- \( b(g) \) = the smallest value of the average distance between documents \( g \) and all objects in another cluster.
- \( C_g \) = cluster of one data object \( g \)

The best cluster performance results in the clustering method can use the icd-rate value. The icd-rate (internal cluster dispersion rate) is an indicator that can be used to measure the level of dispersion between clusters formed thereby [18]. The icd-rate value calculation can use the following formula:

\[ \text{icd-rate} = 1 - \frac{\text{SSB}}{\text{SST}} = 1 - R^2 \]  

(2)

with SSB is the total sum of square between clusters while SST is total sum of squares from partition [19].

2.4. Word Cloud
Word Cloud is a simple visualization technique that purpose to provide information on a word that often appears in a document. Technically, the word cloud describes the words that have the most frequent occurrences in a document. The word is displayed in size and color that will attract the reader's attention at the very first time they see the word cloud presented [20].

2.5. PT. X
PT. X is a multinational company engaged in global logistics. In 2018, PT. X already has 39 warehouses that are fully operational and actively running a goods logistics system. Like other industrial companies, PT. X has several reports that need to be analyzed, one of which is an incident report which is reported regularly in a monthly basis. The report is differentiated based on its level of treatment provided thereupon which is described in brief writing explaining the condition at the time. There are three categories of reports according to its level of treatment, namely low severity, mid severity and high severity. Each of these categories is subdivided into several types of accidents, for example unsafe act and unsafe condition in the low severity category. Unsafe act is a type of action which is dangerous for workers while unsafe act is a condition which is considered dangerous for workers [21].

3. Methodology

3.1. Data Sources
In this research, the data sources that will be used are primary data obtained from the monthly reports of PT. X regarding the conditions of damage that occurred in its work environment at 39 warehouses owned by PT. X. The description of incident is reported directly by employees in the company’s reporting system at each warehouse who has been granted the access such system.
3.2. Data Structure
The data taken came from the Hazard Report owned by PT. X from January 1st, 2018 to September 29th, 2018 with total of 6,598 data. The sector categories contained in the incident report are described in table 1.

| Sector Description | Number of Warehouse | Number of Report |
|--------------------|---------------------|------------------|
| Consumer           | 16                  | 2,591            |
| Life Style         | 3                   | 321              |
| Retail             | 6                   | 640              |
| Spare Part Logistic (SPL) | 8               | 2,740            |
| Technology         | 5                   | 281              |

3.3. Steps of Analysis
The analysis steps that will be carried out in this research are as follows:

- Collect incident description report data in each warehouse which is presented in the form of text data;
- Perform text pre-processing of incident report data, which includes data cleaning, stemming, bigrams, weighting (weighting), and feature selection;
- Obtain the optimum number of clusters based on the overall data as well as data from the feature selection using the Silhouette Coefficient value;
- Perform clustering analysis using Ward’s Clustering and K-Means Clustering method;
- Evaluating the best cluster results;
- Make conclusions on the results of research analysis.

4. Result and Discussion
4.1. Characteristics of Incident Report Data
The characteristics of the incident report data can be described using a descriptive statistical analysis. The analysis is carried out based on the number of reports submitted as well as the classification of the description of the incident that occurred.

Figure 1. Frequency of Incident Reports for January 1st, 2018 – September 29th, 2018
The number of incident reports recorded on January 1st, 2018 – September 29th, 2018 has a pattern that tend to decrease every month. It appears that most of reported incidents occurred in March by 1,050 incidents or about 16% of the total reports. Meanwhile, the lowest number of reports was in September with only 209 incidents reports. This condition can be expected as a result of any corrections and preventive measures that the company regularly takes to minimize the number of
incidents. In addition, there is a classification on the description data incident. The classification can be divided into two categories, namely unsafe act and unsafe condition.

![Figure 2. Proportion of Reported Incident Category](image)

When compared, this two proportions shown in the figure 2 have an unbalanced proportion. It is known that the category with the highest proportion based on the total reports received by the system is incidents with the category of unsafe condition, which is 75% or 4,958 reports, while for the category of unsafe act it is only around 25% of the total number of reports.

