CHAPTER 2

LITERATURE REVIEW

2.1. Empirical Study

Empirical study will elaborate the relevant previous researches that is correlated with the methods used in this research to be used as the references to develop the methods and solving the problems in this research.

(Momtazi 2018) implemented the use of topic modelling Latent Dirichlet Allocation in community-based questions answering. The study is aimed to provide the most probable questions answering in a community. The topic modelling model classifies the questions and provides the system with information about the question categories in topics. The topic is gained from using LDA method then send it to the system to be learnt and then the system will predict the label of unseen documents using topic-categories. The result shows that the model used by researcher is better to another multi-label classification algorithms.

(Bastani, Namavari and Shaffer 2019) wrote a research about the implementation of LDA topic modelling in Consumer Financial Protection Bureau

(CFPB) datasets of consumers complaints in various financial services. The researcher creates an intelligent system to analyse those complaints. LDA generates semantically useful topics from customer complaints, related to the available financial services and the topics then used to monitor the current complained services from the available complaints in any period of monitoring. LDA is proven useful and creates a more efficient complaints' monitoring in this topic modelling model used by the researcher.

(Akhtar, et al. 2017) in this conference proceeding, created a summarizing system of hotel booking website reviews overview. Providing the system with the sentiment analysis result, the dataset used is taken from *TripAdvisor* website. Then LDA topic modelling technique is applied to identify hidden information and aspects of each reviewed hotel. This system is aimed to assist the potential customers in summarizing the reviews available and topic modelling technique that is implemented to collect the information for the customers as well for the business owners' advantage on findings what aspects that are criticized by reviewers.

(Shams and Dastjerdi 2017) applied enriched LDA method, combination of LDA topic modelling applied in aspect extractions in a textual data. The research experimented on applying the method on a shopping websites product dataset. The result shows there is no significance result differences from basic LDA aspect extraction of a dataset. However, ELDA generates the result the extracted aspects with richer knowledge to be judged by human perception.

(Li, et al. 2017) studied the implementation of LDA Model to assist financial stability report visualization and quantification. The type of the report data is basically a textual data, which is LDA Model will be useful to visualize those data. The main

purpose of this study is to get the tendency data of financial stability in the development role of a country which in this case is china. Later, LDA will get the topics on the report that is simplify the report and make it possible to create a mapping of description on the report and finding the embranchment on the financial aspect development and enable the study of the financial stability tendency easier.

(Rao, et al. 2014) studied the implementation of sentiment topic modelling focused on social media emotion mining. Mostly this research tries to compare the use of two-reader oriented sentiment topic models, which are Multi-label supervised topic model (MSTM) and Sentiment Latent Topic Model (SLTM), to connect the latent topic gained from topic modelling in the previous steps that later used to evoke as the lexicon in emotion mining. The result shows that the combine modelling method proposed by author is more stable than the baseline Emotion Topic Modelling with a better p-value which is equal to 0.0005 and 8.0e-22 respectively. This indicates that the proposed model of combine method in emotion mining can discover meaningful latent topics in strong social emotions.

(Srinivas and Rajendran 2019) studied about the problems in retention rates in a university that over the years resulting million dollars of revenues in the perspective of the campus. Aiming to assist the strategic planning of enrolling in a university providing the planning strategy with data-driven decision for students. Using topic-based knowledge mining on the information of universities reviewed online and available all over the platforms on the internet. The idea is to assist the information availability with a precise topic knowledge on universities so that student can enrol based on these data. It uses E-LDA (Ensemble Latent Dirichlet Allocation) as the method to analyse such data available in textual data form. The result shows E-LDA

can automatically generate accurately the review data to their corresponding topics, which in this case author identified 12 meaningful topics. Then, additional method to assist the planning decision for students is applied, such that SWOT analysis that is employed to the selected university as study case to determine the critical factors that influence student's perspective towards the university.

(Twinandilla, et al. 2018) used the advantage of LDA and combined it with K-Means Clustering, in generating Significance Sentences in Multiple documents. The study case presented by this research is applied in case of yellow journalism case, which is the difficulty of separating opinion and facts in terms of blurry and difficult textual journalism reports, in terms of human perception. It takes time to understand this case. The result gives good result of summarizing each specific topic on the analysed case, as shown by the value of ROUGE-1 (value of resulted cluster) showing the value 061991 for the first model cluster and 0.6139 respectively.

The research position is shown by Table 2.1. This table shows the related researches about Topic modelling with various methods and their applications.

