

**ASPECT BASED SENTIMENT ANALYSIS FOR EXTRACTING KANSEI
WORD USING SPACY LIBRARY
(A CASE STUDY ON SMARTPHONE PRODUCT)**

THESIS

Submitted to International Program

Department of Industrial Engineering

The Requirements for the degree of

Sarjana Teknik Industri

at

Universitas Islam Indonesia



Submitted by :

Jeffri Surya Dharma (15522184)

INDUSTRIAL ENGINEERING DEPARTMENT

INTERNATIONAL PROGRAM

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2021

AUTHENTICITY STATEMENT

For the sake of Allah, I confess that this research was conducted by me except for the summaries of the sources that have been cited and mentioned. If in the future my confession is proved to be wrong and dishonest resulting in the violence of legal regulation of the papers and intellectual property rights, then I am willing to return my degree I received to be withdrawn by Universitas Islam Indonesia.

Yogyakarta, December 2021



Jeffri Surya Dharma

الجمعة الائمة الاندونيسية

THESIS PROPOSAL OF SUPERVISOR

**ASPECT BASED SENTIMENT ANALYSIS FOR EXTRACTING KANSEI
WORD USING SPACY LIBRARY
(A CASE STUDY ON SMARTPHONE PRODUCT)**

THESIS

Arranged by:

Name : Jeffri Surya Dharma

Student Number : 15522184

Yogyakarta, August 2021

Supervisor,



Muhammad Ridwan Andi Purnomo, ST., M.Sc., Ph.D.

THESIS APPROVAL OF EXAMINATION COMMITTEE

**ASPECT BASED SENTIMENT ANALYSIS FOR EXTRACTING KANSEI
WORD USING SPACY LIBRARY
(A CASE STUDY ON SMARTPHONE PRODUCT)**

By:

Name : Jeffri Surya Dharma

Student No. : 15522184

Was defended before Examination Committee in Partial Fulfillment of the requirements
for the degree of Industrial Engineering Department

Universitas Islam Indonesia

Examination Committee

Muhammad Ridwan Andi Purnomo S.T., M.Sc., Ph.D.

Examination Committee Chair

Ir. Winda Nur Cahyo, S.T., M.T., Ph.D.

Member I

Arif Bramantoro, M.IT., Ph.D.

Member II

Acknowledged by,
Head of Department

Industrial Engineering – International
Program Universitas Islam Indonesia



(Dr. Taufiq Imawan, S.T., M.M)

PREFACE

Assalamualaikum Warahmatullahi Wabarakatuh

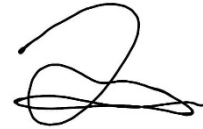
Alhamdulillahirabbil' alamin, praise to Allah Subhanahu Wata'ala for the strength, grace, and guidance, to help the Writer in completing this undergraduate thesis entitled "Aspect Based Sentiment Analysis for Extracting Kansei Word Using Spacy Library (A Case Study on Smartphone Product)". During the thesis project and the process of writing this report, the writer can't finish all the work if there is no help from Allah SWT and all the people that support the writer. On this occasion, the author would like to appreciate and thank to all the parties below:

1. My beloved family members who always give prayers, support, and encourage the Author during the completion of the Undergraduate Thesis.
2. Mr. Muhammad Ridwan Andi Purnomo, ST., M.Sc., Ph.D. as the supervisor who always guides and provides knowledge to assist Author in completing this Undergraduate Thesis.
3. Dean of the Faculty of Industrial Technology, Universitas Islam Indonesia.
4. Head of Undergraduate Program Department of Industrial Engineering, Universitas Islam Indonesia.
5. All lecturers of the Department of Industrial Engineering, Universitas Islam Indonesia who have given knowledge.
6. My friends of International Program Industrial Engineering batch 2015, my senior and my junior for the support, spirit, and enthusiasm for the Author.
7. All parties who cannot be mentioned one by one by Author for the assistance in completing this Undergraduate Thesis.

The author realizes that this undergraduate thesis is still not perfect and still has some weaknesses so the Author expects any criticism and suggestions from readers for the perfection of this report. Hopefully, this report and information included will be useful for the Author and give benefit to other parties who read this.

Wassalamu'alaikum Warahmatullahi Wabarakatuh

Yogyakarta, September 2021



(Jeffri Surya Dharma)

15522184



ABSTRACT

Conventional Kansei engineering is has challenging to conduct. Mostly, Kansei engineering is conducted by employing a questionnaire, survey, or interview. That method needs lot of time to be done and from the previous research is only a few data was acquired. Currently, data availability on the internet such as product reviews is easily found on the online marketplace. Product review from customers has drawn special attention from the product owner. Product review can convince the customer to buy the product or find other products. One of the platforms that provide the review accessibility is Amazon. Amazon can provide a lot of product reviews generated from the user that buy products from Amazon. One of the products is the Samsung Galaxy S9, which is reviewed a lot by the user. This paper discussed aspect-based sentiment analysis for Kansei engineering. The main advantage of this method, it can process larger data in a short time with the help of a programming tool. It can be used to analyze the product review that eventually can be used by the product owner to identify what the customer says about their product. Those reviews are gathered using scrapper and analyzed using SPACY library that employs machine learning to do the analysis.

Keyword: Customer reviews, Amazon customer reviews, Samsung S9, Machine learning, Spacy library, Kansei engineering,

TABLE OF CONTENT

AUTHENTICITY STATEMENT	II
THESIS PROPOSAL OF SUPERVISOR.....	III
THESIS APPROVAL OF EXAMINATION COMMITTEE	IV
PREFACE	V
ABSTRACT	VII
TABLE OF CONTENT.....	VIII
LIST OF TABLE.....	X
LIST OF FIGURE.....	XI
CHAPTER I INTRODUCTION	- 1 -
1.1 Background	- 1 -
1.2 Problem Formulation	- 4 -
1.3 Research Objectives	- 4 -
1.4 Benefits of Research	- 4 -
1.5 Limitations.....	- 5 -
CHAPTER II LITERATURE REVIEW	- 6 -
2.1 Literature Review	- 6 -
2.2 Inductive Study.....	- 6 -
2.3 Deductive Study	- 9 -
2.3.1 Samsung Galaxy S9.....	- 9 -
2.3.2 Web Scraping	- 10 -
2.3.3 Data Mining.....	- 10 -
2.3.4 Text Mining	- 13 -
2.3.5 Sentiment Analysis	- 14 -
2.3.6 Machine Learning.....	- 14 -
2.3.7 Classification	- 16 -
2.3.8 Word Embedding	- 17 -
2.3.9 Kansei Engineering	- 17 -
CHAPTER III RESEARCH METHODOLOGY.....	- 22 -
3.1 Research Object	- 22 -
3.2 Literature Review	- 22 -
3.3 Data Collection	- 22 -
3.4 Data Processing.....	- 22 -

