CHAPTER 1
INTRODUCTION

This chapter discusses seven points, namely first, the rationale which explains the background of this study and the related situation, Second, the conceptual framework, it explains the method of this study in brief. Third, the statement of the problem, it explains problems that need to be solved in this study. Fourth, the objective, it describes aim of study, and fifth, is the hypotheses which discusses the proposed approach to solve the problem. Sixth, The scope and delimitation, it describes the limitation of data, and finally, significance of study which explains the contribution of this study in the field of data text mining.

1.1 Rationale

Dissemination of information on the internet is very fast that the informations becomes large so that the information needed is difficult obtain [1]. The growth of information affect social development that makes long distance become shorter so that it is not a problem, it also change someone of doing business activity through internal media or often called electronic commercial or more popular with the name of e-commerce. With e-commerce, buying and selling activities between buyers and sellers can be done as if they meet each other directly through different district, cities, provinces and even countries through the internet network. Information spreads over an unstructured text network via websites. Information about a particular product is called a review, whereas information about certain products obtained from other customers is customer review. Information in the form of a customer review is useful both for the customer and the manufacture industries. For customer, this information can be used to get a review of product view in term of the advantage or disadvantages of particular products. As for manufacturers, it is used to get complaint from customers of the product's so as to improve service, innovation and product quality from previous products.

According to Hu and Liu[2006] there are three types of review formats available on the website:

1. Format (1) Pros and Cons. Reviewers describe the product separately between pros and cons. Example C|net.com
2. Format (2) Pros, Cons, and detail review. Reviewers provide separated product description between pros and cons as well as explaining the product review details. Example epinions.com and MSN.
3. Format (3) free format. Reviewers describe a particular product freely without being restricted by pros and cons. Example amazaon.com

According to [Kurniawan, 2013], there are two types of persons searching for the information that they need. First, peoples already know clearly the information of a particular product feature, This type of people usually are helped by a machine called the search engines. While the second is the person who does not know the product feature of particular products, this type of people usually need more than search engines, they need information from various sources or forms and they need to compare the features of similar products.

The product sought by the user on e-commerce varies according to the needs of the users. [Sohail et al., 2013] conducted research on the book domain, [Hu and Liu, 2006] and [Cao and Li, 2007] research on the domain of electronic products especially computers. [Hu and Liu, 2004], [Hu and Liu,2006] and [Aciar et al.,2007] [Zhang and Xu,2016] research on domain camera, [Aciar,2010] research on tourism and [Rozi et al.,2014],[Widyantoro and Baizal,2014], and [Baizal et al.,2016] using the domain mobile phone. This study proposes features extraction explicit and implicit feature products review on format (3) with the product domain of the mobile phones, because mobile phones are currently as primary requirement, the develop and change quickly and have many features.

Before deciding to determine the products, users search for information on the internet in the form of customer review from various websites. Information sought by the user in looking for the mobile phone needed is by reading reviews of various users who have used the phones before. The review usually contains the quality, advantages and lack of features a particular product.

Customer review has a large number of reviews, there are reviewers who provide a review consisting of several features in the reviews column either composed of paragraphs or sentences. In several sentences, the reviewers are provided with different feature description, therefore it needs a system that can summarize the review so that they can become a conclusion based on product features [5] [12] [13] [7] or called feature extraction. To determine a sentence that contains a particular feature of extraction on a sentence can be seen from words that contain product features directly called explicit, but there are some words that indirectly say the product feature is to show the characteristic of features called implicit. An example of a review showing the characteristics of product features is something like "The battery life of this phone is too short." This sentences contains explicit feature because the sentence contains the word "battery life" that appears directly on sentence. However, the implicit message on sentence is something like "There are other phones which are cheaper and better, I think it's too expensive." People know that the sentence has a price feature but the system does not know because the price does not express directly. However, the word "expensive " indicates the price of product.

1.2 Conceptual Framework

The basic concept in this study divides three block processes (input from free text , processing, output) which can be seen in Figure 1.1.

