

Online data preprocessing: a case study approach

Mohammed Zuhair Al-Taie¹, Seifedine Kadry², Joel Pinho Lucas³

¹UTM Big Data Centre, Universiti Teknologi Malaysia, Malaysia

²Department of Mathematics and Computer Science, Faculty of Science, Beirut Arab University, Lebanon

³Tail Target, Brazil

Article Info

Article history:

Received Jan 28, 2018

Revised Aug 11, 2018

Accepted Mar 4, 2019

Keywords:

Data preprocessing

Data science

Flight MH370

Social networks

ABSTRACT

Besides the Internet search facility and e-mails, social networking is now one of the three best uses of the Internet. A tremendous number of volunteers every day write articles, share photos, videos and links at a scope and scale never imagined before. However, because social network data are huge and come from heterogeneous sources, the data are highly susceptible to inconsistency, redundancy, noise, and loss. For data scientists, preparing the data and getting it into a standard format is critical because the quality of data is going to directly affect the performance of mining algorithms that are going to be applied next. Low-quality data will certainly limit the analysis and lower the quality of mining results. To this end, the goal of this study is to provide an overview of the different phases involved in data preprocessing, with a focus on social network data. As a case study, we will show how we applied preprocessing to the data that we collected for the Malaysian Flight MH370 that disappeared in 2014.

*Copyright © 2019 Institute of Advanced Engineering and Science.
All rights reserved.*

Corresponding Author:

Seifedine Kadry,
Department of Mathematics and Computer Science,
Beirut Arab University,
Lebanon.
Email: s.kadry@bau.edu.lb

1. INTRODUCTION

Online social networks or social media websites -the recent form of traditional social networks- have provided the mechanisms that allow people to comprehend, interact, engage and collaborate with each other. Platforms such as Facebook, YouTube, LinkedIn, and Twitter are used these days heavily either for fun or business processes. They allow users to share text, audio, video, hyperlinks, and photographs with others and therefore large volumes of their social activities are stored. Facebook, for instance, stores more than 30 billion new pieces of content every month, produced by more than one billion active users.

The analysis of data that come from different social media sources (for example, Facebook, Twitter, and YouTube) is critical for effective decision making [1]. It can be used to identify trends, develop business opportunities, predict customer behavior and market shifts, crime investigation, and natural disaster risk management.

Data preprocessing consumes most of the time and implementation efforts and can be more critical than the machine-learning algorithm itself. If any of the data preprocessing phases is not correctly handled, machine-learning algorithms will not run or can give misleading results [2]. Having in mind the significance of data preprocessing to data science [3], this study discusses data preprocessing and how we applied it to Flight MH370 social data.

The rest of the study is organized as follows: Section 2 discusses what is meant by social networks and what techniques are used to collect online social data. Section 3 gives an overview of the four tasks involved in data preprocessing. Section 4 describes how we applied data preprocessing to a dataset that we collected describing the Malaysian Flight MH370 social structure. Section 5 concludes the study.

2. ONLINE SOCIAL DATA COLLECTION

Typically, social networks are theoretical models for analyzing and visualizing relationships between actors [4]. A tie between two actors should represent some relationship such as kinship, affection, enmity, exchange of favours or loans, club membership or event attending. Ties can also vary in intensity in the way that some ties are stronger than others. It is also possible to have more than one tie between two individuals in one network. For example, a group of students in a college may be connected to each other through friendship, joint courses, club memberships, etc.

In the past, social data was challenging to collect and hard to come by, and many people in the field limited themselves to using only small amounts of data. However, things have changed with the advent of online social websites such as Twitter, Facebook, and Flickr that generate much more data than anyone would expect.

Beyond the traditional methods of collecting social data, online data can be collected using Application Programming Interfaces (APIs), web crawlers, online surveys, and deployed applications. However, collecting online social data is not always straightforward as it faces some problems [5] such as dealing with unstructured and heterogeneous data, dynamic networks, processing power and data storage. However, and in some cases, taking measurements on all actors in the relevant actor set is not possible. Therefore; taking a sample set of actors from the complete set becomes enough, and inference about the population is made later based on the sample [6]. Sampled data in this case, which can be viewed as representative of the broader population, is called the probability sample.

3. DATA PREPROCESSING

In contrast to organizations, where both the data and the hierarchy of knowledge are well-organized, online social data are rich with user-generated annotations and free-style handwriting. Given that people in online social platforms are entirely free to write what they want [7], data quality in such environments ranges from valuable to commercials and rubbish [8].