3.5	Discussion.....	- 23 -
3.6	Conclusion and Recommendation	- 23 -
3.7	Research Flow Chart	- 23 -
CHAPTER IV DATA COLLECTION AND PROCESSING		- 25 -
4.1	Framework Development	- 25 -
4.3	Data Processing.....	- 36 -
4.3.1	Pre-Processing.....	- 37 -
4.3.2	Dependency Parsing.....	- 38 -
4.3.3	Aspect Clustering	- 43 -
4.4	Data Visualization.....	- 48 -
Chapter V RESULTS AND DISCUSSION		- 52 -
5.1	Kansei words identification.....	- 52 -
5.1.1	Screen Feature	- 56 -
5.1.2	Camera Feature.....	- 58 -
5.1.3	Charger Feature	- 60 -
5.1.4	Battery Feature	- 62 -
5.1.5	Speaker Feature	- 64 -
5.2	Product Improvement Guideline	- 66 -
CHAPTER VI Conclusion and Suggestion		- 67 -
6.1	Conclusion.....	- 67 -
6.2	Suggestion.....	- 68 -
REFERENCES.....		- 69 -

LIST OF TABLE

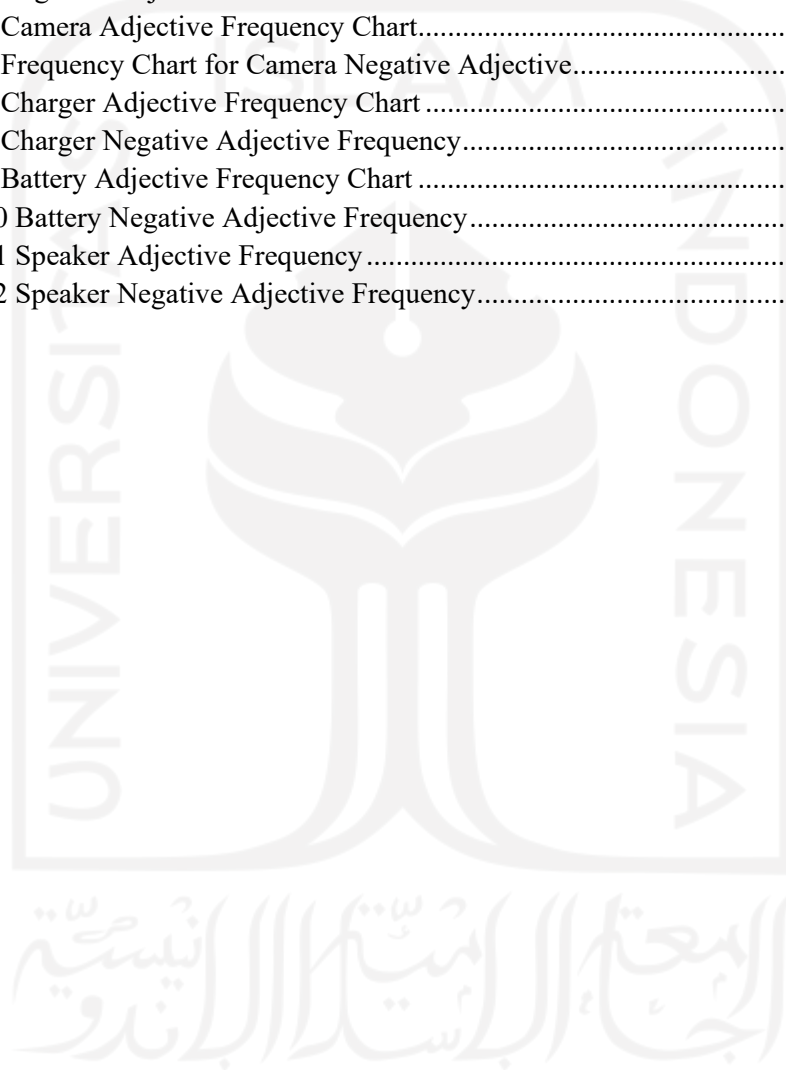
Tabel 5. 1 Table Aspect Vector Gupta Research, adapted from (Gupta et al., 2019)..... - 55 -
Tabel 5. 2 Product Improvement Guidelines**Error! Bookmark not defined.**



LIST OF FIGURE

Figure 2. 1 Classified diagram of three subsets, originally from (Raschka & Mirjalili, 2017).	- 16 -
Figure 2. 2 Kansei engineering model, adapted from (Schütte et al., 2004)	- 18 -
Figure 2. 3 Spanning the semantic space and space of properties breakdown into three steps, adapted from (Schütte et al., 2004)	- 19 -
Figure 3. 1 Research Flowchart	- 24 -
Figure 4. 1 ASP Kansei Engineering	- 26 -
Figure 4. 2 Amazon Customer Review Website	- 28 -
Figure 4. 3 Scrapping bot flow process	- 29 -
Figure 4. 4 Data-hook for Scrapping from Website	- 30 -
Figure 4. 5 The elements of review body on the website	- 31 -
Figure 4. 6 The elements of review rating on the website	- 31 -
Figure 4. 7 The elements of review title on the website	- 31 -
Figure 4. 8 The elements of review author in website	- 31 -
Figure 4. 9 The elements of review date on the website	- 32 -
Figure 4. 10 The elements of review helpful on the website	- 32 -
Figure 4. 11 Function Define URL with Asin and Run Bs4	- 33 -
Figure 4. 12 Web page navigation	- 34 -
Figure 4. 13 Naming the data-frame & Store in Json	- 34 -
Figure 4. 14 Function to bypass Amazon security check	- 35 -
Figure 4. 15 Scrapped data result	- 35 -
Figure 4. 16 Data Processing Flowchart	- 36 -
Figure 4. 17 Clean data function	- 37 -
Figure 4. 18 Noun Adjective Visualizer	- 38 -
Figure 4. 19 Rule 1	- 39 -
Figure 4. 20 Rule 2	- 40 -
Figure 4. 21 Rule 3	- 41 -
Figure 4. 22 Rule 4	- 41 -
Figure 4. 23 Rule 5	- 42 -
Figure 4. 24 Rule 6	- 42 -
Figure 4. 25 Store Json Data	- 43 -
Figure 4. 26 Main Clustering Function	- 43 -
Figure 4. 27 Update review data with cluster	- 44 -
Figure 4. 28 Get Aspect Function	- 44 -
Figure 4. 29 Get Aspect Frequency	- 45 -
Figure 4. 30 Get Word Vector	- 45 -
Figure 4. 31 Get Word Clusters	- 46 -
Figure 4. 32 Get Cluster Names Map	- 46 -
Figure 4. 33 Add Cluster to Review	- 47 -
Figure 4. 34 Convert JSON to Dataframe	- 48 -
Figure 4. 35 Get Wordcloud Function	- 49 -