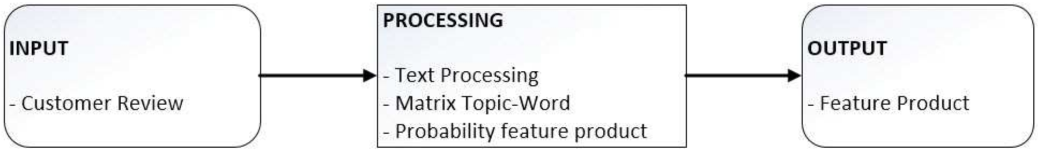

FIGURE 1.1: Conceptual Framework of Summary of Products

Firstly, customer review is used as input. Review on this research using review of each product smartphone. The review collected with user generate content by ASIN (Amazon Serial Identification Number) on amazon.com[14]. Secondly, on processing review from customer using text mining processing [13] [15] [7] such as tokenize, stop word, Stemming and Part-Of-Speech tagging used to facilitate computing separate review sentences that do contain a reviews of certain product specifications. The expected output is feature extraction of particular products such as camera, battery, screen, price, etc.

1.3 Statement of the Problem

Considering point 1.1, review information is available on the internet from free text that can be accessed with a very large number raises information overload. The initial steps are taken in this study from all the reviews that have been collected [14], taken at the reviews that are relevant to a particular product. Every review contains a few sentences. Sentences which are a subset of the reviews contain information that show sentence product features (example: camera, battery, screen, etc) or explicit feature and the words do not appear on sentence or implicit features product (example: blurry, sharp, resolution, etc). Sentences that belong implicit feature products on research [7] is not mentioned, therefore the word that show explicit features of certain product features need to be analyze so that it can extraction of product features. Analysis on the review found that there is a feature word product that has a multi-word that shows product features such as word “blur”, it can show the camera or screen product feature.

1.4 Objective

This study uses the format (3); it is a free format on amazon site. The product feature of a review is extracted from cell phone and accessories. After analyzing the words that have multi-word on product features, dummy data of each product feature need to be created. This is to separate the characteristics of product features and to measure the performance of Sentence Level Topic Model (SLTM) method as a feature extract on the sentence. This system will help user to obtain information about the features of the product easily and faster, and also the manufacturer can respond the complaint from their customers, so that they can increase their services and improve the feature products, and to innovate their next generation of their mobile phones.

1.5 Hypotheses

The information in the review on the internet has two types of product features, explicit and implicit features. The Sentence Level Topic Model method can extract the sentences information in the review. The word tagging used in the SLTM method uses the noun parameter to show product features, the adjectives expressing from perceived users to product features and verbs are how to use the product features.

1.6 Scope and Delimitation

The scope and delimitation of the study are:

1. This research uses data specification of product feature from research Baizal[2015] and based on domain expert with features product camera, battery, endurance, screen, network, audio, price, and system.
2. This research uses data specification products from `http://www.gsmarena.com/` which this page provide detail specification of smartphones.

3. This research uses data of customer reviews from e-commerce site https://www.amazon.com/ to obtain data of reviews to generate the content of each product base on ASIN (Amazon Standard Indentification Number) [14].

4. The preprocessing getting word in basic form using lemmatization because it refers to word in dictionaries and morphological analysis and it does not only removes the affixes to get the word base.

5. The data used as the input on this review are only from a mobile phone domain, otherwise the system will not recognize the product features.

1.7 Significance of Study

This research can improve the information extraction review from huge amount of review which causes the information to be overload. Customer reviews dataset that available on the internet are often found to be irrelevant features, noise the data, redudancy and interacition between attribute, as well as a small ratio between number of samples and number of features [17].Therefore, in this study dummy dataset is created so that it can be controlled to resolve the characteristics of customer reviews datasets and also to know the performance of the extraction feature method that we used.

Preprocessing stage is used first, the data can be established in the process of data mining. There are several processes such as lower, tokenize, stop-word, stemming and parts-of-speech tagging. The output of preprocessing is the words with type of tagging, which determine tagging result is the stemming process. The previous research applies stemming Porter but the tagging result is still roug. This research proposes lemmatization stemming so that output is smooth. For example the word “goes”result from the porter is “goe” while using lemmatization is “go”.The resulting tagging output using the word porter has noun tagging, whereas using lemmatization it has verb tagging.

This study focuses on the feature extraction product, method sentence level topic model (SLTM) is used with word or term as learning using seven parameters of tagging such as noun, adjective, verb, noun with adjective, noun with verb, adjective with verb and combination of three tagging verb, adjective and verb.