Four steps are typically involved in data preprocessing: data cleaning, data integration, data transformation, and data reduction. We will address each step in detail and show how it can be used for knowledge discovery. These processes are not isolated from each other but instead are integral components of data preparation. If any of these steps are not performed as planned, data mining algorithms will not run and will probably give unexpected results. We may safely say that having good and robust data is more important than having an efficient algorithm that is applied to a large quantity of poor data.

3.1. Data cleaning

Data cleaning (also known as data cleansing or scrubbing) aims at removing noise, filling out missing values, and fixing inconsistencies in data. Dirty data in the database can be the result of wrong data entry, update, or transmission [9]. This phase also includes identification and removal of outliers [10]. Performing data cleaning for social network data is not always straightforward as it can face several issues [5]. For example, it requires an understanding and realization of the data. Images, video, audio, failed HTTP records, HTML tags, and white spaces should be removed. Spam detection should also be pursued.

Machine learning methods have been commonly used to perform data cleaning in social data [11]. For example, expectation maximization with Gaussian mixture models is used to manage missing data, while Bayesian models are used for cleaning data. In the blogosphere domain, [12] performed data cleaning: trimming out white spaces, punctuations, and stop words to blog posts. [13] proposed the removal of inactive users from the social network. In the trust-based information propagation model they built, the authors isolated inactive users from other users because they never or rarely send messages and had little influence on information propagation. To deal with noise in Twitter data, the authors in [14], and before applying sentiment analysis, removed questions, URLs, special characters as well as retweets. Targeting Twitter service also, [15] proposed the use of a number of methods to clean data. They used regular expression tools to remove unnecessary tags, English lexicon to filter out words in languages other than English, and an external source to eliminate stop words. They aimed to discover peoples' opinions by using topic models, sentiment analysis as well as geolocation information. In the email communications, data cleaning was applied by [16] to remove non-essential information from the corpus, such as message ID, message timestamp, sender and receiver information as well as all messages with invalid email addresses.

3.2. Data integration

Inconsistency and redundancy in social data are very likely because users have different perspectives and show a different behaviour [13]. Data integration aims at combining data from several sources into coherent data storage [10]. This task is not simple [9] as it requires achieving a match between different schema types. Incomplete or inefficient data integration can cause inconsistency and redundancy,

while a proper implementation will undoubtedly enhance accuracy and speed up later processes. Techniques of data integration include entity identification, tuple deduplication, redundancy and correlation analysis, as well as data value conflict detection and resolution.

3.3. Data transformation

Data transformation aims at converting data into a usable and understandable format. The goal is to have more efficient data mining operations [10]. Strategies for data transformation include data smoothing, feature construction, normalization, discretization, generalization, as well as concept hierarchy generation of nominal data. These subtasks require human supervision and are highly dependent on the data being preprocessed [17]. However, data transformation for social data faces a number of challenges [5] including:

- a. Data having different formats. This is because the social data extracted from online websites can be both categorical and continuous. A proper conversion between the two types, in this case, is required.
- b. Application dependency, which means that each data management tool has its uses and can deal with only particular types of data. For example, some tools (e.g., Pajek and UCINET) are standalone applications, while NodeXL is a utility toolkit. Some packages are provided with programming utilities (e.g., JUNG and SNAP-GAUSS) whereas others provide a graphical user interface that requires no scripting utilities. Finally, some tools are free software (e.g., STRUCTURE and StOCNET) while others have commercial licenses, such as InFlow and SocioMetric LinkAlyzer.
- c. Data privacy and user personal information that can be broken throughout data transformation. One of the remedies is to use data anonymization.

Several studies have addressed this issue. For example, and regarding data transformation of email corpus, [16] proposed the transformation of data into a hierarchical structure, which can be achieved by grouping the object-based relationships of people into a hierarchy. Objects in this unified structure represent emails or threads, while each one can perform different roles, such as a mail sender, mail receiver, thread creator or thread participant. The next step of social data preprocessing would include flattening the hierarchy of objects into one selected level, followed by multi-layering of the social network (see the same previous reference for more details). The authors in [12] performed data stemming on blog posts to transform all the terms into their morphological roots with the help of the stemmer portal. In a different study, [15] applied regular expression techniques to normalize some words to their standard form. They also applied Porter stemming to stem words into a unified form.