Figure 4. 36 Wordcloud Result	- 49 -
Figure 4. 37 Get Top Words Function	- 50 -
Figure 4. 38 Draw Chart Function	- 50 -
Figure 4. 39 Frequency Chart.....	- 50 -
Figure 5. 1 Most Frequent Word in Cluster	- 53 -
Figure 5. 2 Cluster Noun Frequency	- 54 -
Figure 5. 3 Most Frequent Adjective in Screen Cluster	- 56 -
Figure 5. 4 Negative Adjective in Screen Feature	- 57 -
Figure 5. 5 Camera Adjective Frequency Chart.....	- 58 -
Figure 5. 6 Frequency Chart for Camera Negative Adjective.....	- 59 -
Figure 5. 7 Charger Adjective Frequency Chart	- 60 -
Figure 5. 8 Charger Negative Adjective Frequency.....	- 61 -
Figure 5. 9 Battery Adjective Frequency Chart	- 62 -
Figure 5. 10 Battery Negative Adjective Frequency.....	- 63 -
Figure 5. 11 Speaker Adjective Frequency	- 64 -
Figure 5. 12 Speaker Negative Adjective Frequency.....	- 65 -



CHAPTER I

INTRODUCTION

1.1 Background

The smartphone industry is one of the industrial sectors, which has rapid development in terms of technological discoveries. This rapid change has led the industrial company to face a great challenge and a competitive market. First, the need of the customer is following trends under the provision of advancement. The nature of human beings will always need to simplify their burden or job. Thus, this has become a strong motive for them to demand something better. Second, the nature of the company is to own recognition. This can be streamed by launching the new feature and promoting it to its market area. Having this as a fact that the company releases it first before rivalry will note the company in the public eyes to be a good existence. Taking one of the recent examples, 10x lossless zoom camera, is owned by OPPO smartphone by Blain (2019). This improvement has led many smartphone industries to follow the improvement alike, such as Huawei, Samsung, and Xiaomi. However, these all products have finally owned their reputation by simply obeying rapid change.

Samsung is one of the smartphone manufacturers with the largest market share in the world. Based on StatCounter, the market share of Samsung mobile phones is 32 %, followed by Apple at 22%. However, when Samsung released their flagship smartphone galaxy s9 in February 2018, based on the report from Counterpoint Research Market Monitor, they reported that Samsung net profit decreased by 11%. The sluggishness of Galaxy S9 sales is the leading cause. Shipping the Galaxy S9 only reached 31 million units, making it the longest selling since the Galaxy S3 which was released in 2012. According to Counterpoint's new weekly US smartphone model tracker, the Galaxy S10 series sold more than 16% in the first week of its sales compared to last year's sales of the Galaxy S9 series. This is a comparison of sales in the US between March 8th to 14th in 2019 with March 16th to 22nd in 2018. From those data, the researcher is curious about what happened to the Galaxy S9 series. One of the ways to evaluate a product is from what the user says about the product.

In this era of online shopping, the user can see online reviews anywhere, such as Youtube, e-commerce, Instagram, Twitter, and Facebook. Presently, consumers have a strong bond with social media. Thus, in making an online purchase decision, consumers rely more on information generated by other users in the social media and networks Thoumrungroje (2014). Online reviewers/consumers will not only write what they experienced about the product but also the user will write the pros and cons of the product as well. From those data/ online reviews, the researcher sees the opportunity to dig more in-depth about the product, because the text data from the user is compelling and can help a company to develop their product based on the needs of a customer. The company needs to give attention to customer feedback because consumers trust online reviews more than advertisements and rely on online reviews when choosing products by Nielsen (2012). One popular e-commerce site is amazon.com with more than 300 million users by Eugene (2017). Amazon provides a platform for the user to give their experience about the product that the customer buys, its social media for the user to share their information with other users. The researcher can gather that information by using a web scraping technique, then analyze the data to obtain valuable information about the product.

Text is classified as unstructured data. Unstructured text is easily processed and perceived by humans, but significantly harder for a machine to understand. This volume of text is an invaluable source of information and knowledge. For processing text data, the researcher needs to design methods and use an algorithm in order to process text data effectively by Mehdi (2017). Some text mining and opinion extraction applications such as text summarization, text classification, chatbot for customer service, product recommender based on user review. One interesting application field is sentiment analysis for the product based on the user review in e-commerce. Sentiment analysis can analyze the online review and classify the review as a positive or negative review, and this information can reveal the product's strengths and weaknesses. In the process of product design, the designer must know what the customer needs or the designer can see the strength and weaknesses of the past product, to improve the next product. Sentiment analysis is a technique for extracting information/features from the text by Ireland & Liu (2018).

The sentiment is a view, feeling, opinion, or assessment of a person for some product, event, or service by Damereau (2010). Sentiment analysis or opinion mining is a combination of text mining and natural language processing for extraction,

classification, and summarization of sentiment and emotions expressed in the text by Damereau (2010). Sentiment analysis can replace conventional ways such as web-based surveys to gather public opinion about products or services. Feature-based sentiment analysis includes feature extraction, sentiment prediction, and sentiment classification by Hu (2004). Feature extraction is a process to identify product aspects that are commented on by the customer on an online review. In the product development process, start with finding what features are needed by the customer in the analyzed product. In order to create a good product, engineers and designers need much information from the customer's voice. Sentiment analysis can be used in this scenario. Sentiment analysis can reveal the product feature that has a positive or negative sentiment. For the product feature with negative sentiment, the designer needs to improve the quality of the product for the next product. There are a lot of techniques for sentiment analysis, such as topic modeling and feature extraction, and can be classified by lexical method or machine learning method. A lexical method is based on the insight that the polarity of a piece of text can be obtained on the ground of the polarity of the words which compose it by Cataldo (2014). The machine learning method is more complicated than the lexical method; one of the machine learning methods is convolutional neural networks. Convolutional neural networks are supervised learning.

One library that can be used to explore and analyze text is the Spacy library. Spacy is advanced natural language processing based on python. The features in space are quite complete such as Tokenization, Part-of-Speech detection, Dependency parsing, Lemmatization, Named Entity Recognition, and text classification. Spacy is built on CNN that is trained on the text that is written on the web. There are several options based on the size of the trained pipeline such as sm, md, lg, and trf. In this research, the lg one is used since it is enough for this research with a large word vector table of around 500k entries.