Table 2. 1 Literature Survey

No	Author	Year		Research Focus				Object		
			Self-Service	Business	Performance	Design & Development	Critical Success	Manufact	Re	Service
1	Momtazi	2018		✓	✓	✓				✓
2	Bastani et al	2019		✓		✓	✓			✓
3	Akhtar et al	2017		✓		✓				✓

4	Shams and Dastjerdi	2017			✓	✓				
5	Li et al	2017		✓		✓				✓
6	Rao et al.	2014		✓	✓	✓				
7	Srivinas and Rajendran	2019	✓	✓	✓	✓	✓			✓
8	Twinandilla et al	2018		✓	✓	✓				✓
9	Elrienanto	2019		✓		✓		✓		

2.2. Electronic Commerce (E-Commerce)

Electronic Commerce (E-Commerce) is a new concept that came into business terms no sooner than 1970s. The idea is the presence of internet becomes the connector of e-commerce for business activities by the year of 2000. (Wigand 2014). E-commerce is basically a business interaction inside a system that poses inter-connected communication systems which consists of data management and system security in relation to commercial information that the system conducts that includes product sales information or services will be provided by the system. That enables people to do business activities literally everywhere in which internet connection and information system devices is available. (Nanehkarana 2013).

The frameworks of E-commerce which generally consists of internet connections, software, hardware, and database, that serve the purpose of information provider of business' product availability, negotiation and agreements process and trading partners which is presented in real-time in terms of availability of the business activity. And serve the purpose of product sales in terms of availability of the product in the current market, transactions, and chain of supply and support for both customers and business owners. Not only products can be sold in E-commerce some serve the purpose of service information availability. (Nanehkarana 2013).

2.3. Samsung Galaxy S9 Smartphone

Samsung Galaxy S9 was first released on March 16th, 2018. With body dimensions of 147.7 x 68.7 x 8.5 mm (5.81 x 2.70 x 0.33 in), weight 163 g (5.75 oz) built with Front/back glass (Gorilla Glass 5), aluminium frame, equipped with Single SIM (Nano-SIM) or Hybrid Dual SIM (Nano-SIM, dual stand-by) Samsung Pay (Visa, MasterCard certified) IP68 dust/water proof (up to 1.5m for 30 mins). Display specifications of this phone equipped with Super AMOLED capacitive touchscreen, 16M colours, with the size of display is 5.8 inches, 84.8 cm2 (~83.6% screen-to-body ratio) and screen resolution ratio is 1440 x 2960 pixels, 18.5:9 ratio (~570 ppi density). And software specifications equipped with Android 8.0 (Oreo), upgradable to Android 9.0 (Pie). With some card slot and memory slot of microSD, up to 1 TB (uses shared SIM slot) - dual SIM model only Internal 64GB 4GB RAM, 128GB 4GB RAM, 256GB 4GB RAM.

Figure 2. 1. Samsung Galaxy S9 Smartphone

Source: *https://www.amazon.com/Samsung-Galaxy-S9-Unlocked-Smartphone/dp/B079H6RLKQ*

2.4. Amazon Customers Review

Amazon was first founded by Jeff Bezos in 1995. Serving as a media of virtual shopping platforms for book lovers. It becomes a major success back in that time, then nowadays it grows to serve not just selling books but other stuffs as well. Some notable features of amazon such that the information system that amazon poses creates the condition of a safe transaction for customers, for example a web service from amazon called Amazon Web Service (AWS) that maintain the big number of products posted and millions of customers within in their system. Amazon becomes world dominator due to some distinct features of handling customers' needs with the features they possess.

One of the parts of Customer Relationship Management (CRM) that amazon conduct in their business process is the existence of product review in amazon. This system serves the purpose as the benchmarking start before a customer decide whether to buy the product or not. Product review by customer not only be help customers to find and select product but also acts as direct marketing event for amazon. Amazon allows making connection in between the customer, communicating and viewing their interest help to amazon to gain customer trust and initiate world of mouth. (Al Imran 2014).

2.5. Feature-Based Opinion Mining

Feature based opinion mining is one of the applications of opinion mining to classify the polarity of a given text and extract the feature of given dataset whether it indicates a positive or negative feedbacks. Product feature extraction is an important task of this

method. The goal is to get the idea of what features the customers talk about and express themselves. There are many approaches in conducting feature-based opinion mining, such as using maximum entropy that applies probability distribution that is widely used for NLP process, unsupervised learning like LDA can also be applied to this method. The general steps in conducting feature-based opinion mining are (Vinodhini, Srisubha and Chandrasekara 2012);

1. Dataset Scrapping from the desired web sources that must contain text data about reviews of a product,
2. Pre-processing documents are then cleaned to remove tags, after that, extract only text of reviews. Reviews are split into sentences and make a bag of sentences. After extraction reduplicate reviews are removed and the rest of the reviews are stored in the database,
3. Text Extraction uses the basic NLP library and other libraries that is available for text processing technique,
4. Association Mining to find the most likely frequent words or sentences after the words from dataset are separated and so the feature can be extracted,
5. Feature Ranking is to rank the most frequent features that appears from the process of association mining, and getting the ideas of other words associated from the features that mostly nouns, for example good, bad, excellent, etc.