3.4. Data reduction

Data reduction aims at reducing the size of data while keeping information loss at a minimum [10]. Techniques of data reduction include:

- a. Dimensionality reduction: reducing the number of attributes or random variables by using techniques such as principal component analysis, wavelet transformation, attribute subset selection and attribute creation. Text data extracted from online social networks can be extremely sparse and high-dimensional [18]. High dimensionality typically means that there are 16 or more data attributes. In such data orientations, the nature of geometrical objects and the concept of proximity are not the same as two or three dimensions. Data cleaning, which aims at removing undesired content, can be considered one method for reducing data size.
- b. Numerosity reduction: achieving smaller data representations by using special techniques such as parametric models (such as regression and log-linear) and nonparametric models (such as sampling, histograms, clustering, and data cube aggregation).
- c. Data compression: compressing data to obtain a reduced copy of the original data. Resultant data can be either lossless or lossy, depending on whether there was a loss of information during data compression.

Both dimensionality reduction and numerosity reduction can be applied to social data to have a smaller data volume while at the same time producing the same (or near the same) mining results. For example, to provide smaller representations of the original data, [13] performed dimensionality reduction on a Twitter corpus to remove useless keywords from the old messages, and used numerosity techniques to replace user's messages with the least possible keywords that represent the main types and significant sections of the messages. In a different study, [19] applied attribute reduction to improve the performance of spam detectors on Twitter data by computing correlation coefficients between attributes [20-22].

4. CASE STUDY: FLIGHT MH370

This section discusses how we implemented preprocessing to Flight MH370 social data. After preprocessing, we used the resultant data to examine the flight community structure, discover types of social relationships, reveal the truth behind some of the unusual events, and study people coping behavior (adaptation patterns) during disaster time [23].

The Boeing 777-200 ER, which carried 239 individuals, left Kuala Lumpur International Airport (KLIA) at 12.41 am. It was supposed to land at Beijing Capital International Airport after six hours. However, the Boeing disappeared from radar after only 40 minutes from its take-off. The passengers were from China, Malaysia, Indonesia, Australia, France, India, United States, New Zealand, Canada, Ukraine, Russia, Taiwan, Italy, Netherlands, and Austria. A group of employees from the American-based Freescale Semiconductor Corporation was on a visit to China to take a technical course for one month. A group of Chinese artists was on the way back from a visit to a cultural exhibition in the Malaysian capital, themed Chinese Dream: Red and Green Painting. Some passengers were on travel from China to Malaysia to attend a Buddhist religious ceremony. The airplane also carried a group of workers who worked in Singapore, a group of tourists who were on a trip to Nepal, a number of families, as well as passengers just making a stopover. The crew of the airplane has 12 Malaysian citizens (2 pilots and 10 attendants).

4.1. MH370 social data collection

As we said earlier in “online social data collection,” collecting social data is not straightforward and requires much effort, in particular, that such data are unstructured and come from various sources, added to other issues such as storage, processing, and dynamicity of the network.

The dataset that we collected came from tens of different online news websites (such as The New Indian Express, YAHOO! News Malaysia, the Economic Times, CNN.com, the Daily Express, and India Today). Data gathering was done over March and April in 2014 covering the events from when the airplane disappeared until late of April, when most of the international efforts to locate the missing flight stopped, without achieving their goals.

Online news media addressed this event in three different ways: how the airplane disappeared, the background of the passengers, how relatives were interacting with news and the efforts made by the international society to locate the missing airplane.

4.2. MH370 social data preparation

As we said earlier in this study, in order to efficiently mine and analyze collected data, data preprocessing must be implemented as a series of sequential steps. First: data cleaning, which aims at eliminating irrelevant and redundant content? Second: data integration, which aims at integrating data from different sources into a unique and uniform data endpoint. Third: data transformation, which aims at converting the resultant data into a format that is easy for analysis. Finally: data reduction, which aims at having a smaller (both in volume and dimension), concise and clean, representation of the data that will undergo analysis later.

4.2.1. MH370 social data cleaning

The first step of Flight MH370 data preprocessing was to check and evaluate all the attributes of the original data. As we do in other data cleaning processes, we prioritize data quality over data scale. It is preferable to reduce data volume to have more representative information, rather than poor and inconsistent data.

In the case of Flight MH370, where data were collected from various online sources, passenger demographic information was mostly inconsistent, ambiguous or missing. Accordingly, we preferred to remove all demographic attributes except “Age.” As will be described in subsequent sections, we used ages to differentiate users having the same name and to help explore groups and connections.