Kansei engineering is a product development method used to investigate human feelings and discover quantitative relationships between the affective responses and design features by Nagamachi (1989). By using Kansei engineering, much research has been done to improve product and service design. The data collection method of the existing studies is similar, by collecting Kansei words manually from various sources, such as customer interviews, expert interviews, articles, and more. In this research, the

researcher tries to use natural language processing and machine learning algorithms to extract Kansei words from customer reviews.

Traditional survey-based methods produce high-quality sentiment data, which has been widely used in numerous sentiment design research. Early Kansei knowledge extraction relied heavily on complicated procedures including surveys, interviews paired with oral protocol analysis, semantic discrimination methods, conceptual sketching, and picture scaling methods. The majority of existing research, however, is conducted on a limited scale because it relies on users' active participation in the study. Based on this problem, this paper conducts Kansei engineering with the help of aspect-based sentiment analysis, in order to speed up the process, and analyze more data in a shorter time. This paper also aimed to create a new framework for doing Kansei engineering with aspect-based sentiment analysis.

1.2 Problem Formulation

Based on the problem in the background, the formulation of the problem to be addressed in this study are: how is the procedure of aspect-based sentiment analysis for Kansei Engineering?

1.3 Research Objectives

The objective of this research is to identify and formulate the procedure of Kansei engineering with aspect-based sentiment analysis.

1.4 Benefits of Research

The expected benefits of this research to some other involved parties are:

1. For the Researcher

The researcher can gain more knowledge about python programming and machine learning and data scientists in the industry.

2. For Manufacturer

The result of this research can be used as decision support for the manufacturers to product's further development.

1.5 Limitations

The researcher has to set several limitations to avoid widespread discussion and understanding about this concept of research, which are:

1. The data for this research is the customer reviews on a product, namely Samsung Galaxy S9+, data are derived from Amazon.com. The researcher filters only verified purchases and only reviews for the data, and only US reviews can be gathered.
2. This research involves the implementation of the spacy framework for extracting Kansei words.
3. This research only formulates the schematic process of Kansei engineering using aspect-based sentiment analysis. For the experiment, this research is limited to the interpretation of Kansei word.
4. The research ignores data about price, username for privacy, and smartphone color. For this research, aspect pairs extracted from review text, and sentiment analysis will be using aspect modifiers to pass into Vader Sentiment.

CHAPTER II

LITERATURE REVIEW

2.1 Literature Review

The literature study explains the foundation of the theory used in conducting the research. The foundations of the theory in this study include Amazon, Web-scraping, Python, web Online Customer review, online reviews, Data Processing, sentiment analysis, Kansei engineering. There will also be inductive studies on previous research studies that have been conducted and similar with this research.

2.2 Inductive Study

There are many previous research about Sentiment analysis. (L. Zhang, Hua, Wang, Qian, & Zheng, 2014) conducted a study titled “Sentiment analysis on Reviews of Mobile Users”. The data for the research is a mobile user review for WeChat from the iTunes store for Apple devices. The researcher uses SVM LibLinear as the tool for analysing the data. From the research conclusion, the Bayesian method is more accurate rather than SVM method for classifying the review, second, longer review text is more difficult for the model to classify. Beside, the shorter review text gives more accuracy for the classification.

The research titled Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics by Mohammad Salehan, Dan J. Kim (2015) is about the effect of review sentiment on readership and helpfulness of online reviews. This paper stated that online review can be useful for vendors and customers. For consumers, a useful online review can provide them with more valuable information and save time and energy for searching the product they needed. For vendors, the online review can satisfy their customer's need for information that can persuade the customer to decide to buy the product quickly and lead to increase sales of vendors.

The research by (Sammons et al., 2016) titled Feature Extraction for NLP, simplified. The research conduct feature extraction from text using Edison. Edison is a

java library for feature extraction in natural language processing. Edison feature extractors can be easily integrated with learning frameworks such as LB Java, Mallet, Weka. Feature extraction in the context of the machine learning application. Machine learning algorithms are used to generate decision functions that would be hard to implement programmatically. Feature extraction is done by building statistical models that map some representation of the input data to some predefined set of meaningful outputs.

The research by (Bello Garcés, 2018) titled Implementasi SVM dan Asosiasi untuk sentiment analysis data ulasan the phoenix hotel Yogyakarta pada situs tripadvisor. The research conducts experiments by using svm for classifying user reviews into positive or negative review, and conducting associations to see the correlation between words in the review, this process can reveal what customers said in the review. The research also tried to solve the problem from the result of the sentiment using a fishbone diagram to see the root cause of the problem. From the experiment conducted by the researcher, SVM accuracy is 84%. The data for testing cover 302 data, yet only 296 reviews are accurately classified by SVM model, and the rest is classified as wrong by SVM model.

The research by (Tan, Wang, & Xu, n.d.) titled Sentiment Analysis for Amazon Reviews. The researcher conducts sentiment analysis on Amazon review using various machine learning algorithms. For feature extraction, the researcher conducts two methods. For the traditional method, the researcher said that the traditional method produces lower accuracy than the machine learning algorithm. For the machine learning algorithm, the researcher conducts a 50-d glove dictionary which was pretrained on Wikipedia. 50-d glove dictionary has higher accuracy than the traditional method, from the experiment, LSTM without Glove give an accuracy rate of 73% on training and 71% on testing, and for LSTM with Glove has the accuracy rate of 85% on training and 65% on test phase, its means the model of LSTM with glove is overfitted.

The research titled Recommendation system in E-Commerce using sentiment analysis by Lydia Priyadharsini, Lovelin Ponn (2017). The research suggested implementation of a product recommender system based on hybrid recommendation system sentiment analysis on online review at amazon. The research also conducts fake review monitoring by filtering the MAC address of user devices, so the user only can post one review, if the user posts another comment, the system automatically blocks the

comment. The model is developed using C#.NET framework as front end design and SQL server as the back end. the research is done by using 2 algorithms: collaborative filtering and stochastic learning, from the performance metric, the researcher measures recall and precision. For collaborative filtering, precision is resulted as 77,5% and recall 35%. For stochastic learning, precision is calculated as 80% and recall 37,5%.