4.2. Data Pre-processing
Data pre-processing purpose to collect data which is ready to be analyzed. Data pre-processing includes data cleaning, stemming, and term-weighting. Data cleaning is a process of removing the unimportant characters so that the data is free of unnecessary symbols, numbers and words. The methods used to perform this stage includes changing capital letters to lowercase letters (case folding), removing symbols (punctuation), deleting numbers, deleting irrelevant words (stop words removal), and removing double spaces (whitespace). Then the words resulted from the data cleansing process shall be changed to their base word form. After being converted into their base word form, then these words are combined with each other using the $n$-Grams technique as many as $n = 2$ (bigrams). The next step is to calculate the frequency of appearance in the document. The frequency of each word is formed in a matrix called Document-Term Matrix. The set of frequencies obtained in the Document-Term Matrix is then weighted using the Term Frequency-Inverse Document Frequency (TF-IDF).

4.3. Data Exploration
The incident characteristics observed are the number of words formed into a feature based on the description of reported data in each warehouse sector. It is known that the number of features that are formed in each warehouse sector are as in Table 2.

| No | Warehouse Sector | Number of Features |
|----|------------------|--------------------|
| 1  | Consumer         | 523                |
| 2  | Life Style       | 923                |
| 3  | Retail           | 1,253              |
| 4  | SPL              | 1,483              |
| 5  | Technology       | 561                |

The words (features) with the highest occurrence frequency can have a significant effect on cluster formation. The frequency of these words was previously obtained based on calculations in the Document Term-Matrix. The appearance of the words is presented with the word cloud visualization as figure 3 to 7.
Based on figure 3 and figure 4, in the Consumer Sector, event that often results in incidents is pallet damage due to careless placement or overuse. Pallets and other facilities shall be regularly maintained and controlled in order to maintain their quality. In the Life Style Sector, there are many incidents that occur in the pedestrian area, such as employees who do not walk on the determined existing lanes, as well as loading and unloading activities that are not carried in appropriate places so that the activities disturb other road users who want to pass. Furthermore, in the warehouse entrance area, there were many officers who did not place the goods according to the determined standards and prevented other officers from entering the warehouse.

In the Retail Sector, there are many employees who do not comply with safety requirement standards, such as wearing shoes improperly by stepped on their folded end part of their shoes. On the other hand, there are also several damaged pallets that actually needs for replacement and or repaired/maintained. In the SPL sector, many incidents are caused by broken lights and scattered objects interfering the workers’ daily activities. The employees shall improve their awareness on the importance of keeping the warehouse clean and tidy. In the Technology Sector, a lot of damage has occurred in the warehouse or warehouse area, such as damaged floors and facilities therein and many workers who smoke while at works which is endangering the surrounding environment.

4.4. Data Clustering

Data clustering purpose to classify incidents that occurred in the warehouse environment as of January 1st, 2018 to September 29th, 2018 according to the similarities of the events. The clustered data is the incident description data in each warehouse sector. Clustering is carried out using two approaches, Hierarchical Clustering and Partitioned Clustering. In the Hierarchical Clustering approach, by using the Ward's Clustering method, while for the Partitioned Clustering approach, by using the K-Means Clustering method. The results of these two methods will obtain the optimal number of clusters in each method based on the Silhouette Coefficient value. In addition, optimal cluster results are also determined based on the comparison between the clusters with Genetic Algorithm as feature selection.
and without the same. The purpose of feature selection is to get features that match the conditions in the warehouse environment. The following is the number of features obtained using the Genetic Algorithm method.

| No | Warehouse Sector | Method      | Number of Features |
|----|-------------------|-------------|--------------------|
| 1  | Consumer          | K-Means     | 52                 |
|    |                   | Hierarchical| 21                 |
| 2  | Life Style        | K-Means     | 452                |
|    |                   | Hierarchical| 369                |
| 3  | Retail            | K-Means     | 557                |
|    |                   | Hierarchical| 464                |
| 4  | SPL               | K-Means     | 305                |
|    |                   | Hierarchical| 194                |
| 5  | Technology        | K-Means     | 235                |
|    |                   | Hierarchical| 159                |