Using Pajek Software, it was possible to detect visually and correct some outliers and inconsistencies related to the passenger age. We were able to probe some data sources and identify inconsistencies, and using other data sources and relying on most common data among such sources; we could smooth out noisy data like that.

In this phase, we removed unwanted and irrelevant information from data. For a study on social networks, we were concerned only with the information that would help reveal relations and connections between passengers. It was a time-consuming task since online social data tend to have noise, outliers, missing values and other inconsistencies. Other methods and packages can also be used by data scientists for this purpose, such the R Software; and Python packages for data manipulation, such as Pandas and NumPy.

4.2.2. MH370 social data integration

We collected our data from different social media websites over two months. Because the data were mostly unstructured text, we had to run some data manipulation scripts to process the text. The goal was to extract a simple and more unified data.

In this phase, we detected and solved data value conflicts on passenger data such as the use of different representations, and different scales (i.e., metric vs. British units, or others). Afterwards, we

reformulated and unified the data such that we stored passenger information in more recognizable formats. All the information about the flight passengers and their families that we were able to integrate are shown in Figure 1.

Item	Passenger Name	Nationality	Age	Connected-to	Alternative Name	Symbol	Background
1.	WATTRELOS/AMBRE	French	14	Laurence Wattrelos(52)	Ambre Wattrelos(14)	P ₁	Ambre Wattrelos, Student in Beijing at Lycee Francois International de Pekin. Moved to China 6 years ago. At the plane with her mother Laurence Wattrelos, her brother Hadrien, and his girlfriend Zhao Yan.
				Hadrien Wattrelos(17)	Ambre Wattrelos(14)		
				Zhao Yan(18)	Ambre Wattrelos(14)		
				Flight	Ambre Wattrelos(14)		
2.	ANWENLAN	Chinese	65	Zhang Zhonghai(43)	An Wenlan(65)	P ₂	An Wenlan, Painter. Devout Buddhist. Linked to her daughter Zhu Junyan. She studied painting with Yao Jianfeng. One of a group of 24 Chinese artists who had travelled to Kuala Lumpur to attend an art exhibition. Resident of Hababe county in Northern Xinjiang province
				Bao Yuanhua(63)	An Wenlan(65)		
				Che Junzhang(68)	An Wenlan(65)		
				Dong Guowei(48)	An Wenlan(65)		
				Feng Jixin(70)	An Wenlan(65)		

Figure 1. Raw data of Flight MH370

In Figure 1 above, the data that describe names, nationalities, and ages were expanded from MH370 Passenger Manifest, released by Malaysia Airlines the same day of the disaster. Column “Connected-to” represents the people that a particular passenger is connected to. This includes family members, co-workers, and friends. The data in that column were collected based on statements from family members, journal reports or officials. However, officials were accused of giving little information about the accident.

Column “Alternative Name” has the names that we used later for analysis instead of the list of names provided by the MH370 Passenger Manifest which, in turn, are shown in column “Passenger Name.” Each name in column “Alternative Name” begins with the passenger’s first name followed by the family name (names with two parts were also considered). We also removed unnecessary word capitalizations and suffixed the second name of each passenger with his/her age to avoid name confusions (for instance, the name *Zhang Yan (36)* is also the name of another passenger, *Zhang Yan (45)*). For simplicity, we preferred to use only the passenger’s first and second name followed by age.

Column “Symbol” has the labels that we gave to the passengers. For example, P_1 , P_2 , P_3 refer, respectively, to the first three passengers *Ambre Wattrelos (14)*, *An Wenlan (65)* and *Andrew Nari (49)* in the table. These symbols were used to label nodes when we built the graphical representations. Column “Background” contains information over two months about the passengers. It tells about the career, marital status, family members, and friendships of each one.

4.2.3. MH370 Social Data Transformation

One of the tasks that were actively present when we built our database was data transformation of the passenger name. We used a simple heuristics for this purpose. The task included reviewing the table row by row, transforming each name into more a proper structure (i.e., alternative name), and giving it a proper identifier (i.e., label).

One of the things that we did for data transformation was adding extra nodes that serve particular purposes. For example, in addition to the original 239 nodes, passenger nodes were tied to a central node, called “Flight,” to show that all the passengers were present at the same event and were part of the flight network structure. Another node “Singapore” was added to connect the Chinese workers who worked in Singapore altogether. Three of those workers were confirmed to be friends and to know each other, and therefore, they were directly connected to each other and the central node of the group. The resulting network is a small undirected network that has no loops, having every two passengers connected.