The research titled sentiment analysis of online reviews using bag-of-words and lstm approaches by (Barry, 2017). The research explained the comparison between bag-of-words and neural network-based approach for sentiment classification. The researcher does a comparison between support vector machine, multinomial naive bayes, and long short-term memory. From the result, svm models outperform the multinomial naïve bayes classifiers, the lstm model generates the highest performance than both algorithms. The LSTM models uses pre-trained GloVe embeddings and Word2vec embeddings. The LSTM models with GloVe embeddings provide the best performance with the accuracy of 95.84% and AUC .9832.

In Guo et al. (2009), a method called multilevel latent semantic association was presented. At the first level, all the words in aspect expressions (each aspect expression can have more than one word) are grouped into a set of concepts/topics using LDA. The results are used to build latent topic structures for aspect expressions. For example, we have four aspect expressions “day photos”, “day photo”, “daytime photos” and “daytime photo”. If LDA groups the individual words “day” and “daytime” into topic10, and “photo” and “photos” into topic12, the system will group all four aspect expressions into one group, call it “topic10-topic12”, which is called a latent topic structure. At the second level, aspect expressions are grouped by LDA again but according to their latent topic structures produced at level 1 and their context snippets in reviews. Following the above example, “day photos”, “day photo”, “daytime photos” and “daytime photo” in “topic10-topic12” combined with their surrounding words form a document. LDA runs on such documents to produce the result.

In Zhang and Liu (2011), is purposed a method to identify nouns and noun phrases that are aspects and imply sentiments in a particular domain. These nouns and phrases alone indicate no sentiments, but in the domain context, they may represent desirable of undesirable facts. For example, “valley” and “mountain” do not have any sentiment connotation in general, they are objective. However, in the domain of mattress reviews.

They often imply negative opinions as in “within a month, a valley has formed in the middle of the matters.” In that case, “valley “implies a negative sentiment on the aspect of matters quality. Identifying the sentiment orientations of such aspects is very challenging but critical for effective sentiment analysis in these domains.

In the paper titled Aspect-based sentiment analysis of mobile reviews by Vedika and viveka, (2019), the researcher conducts an aspect-based sentiment analysis experiment. The researcher uses 3 different datasets of mobile phone reviews that crawled from amazon.com. the researcher in the paper explains also about aspect vector. Aspect vector contains 2 entities which are aspect category and aspect terms, the data for the aspect vector is obtained from GSMarena. The name of identified aspect categories is network, body, display, performance, memory, camera, sound, communication, feature, battery, overall, and accessories. The researcher, in this paper for their model, uses 2 algorithms, Rouge-L algorithm for evaluating the summaries, and LexRank for ranking sentences.

In a paper titled extracting and summarizing affective feature and responses from online product description and reviews: A Kansei text mining approach by Wang, Li, Tian, and Cheng (2018). The researcher tries to use a text mining approach to extract and summarize useful information from Amazon user reviews. This research can help consumers for making a better purchase decisions while, product designers can use this research for improving their products and strategies. In comparison to this work, the major purpose is to establish a framework for conducting Kansei engineering using sentiment analysis. With the use of the spacy library, this paper attempts to extract aspect-modifier pairs from nouns, adjectives, adverbs, and verbs.

2.3 Deductive Study

2.3.1 Samsung Galaxy S9

The Samsung Galaxy S9 features a metal and glass design, with tapering edges on both sides. It has a 5.8-inch Quad HD+ Super AMOLED display with an 18.5:9 aspect ratio. Touch response and colours are superb, and it also supports HDR. In Indonesia, the phone uses an Exynos 9810 octa-core SoC and comes with 4GB of RAM a choice of 64GB or 256GB internal storage, which is expandable. The stereo speakers deliver good sound

quality and there's wireless charging here too. The main highlight is the camera, which features a variable aperture, super slow-motion videos up to 960fps. and AR Emojis. The camera performance is good too, especially in low light. The 3000mAh battery supports fast charging and should stretch an entire day on a single charge.

2.3.2 Web Scraping

Web scraping is a process involving the retrieval of a semi-structured document from the internet, generally a web page in a markup language such as HTML or XHTML, and analysis of that document in order to extract specific data from it for use in another context (Turland, 2010).

Web scraping has several steps, including:

- 1) Create Scraping Template: Programmer survey HTML document from the target website.
- 2) Explore Site Navigation: The programmer surveys the site navigation in the target website, then implements the site navigation technique in programming applications.
- 3) Automate Navigation and Extraction: Programmer must implement a technique to create an application that automatically navigates through the website page and extract the targeted information.
- 4) Extracted Data and Package History: Information/data that is already gathered in step 3 must be stored in the database (pandas) or saved into csv or txt file.

2.3.3 Data Mining

In simple terms, data mining is mining or finding new information by looking for certain patterns or rules from a very large amount of data (Davies, 2004). Data mining is also referred to as a series of processes to explore the added value of knowledge that has not been known manually from a data set (Pramudiono, 2007). Data mining is often also referred to as knowledge discovery in databases. Knowledge discovery in the database is an activity that includes the collection, use of data, historical to find regularities, patterns, or relationships in large data sets (Santosa, 2007).

Data mining is the activity of finding interesting patterns of large amounts of data, data can be stored in a database, data warehouse, or other information storage. Data mining relates to other fields of science, such as database systems, data warehousing, statistics, machine learning, information retrieval, and high-level computing. In addition, data mining is supported by other sciences such as neural networks, pattern recognition, spatial data analysis, image databases, signal processing (Han, 2006). Data mining is defined as the process of finding patterns in data. This process is automatic or often semi-automatic. The pattern found must be meaningful and the pattern provides benefits, usually economic benefits. Data needed in large numbers (Witten, 2005). There are 7 stages in data mining stages, namely:

1. Data Cleaning

Data sanitization is the process of eliminating noise and inconsistent data or irrelevant data. In general, data obtained from both the company's database and experiment results have imperfect stuffing such as lost data, invalid data, or just a typo. In addition, there are also data attributes that are irrelevant to the hypothesis of data mining owned. Irrelevant data is also better discarded. Data sanitization will also affect the formation of data mining techniques because the data handled will be reduced in number and complexity.

2. Data integration (Data Integration)

Data integration is merging data from various databases into a single new database. It is not uncommon for data mining to not only come from a single database but also from multiple databases or text files. Data integration is done on attributes that identify unique entities such as name, product type, customer number, and other attributes. Data integration needs to be done carefully because errors in data integration can produce deviant results and even misleading action later. For example, if the data integration by product type turns out to be combining products from different categories it will be obtained correlation between products that do not exist.

3. Data selection

The data in the database is often not all used, therefore only the appropriate data to be analysed will be retrieved from the database. For example, in a case that examines the trend factor of people buying in the case of basketball analysis market, no need to take the name of the customer, enough with the customer ID only.