### Table 3. Number of Featured in Each Sector after Feature Selection

$$
\begin{array}{|c|c|c|c|}
\hline
\text{No} & \text{Warehouse Sector} & \text{Method} & \text{Number of Features} \\
\hline
1 & Consumer & K-Means & 52 \\
& & Hierarchical & 21 \\
2 & Life Style & K-Means & 452 \\
& & Hierarchical & 369 \\
3 & Retail & K-Means & 557 \\
& & Hierarchical & 464 \\
4 & SPL & K-Means & 305 \\
& & Hierarchical & 194 \\
5 & Technology & K-Means & 235 \\
& & Hierarchical & 159 \\
\end{array}
$$

4.5. Evaluation of Cluster

Evaluation of cluster is carried out by comparing the Silhouette and icd-rate values obtained from each cluster method in each sector. The following is a comparison of the Silhouette value and icd-rate resulted for each sector after the implementation of cluster analysis.

| Sector     | Method      | Number of Cluster | Feature Selection | Silhouette Coefficient | icd-rate  |
|------------|-------------|-------------------|-------------------|------------------------|-----------|
| Consumer   | K-Means     | 17                | No                | 0.2                    | 0.885     |
|            | Hierarchical| 20                | No                | 0.198                  | 0.854     |
|            |             |                   | Yes               | 0.999                  | 0.0005    |
| Life Style | K-Means     | 20                | No                | 0.063                  | 0.854     |
|            | Hierarchical| 20                | No                | 0.248                  | 0.739     |
|            |             |                   | Yes               | 0.337                  | 0.898     |
| Retail     | K-Means     | 17                | No                | 0.090                  | 0.876     |
|            | Hierarchical| 20                | No                | 0.056                  | 0.924     |
|            |             |                   | Yes               | 0.761                  | 0.790     |
| SPL        | K-Means     | 18                | No                | 0.106                  | 0.911     |
|            | Hierarchical| 20                | No                | 0.054                  | 0.813     |
|            |             |                   | Yes               | 0.919                  | 0.600     |
| Technology | K-Means     | 14                | No                | 0.082                  | 0.880     |
|            | Hierarchical| 11                | No                | 0.365                  | 0.756     |
|            |             |                   | Yes               | 0.117                  | 0.866     |
|            |             |                   |                   | 0.690                  | 0.771     |

In table 4, it is known that the results of feature selection using the Genetic Algorithm method are able to provide the highest Silhouette value. In the Consumer, Fashion, Life Style, Retail, SPL, and
Technology sectors each has the greatest Silhouette value for K-Means and Hierarchical Clustering methods. This proves that feature selection using the Genetic Algorithm can significantly improve the cluster performance. Also, the presence of feature selection is able to provide a relatively low icd-rate value compared to that which uses overall feature data. In term of the goodness of the method, in the Consumer Sector it is known that the lowest icd-rate value is generated by cluster analysis using the Hierarchical Clustering with icd-rate value of 0.0005. It’s similar to the results for the Retail Sector and the SPL Sector. In both sectors, the best cluster results were obtained by using the Hierarchical Clustering method with icd-rate values of 0.790 and 0.600, respectively. Unlike the previous three sectors, the best cluster results in the Life Style Sector and Technology Sector were obtained using the K-Means Clustering method with icd-rate values of 0.739 and 0.756.

5. Conclusions and Suggestions

5.1. Conclusions

The characteristics of the incident data obtained indicate that the company has played an active role in minimizing incidents in its environment. The clusters resulted by using feature selection that have been selected by the Genetic Algorithm method tend to have better results, as evidenced by the value of the Silhouette Coefficient in each sector which tends to be higher than that resulted by using the overall feature data. Based on the goodness of the resulted clusters by using the icd-rate value, the best results in the Life Style and Technology Sector were formed with the K-Means Clustering method, while in the Consumer, Retail and SPL Sectors the best clusters resulted by using the Hierarchical Clustering method.

In the Consumer Sector, the cluster with the most members is related to the damaged warehouse facilities, the pallets, in particular. Similar to the Consumer Sector, in the Life Style Sector there is also a lot of damaged pallets and there are officers who are negligent during the course of their duties such as placing object disorderly. In the Retail Sector, incidents that often occur according to the number of clusters formed are many employees who do not use safety shoes properly while working and there are facilities that need to be repaired immediately because they can interfere with the company performance itself. In the SPL sector, the incidents that often occur based on the resulted clusters are that related to broken lights and other facilities requiring repair immediately. In addition, there are scattered objects around the warehouse area. In the Technology Sector, there are many incidents related to dangerous activities carried out by employees while working resulting in potential accidents.