Since our data preprocessing was mostly involved in text manipulation, we did not need to perform (except for feature construction) other types of data transformation. Nevertheless, and depending on the

mining purpose, other methods are also used for data transformation. This includes data aggregation (summarize numeric attributes based on their concept hierarchy), data generalization (analogous to aggregation but for nominal attributes), data normalization (scaling numeric attribute values to fall within specified ranges), and others.

4.2.4. MH370 social data reduction

As we mentioned earlier in the data cleaning section, data preprocessing should prioritize data quality over data scale. In this way, data reduction that we performed on flight MH370 data has also given up of some data volume in exchange for more representative and consistent information. It is important to highlight that the operations performed during data cleaning have a different purpose from these performed in data reduction. While the earlier aims at eliminating noisy and inconsistent data, the latter aims at reducing data representation and making it simpler and more useful, particularly for visual analysis.

The table discussed earlier contains some raw data that we did not need during the analysis phase. Therefore, we extracted only the data that show the relationship between the people who joined the flight. We compiled the resultant data into a new Microsoft Excel file as shown in Figure 2.

Connected-to	Names	Labels
Laurence Wattrelos(52)	Ambre Wattrelos(14)	P1
Hadrien Wattrelos(17)	Ambre Wattrelos(14)	P1
Zhao Yan(18)	Ambre Wattrelos(14)	P1
Flight	Ambre Wattrelos(14)	P1
Zhang Zhonghai(43)	An Wenlan(65)	P2
Bao Yuanhua(63)	An Wenlan(65)	P2
Che Junzhang(68)	An Wenlan(65)	P2
Dong Guowei(48)	An Wenlan(65)	P2
Feng Jixin(70)	An Wenlan(65)	P2
Fu Baofeng(28)	An Wenlan(65)	P2
Hou Bo(35)	An Wenlan(65)	P2
Li Guohui(56)	An Wenlan(65)	P2
Li Mingzhong(69)	An Wenlan(65)	P2
Li Zhijin(30)	An Wenlan(65)	P2
Liu Rusheng(76)	An Wenlan(65)	P2
Liu Zhongfu(72)	An Wenlan(65)	P2
Lou Baotang(79)	An Wenlan(65)	P2
Mao Tugui(72)	An Wenlan(65)	P2
Memetjan Abula(35)	An Wenlan(65)	P2
Meng Gaosheng(64)	An Wenlan(65)	P2
Su Qiangguo(71)	An Wenlan(65)	P2
Wang Linshi(59)	An Wenlan(65)	P2
Xiong Deming(63)	An Wenlan(65)	P2

Figure 2. Sample data from Flight MH370 dataset

The dataset is available from the first author's and GitHub (<https://github.com/mohammed-taie/Flight-MH370>) repository. The dataset has 241 vertices and 1563 edges, and contains three columns:

1. Connected-to: individuals that a particular passenger is connected to
2. Names: names that were used during data analysis
3. Labels: labels that we gave to the passengers

It was essential to transform the dataset into the ".net" format before embarking on data analysis and data visualization. The resulting .net file has two parts:

1. Vertices to denote flight passengers. For data visualization, all the vertices were given the same shape and the same size, even though they may have a different number of connections. A few additional vertices were also added to achieve particular benefits.
2. Edges to connect the members of the same group and to connect the different groups. For a better data visualization, edge labels and weights were eliminated. Labels and weights are sometimes used to show the strength (among other things) of a connection between two nodes.

Depending on how big the data are, other methods can also be applied for data reduction such as clustering (grouping sample values in clusters based on a similarity measure) and sampling (taking a sample set of actors from the complete set).

Connecting the nodes helped us spot three large community structures [20]: the "Artists Group," consisting of 29 members, the "Freescale Semiconductor" group, comprised of 15 members and the "Aircraft Crew" consisting of 12 members. Other smaller groups include six people and five people families from China, four people family from Malaysia, a French family, an Australian family, and others. The analysis also showed 8 nodes (two pilots, five toddlers, and one physically challenged elderly woman) that are not directly connected to node "Flight."

5. CONCLUSION

The goal of this study is to understand data preprocessing in the light of Flight MH370 social data. Data preprocessing is a critical part of data science projects, mainly that real-world data often have noise, outliers, missing values and other inconsistencies. The steps should be introduced efficiently, supported by human experience, and to be applied more than once. It is common that the preprocessing steps are reiterated if the results of data mining are significantly different from what is expected.