4. Data transformation

Data is altered or merged into the appropriate format for processing in data mining. Some data mining methods require a specific data format before it can be applied. For example, some standard methods such as association and clustering analysis can only accept categorical data inputs. Therefore, the data in the form of numeric numbers continuing needs to be divided into several intervals. This process is often called data transformation.

5. Mining Process

It is a major process when the method is applied to find valuable and hidden knowledge of the data.

6. Pattern Evaluation

To identify interesting patterns into knowledge-based found. In this stage, the results of data mining techniques in the form of distinctive patterns and predictive models are evaluated to assess whether the existing hypothesis is indeed achieved. If the results are obtained not according to the hypothesis some alternatives can be taken such as making it feedback to improve the data mining process, trying other data mining methods that are more appropriate, or receiving these results as expected result that may be beneficial.

7. Knowledge Presentation

It is the visualization and presentation of knowledge about the methods used to acquire the knowledge gained by users. The last phase of the data mining process is how to formulate decisions or actions from the results of the analysis. Sometimes it should involve people who do not understand data mining.

Therefore, the presentation of data mining results in the form of knowledge that everyone can understand is a necessary step in the process of data mining. In this presentation, visualizations can also help communicate the results of data Mining (Han, 2006).

2.3.4 Text Mining

The initial stage in text mining is information retrieval. It is the process of looking for and retrieving data from a large volume of data. There are certain principles for retrieving information from text, as well as what degree of retrieval should be regarded. This is the process of locating and returning useful data. Text classification is a type of text mining. It is the process of categorizing documents into preset groups. Documents should be categorised according to their purpose. The process of text mining includes information extraction. Identification of specified entities in the text, extraction of entities, and representation of entities in the appropriate format are all part of information extraction. The search for structured data inside documents is an important job of text mining (Anil Zende et al., 2016).

In today's world, we are frequently confronted with activities that are difficult or impossible to do without the use of sophisticated technologies. Text mining has become an important commercial technique for uncovering hidden data trends in recent years. Text mining is used in a variety of industries to evaluate enormous data volumes. In the publishing and media industries, text mining is utilized for the creation and optimization of information retrieval, which is accomplished by extracting, loading, and conveying data. CRM technology is used in banks and financial markets to improve customer communication management by resending messages using a search engine that asks questions in natural language. For categorization and extraction of information from papers, scholarly abstracts, and patents in hospitals and pharmaceutical firms (Anil Zende et al., 2016).

2.3.5 Sentiment Analysis

Sentiment analysis or a commonly called opinion mining is one of Text Mining's research branches. Opinion Mining is computational research of opinion, sentiment, and emotion expressed by text. If given a set of text documents that contain opinions on an object, the opinion mining aims to extract the attributes and components of the object that have been commented on each document and to determine whether the comment is meaningful Positive or negative (Shelby, 2013). Sentiment Analysis can be distinguished based on the data source, some of the levels commonly used in Sentiment Analysis research are Sentiment Analysis at the document level and Sentiment Analysis at the sentence level. Based on the data source level Sentiment Analysis is divided into 2 large groups (Clayton, 2011): Coarse-grained Sentiment Analysis and fined-grained Sentiment Analysis.

In Sentiment Analysis Coarse-grained, the Sentiment Analysis is at the document level. Broadly, the main goal of Sentiment Analysis is to consider the whole document as a positive Sentiment or negative Sentiment. Fined-grained Sentiment Analysis is a Sentiment Analysis at the sentence level. The main goal of fined-grained Sentiment Analysis is to determine the sentiment on each sentence.

2.3.6 Machine Learning

Machine Learning (ML) is an approach in Artificial Intelligence (AI) that is widely used to replace or impersonate human behavior to solve problems or automate. As the name suggests, ML attempts to mimic how human processes or intelligent beings learn and regenerate (Tanaka & Okutomi, 2014).

Purnamasari (2013) in his book explained that Machine Learning is a branch of artificial intelligence, is a discipline that includes the design and development of algorithms that enable the computer to develop behaviors that Based on empirical data, such as from data sensors on the database. The learning system can utilize the example (data) to capture the necessary features of the underlying probability (unknown). The Data can be viewed as an example describing the relationship between the variables observed. The great focus of Machine Learning research is how to automatically recognize complex patterns and make intelligent decisions based on data. His success

happened because of the set of all possible behaviors, of all possible inputs, too large to be covered by the set of observation examples (training data). Therefore, Machine Learning should be a complete (generalized) behavior of an existing instance to produce a useful output in new cases.

Machine learning techniques can roughly be divided into two large classes. Here are the classes:

1. Supervised Learning

Supervised learning algorithms are a class of machine learning algorithms that use previously labeled data to learn its features, so they can classify similar but unlabelled data (Ivan, 2019). This example of the use of supervised learning for understanding this concept better. Let's assume that a user receives many emails every day, some of which are important business emails and some of which are unsolicited junk emails, also known as spam.

A supervised machine algorithm will be presented with a large body of emails that have already been labeled by a teacher as spam or not spam (this is called training data). For each sample, the machine will try to predict whether the email is spam or not, and it will compare the prediction with the original target label. If the prediction differs from the target, the machine will adjust its internal parameters in such a way that the next time it encounters this sample it will classify it correctly. Conversely, if the prediction was correct, the parameters will stay the same. The more training data we feed to the algorithm, the better it becomes (Ivan, 2019).

2. Unsupervised Learning

The second class of machine learning algorithms is unsupervised learning. Here, we don't label the data beforehand, but instead, we let the algorithm come to its conclusion. One of the most common, and perhaps simplest, examples of unsupervised learning is clustering. This is a technique that attempts to separate the data into subsets (Ivan, 2019).