5.2. Suggestions

Suggestions can be given regarding the analysis result is that we need to understand problem accurately and deeply during the course of data pre-processing because the results obtained from this process will greatly affect the results of the cluster analysis formed. While for PT. X, it is advised to consider the incident reporting process based on the cluster formed so that it can make it easier to handle and prevent any recurring occurrence.

Acknowledgement

This work was supported by Ministry of Research, Technology, and Higher Education, Republic of Indonesia and Institut Teknologi Sepuluh Nopember Surabaya Indonesia under “Penelitian Dasar”. The authors greatly appreciate the support.

References

[1] Veltman L M 2008 Incident Data Analysis using Data Mining Techniques (Texas: Texas A&M University)
[2] Nakata T 2017 Text-Mining on Incident Reports to Find Knowledge on Industrial Safety 2017 Annual Reliability and Maintainability Symposium (RAMS) (Orlando) pp 1-5
[3] Safety C f C P 2016 *Guidelines for Implementing Process Safety Management* (New York: WILEY)

[4] Allahyari M, Pouriyeh S, Assaei M, Safaei S, Trippe E D, Gutierrez J B and Kochut K 2017 *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques* (Canada: KDD Blgdas)

[5] Talib R, Hanif M K, Ayesha S and Fatima F 2016 *Text Mining: Techniques, Applications and Issues* International Journal of Advanced Computer Science and Applications (IJACSA)

[6] Jurafsky D and Martin J H 2018 *N-gram Language Models* In Speech and Language Processing 3rd Edition (USA: Stanford University)

[7] Zhang C and Fang Z 2013 *An Improved K-Means Clustering Algorithm* Journal of Information & Computational Science pp 193-199

[8] Murtagh F 2014 *Ward’s Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward’s Criterion?* Journal of Classification pp 274-295

[9] Meena K Y and Gopalani D 2015 *Evolutionary Algorithm for Extractive Automatic Text Summarization* Intelligent Computing, Communication & Convergence pp 244-249

[10] Feldman R and Sanger J 2007 *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data* (New York: Cambridge University Press)

[11] Jannah S Z, Fithriasari K and Usagawa T 2018 *Clustering and Visualizing Surabaya Citizen Aspirations by Using Text Mining* (Surabaya: Institut Teknologi Sepuluh Nopember)

[12] Tala F Z 2003 *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia* M.Sc. Thesis (Amsterdam: Universiteit van Amsterdam)

[13] Nguyen V H, Nguyen H T, Duong H N and Snasel V 2016 *n-Gram-Based Text Compression* Hindawi Publishing Corporation Computational Intelligence and Neuroscience

[14] Abualigah L M, Khader A T and Betar M A 2016 *Unsupervised Feature Selection Technique Based on Genetic Algorithm for Improving the Text Clustering* International Conference on Computer Science and Information Technology (CSIT) IEEE

[15] Tsai C F, Eberle W and Chu C Y 2013 *Genetic Algorithms in Feature and Instance Selection Knowledge-Based System* 39 pp 240-247

[16] Ghareb A S, Bakar A A and Hamdan R A 2015 *Hybrid Feature Selection based on Enhanced Genetic Algorithm for Text Categorization* Expert Systems With Applications pp 31-47

[17] Handoyo R, Rumani M R and Nasution S M 2014 *Perbandingan Metode Clustering menggunakan Metode Single Linkage dan K-Means pada Pengelompokan Dokumen* Jurnal SIFO Mikroskil (JSM) STMK Mikroskil

[18] Mingoti S A and Lima J O 2006 *Comparing SOM Neural Network with Fuzzy c-Means, K-means and Traditional Hierarchical Clustering Algorithms* European Journal of Operational Research pp 1742-1759

[19] Everitt B S 2001 *Cluster Analysis* (New York: John Wiley & Sons)

[20] Castella Q and Sutton C 2013 *Word Storm: Multiples of Word Clouds for Visual Comparison of Documents* Computer Research Repository (CoRR)

[21] X P 2017 *Welcome to My Site PT. X Indonesia Induction 2017* (Jakarta: PT. X)