Handling online social data (regarding data collection and data preprocessing) is a hard task and can become even more complicated if the data that we are concerned with come from online news pages that are commonly known to have problems in credibility and accuracy (except for some cases). One of the problems that we encountered when we performed MH370 social data preprocessing is how to deal with missing data, shortage in uniformity, and lack of some critical information. One of the reasons that the available information was not enough is that officials abstained from revealing much information about the accident for fear of triggering political issues between Malaysia and some other countries. Regarding accuracy, some information that was included in the MH370 Passenger Manifest (released by Malaysian Airlines) was not accurate.

REFERENCES

- [1] S. Kadry and M. Z. Al-Taie, "Social network analysis: An introduction with an extensive implementation to a large-scale online network using Pajek," Bentham Science Publishers, 2014.
- [2] M. S. Brown, "Data mining for dummies," John Wiley & Sons, 2014.
- [3] M. Z. Al-Taie, et al., "Successful Data Science Projects: Lessons Learned from Kaggle Competition," *Kurdistan Journal of Applied Research*, vol/issue: 2(3), 2017.
- [4] S. Wasserman and K. Faust, "Social network analysis: Methods and applications," Cambridge university press, vol. 8, 1994.
- [5] P. Gupta and V. Bhatnagar, "Data preprocessing for dynamic social network analysis," *Data Mining in Dynamic Social Networks and Fuzzy Systems*, pp. 25-39, 2013.
- [6] M. Al-Taie and S. Kadry, "Python for Graph and Network Analysis," Springer, pp. 1-184, 2017.
- [7] M. Zuber, "A survey of data mining techniques for social network analysis," *International Journal of Research in Computer Engineering & Electronics*, vol/issue: 3(6), 2014.
- [8] J. Bian, et al., "Finding the right facts in the crowd: factoid question answering over social media," *Proceedings of the 17th international conference on World Wide Web*, 2008.
- [9] S. García, et al., "Data preprocessing in data mining," Springer, 2015.
- [10] J. Han, et al., "Data mining: concepts and techniques," Elsevier, 2011.
- [11] S. S. De and S. Dehuri, "Machine Learning for Auspicious Social Network Mining," *Social Networking*, Springer, pp. 45-83, 2014.
- [12] N. Agarwal, et al., "Clustering of blog sites using collective wisdom," *Computational Social Network Analysis*, Springer, pp. 107-134, 2010.
- [13] L. Wenxue and G. Sun, "A trust-based information propagation model in online social networks," vol.issue: 8(8), pp. 1767, 2013.
- [14] I. Hemalatha, et al., "Preprocessing the informal text for efficient sentiment analysis," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, vol/issue: 1(2), pp. 58-61, 2012.
- [15] T. H. Wen, et al., "Analysis of combining Topic model, Sentiment, Geolocation information approaches on Social Network."
- [16] P. Kazienko, et al., "A generic model for multidimensional temporal social network," *Communications in Computer and Information Science, CCIS*, vol. 171, pp. 1-14, 2011.
- [17] M. Al-Taie and S. Kadry, "Applying Social Network Analysis to Analyze a Web-Based Community," arXiv preprint arXiv:1212.6050, 2012.
- [18] C. C. Aggarwal and C. K. Reddy, "Data clustering: algorithms and applications," CRC press, 2013.
- [19] M. Klassen, "Twitter data preprocessing for spam detection," *FUTURE COMPUTING 2013, The Fifth International Conference on Future Computational Technologies and Applications*, Citeseer, 2013.
- [20] B. N. Octaviana and J. Abraham, "Tolerance for Emotional Internet Infidelity and Its Correlate with Relationship Flourishing," *International Journal of Electrical and Computer Engineering (IJECE)*, vol/issue: 8(5), pp. 3158-3168, 2018.
- [21] C. Virmani, et al., "Clustering in Aggregated User Profiles across Multiple Social Networks," *International Journal of Electrical and Computer Engineering (IJECE)*, vol/issue: 7(6), pp. 3692-3699, 2017.
- [22] Y. L. S. Rani, et al., "Extensive Analysis on Generation and Consensus Mechanisms of Clustering Ensemble: A Survey," *International Journal of Electrical and Computer Engineering (IJECE)*, vol/issue: 8(4), pp. 2351-2357.
- [23] M. Z. Al-Taie, et al., "Flight MH370 community structure," *Int. J. Advance. Soft Comput. Appl*, vol/issue: 6(2), 2014.