To illustrate this, let's view the spam-or-not-spam email classification as an unsupervised learning problem. In the supervised case, for each email, we had a set of features and a label (spam or not spam). Here, we'll use the same set of features, but the emails will not be labeled. Instead, we'll ask the algorithm, when given the set of features, to put each sample in one of two separate groups (or clusters). Then the algorithm will try to combine the samples in such a way that the intraclass similarity (which is the similarity between samples in the same cluster) is high and the similarity between different clusters is low. Different clustering algorithms use different metrics to measure similarity. For some more advanced algorithms, you don't have to specify the number of clusters (Ivan, 2019). In the following graph, we show how a set of points can be classified to form three subsets:

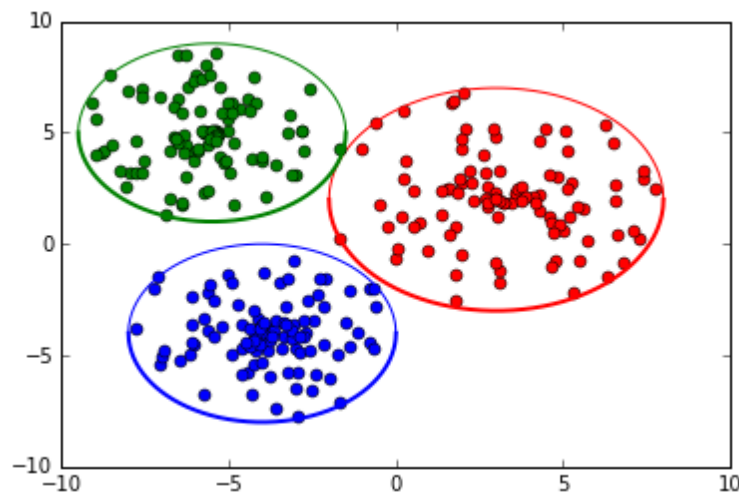


Figure 2. 1 Classified diagram of three subsets, originally from (Raschka & Mirjalili, 2017)

2.3.7 Classification

Classification is a process of finding a model or function that can explain or distinguish the concept or class of data. The purpose of the classification is to estimate the class of an object whose label is unknown. The classification process is usually divided into two phases namely the learning phase and the test phase.

In the learning phase, a part of the data class that has been known is fed to form an approximate model. Then in the test phase of the model that has been formed tested with some other data to know the accuracy of the model. If the accuracy is sufficient, this model can be used for predictive data classes that are not yet known. This technique can provide a classification on new data by manipulating existing data that has been classified and by using the results to calculate the distance between the template image features and the input image (Ulwan, 2016).

2.3.8 Word Embedding

According to Collobert R. et al (2011) in their journal about natural language processing from scratch, Word embedding is needed in many deep learning models in NLP cases. The word embedding results are used as an input feature in deep learning models. Word embedding is a technique for language modeling and feature learning, which transforms words in a vocabulary to vectors of continuous real numbers, for example, word “hat” converts to numerical vector (... , 0.15, ... , 0.23, ... , 0.41, ...). The technique normally involves a mathematic embedding from a high dimensional sparse vector space (one-hot encoding vectors space, in which each word takes a dimension) to a lower-dimensional dense vector space. Each dimension of the embedding vector represents a latent feature of a word. The vectors may encode linguistic regularities and patterns by Lei Zhang, Et al (2010).

2.3.9 Kansei Engineering

Kansei is a Japanese term that represents emotions and impressions. In Kansei engineering studies, surveys are always used to study the relationship between affective attributes and product design features by Llinares and Page (2011). The most commonly used method is the semantic differential (SD) method, which is a rating scale used to measure respondents' opinions and attitudes towards a given object by Osgood (1957). Researchers use the SD method to design questionnaires to measure subjective consumer impressions of the product by Yan (2008). The questionnaire consists of a list of words called Kansei attributes. Each Kansei attribute refers to a particular emotional expression

by Chou (2016). Each Kansei attribute consists of a bipolar pair of Kansei words (i.e. a positive word and a negative word, such as beautiful–ugly) by Friberg (2006).

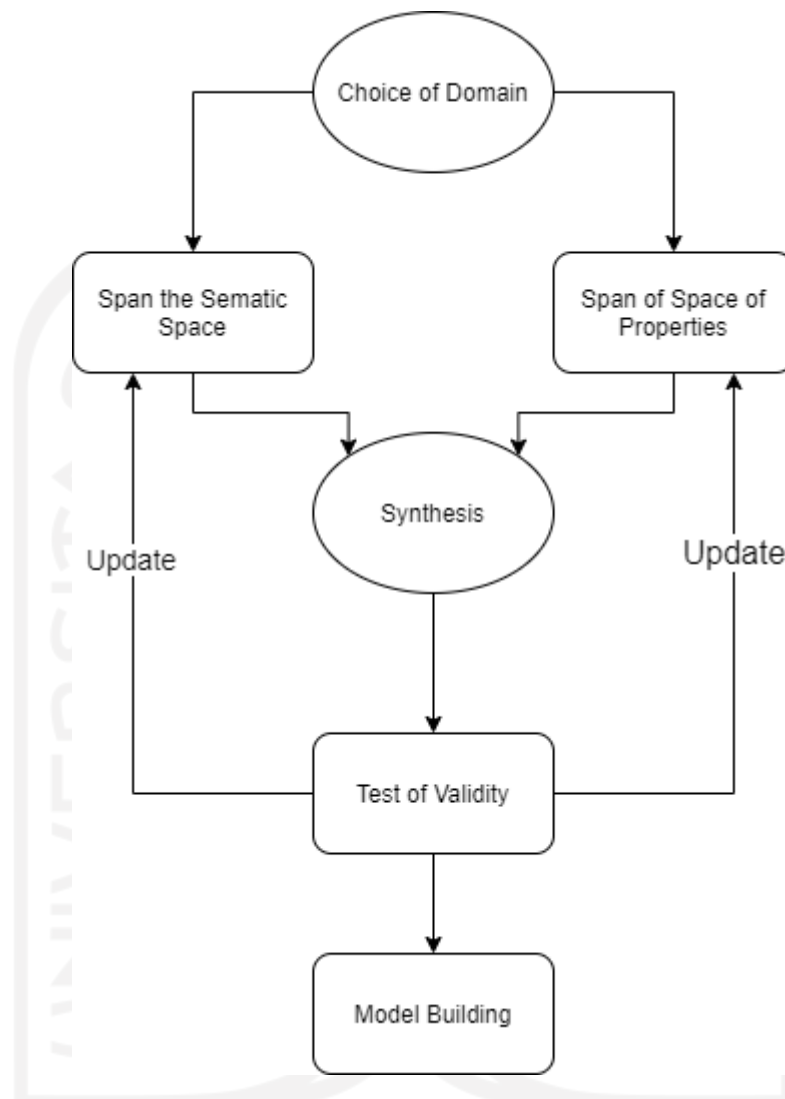


Figure 2. 2 Kansei engineering model, adapted from (Schütte et al., 2004)

Kansei Engineering is a comprehensive subject that combines design science, ergonomics, and engineering. To do this, designers must build a bridge between the user's psychological experience and product design decisions (Jiao & Qu, 2019). Kansei engineering is aimed to help product manufacturer to develop their product based on what user needs. However, some challenges faced in Kansei engineering as purposed by (Nagamachi, 2002) there are 2 points about how to accurately measure users' perception on products and how to combine user's emotions and product design. User's emotions on product are based on the product feature and the functional experience. It is an

information analysing process to convert users' emotional desire into the cognitive image of a product.