2.6. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a simple algorithm for the topic modelling. In machine learning, LDA is included in the group unsupervised learning. LDA appears as one of the methods used for analysis on very large documents. LDA is also used to

summarize, grouping, connecting or processing data. Because LDA produces a set of lists of topics that are weighted for each document (Campbell, Hindle and Stroulia 2014). For each document in the collection, the LDA algorithm takes the topic is based on the distribution of multinomial words in the document. Then the topic is used to generate the word itself based on the multinomial distribution of topics and repeat these two steps for all words in the document. Processing the distribution of words in the document is done to determine whether they are originated from the same topic or not. Distribution process is carried out for obtaining the distribution of these topics. In LDA, document is an object that can be seen, while the distribution of topics per document is hidden structure, therefore this method is called the Latent Dirichlet Allocation (LDA).

LDA is a generative probabilistic model of a set of texts called the corpus. There are three generative processes that are carried out in each document in corpus (Blei, et al. 2003):

1. Choose a topic randomly from the distribution of topics for each document.
2. Choose words from the distribution of words related to the selected topic.
3. Repeat process 1 and 2 for all existing documents.

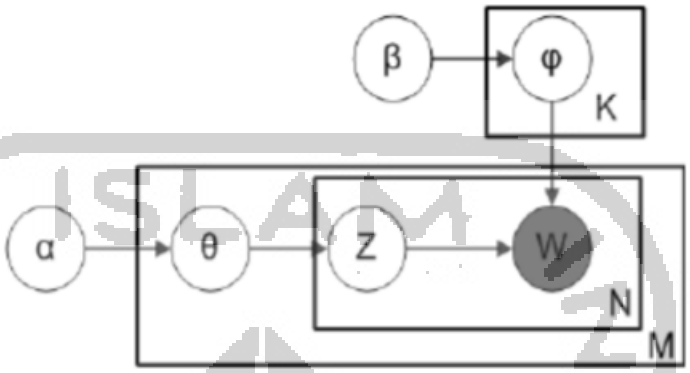

Figure 2. 2. LDA plate notation (Blei, et al. 2003)

From Figure 2.2, there are 3 levels that represent the LDA model. The parameters α and β are parameters existed outside the plate that is included in the level of the corpus, which is a collection of M document. The parameter α is used in determining the number of distributions in the topic document. The higher value α, indicating the more topics are covered. The β parameter is used to determine the number of word distributions in a topic. The higher value β, indicating the more words contained in a topic, whereas the lower the value of β, the fewer words that are on the topic, so the topic contains more specific words. The variable θm is the variable at the document level, M. M, indicates that the variable in repeated M times, for each document. The variable θm represents topic distribution for certain documents. The higher the value θ, the more it indicates many topics existed in the document, whereas the lower the value θ, then the topics are contained in the document are more specific. The variables zn and wn are variables of the word level in the document, which is N. N indicates that the variable is at therein repeated N times, for each word. The variable z represents the topic from certain words. While, the variable w represents words related to the topic certain in the document. Circles represent individual words. Circle

the grey represents the observed variable, while the circle is empty represents latent variables or variables that are not directly observed.

2.7. Kansei Engineering

Kansei engineering is a methods of product development in form of a 2-way communicated information flows, that translates customer's impressions and demands on an existing product into design solutions or some conceptual design parameters. Kansei engineering is mainly acting as a media in a systematically designed solution and innovation findings on a product development process, or it can also play its role in product improvement process of an existing products. (Schutte, et al. 2004).

There are various procedures to conduct Kansei Engineering. This research applies the procedure of Kansei Engineering integrated with text mining procedure, proposed by (Hsia, Chen and Lin 2017). The proposed framework of Kansei Engineering that integrated with text mining are:

a. Selecting Design Domain and Data Collecting

The selection of desired application Kansei Engineering and the data collection method that can be applied later in text mining.

b. Identifying Product Features and the Related Kansei Words

This step is aimed to find the features of the products and the words of impressions related to the products that are known as Kansei Words.

c. Synthesizing Product Features and the Related Kansei Words

A method of evaluating the relation of features and its Kansei words that is conducted to support the data

d. Generating Product Improvement Guidelines

Later, the product development guidelines can be determined as a suggestion provider in a summarized description about the feature and the related Kansei words