Figure 2.2 is purposed Kansei engineering framework by (Schütte et al., 2004). The framework has 2 branches of perspective for describing the product by semantic description and description of product properties. These 2 spaces will merge with each other in the synthesis phase. The framework starts from choosing a domain. Domain selection includes the selection of target groups and sales markets, as well as the specification of new products. Based on this information, product samples representing the domain are collected. A Kansei domain can be understood as the ideal idea for a particular product (Schütte et al., 2004).

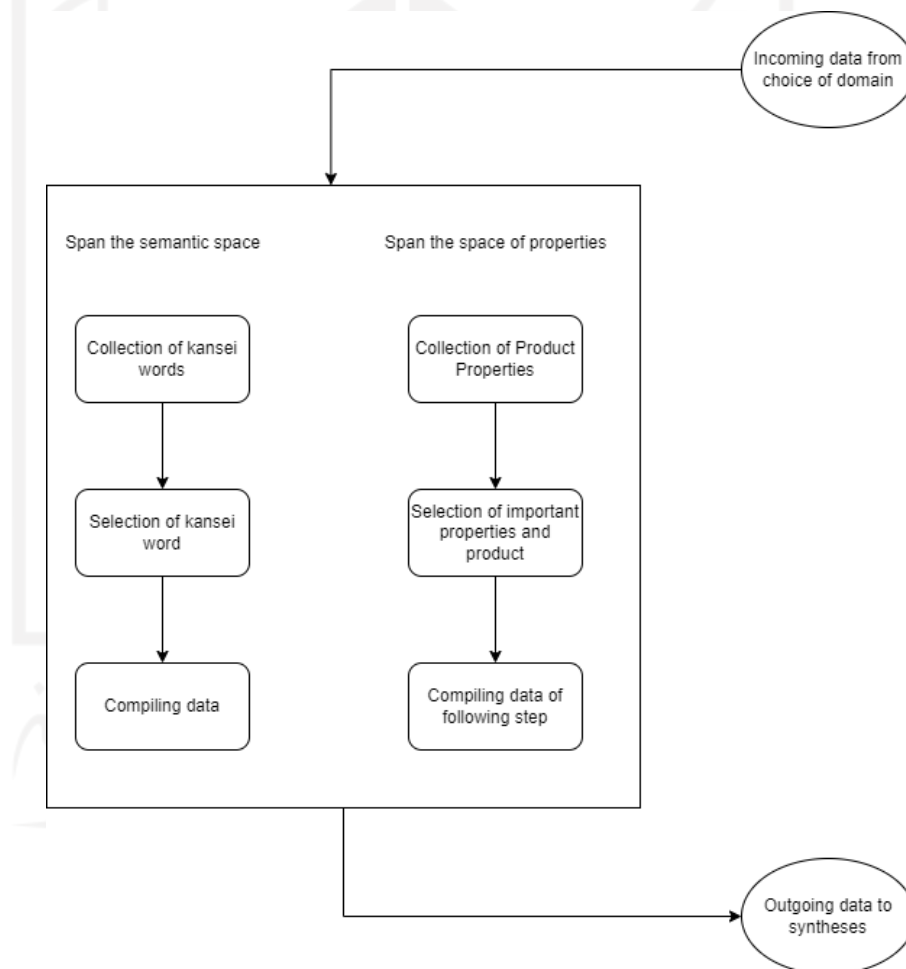


Figure 2. 3 Spanning the semantic space and space of properties breakdown into three steps, adapted from (Schütte et al., 2004)

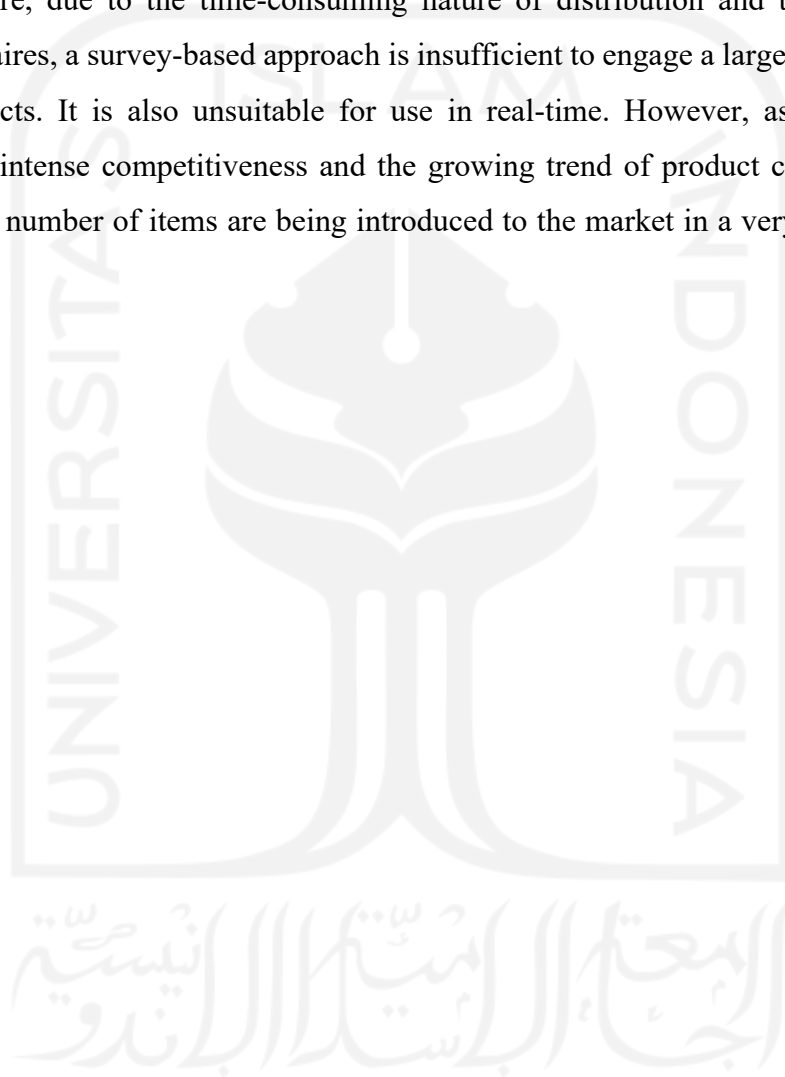
The second step in the Kansei framework is a process to collect Kansei word. There are 2 options to find Kansei word. The first is using span the semantic space. In

practice, the "Span Semantic Space" step is performed in three steps as shown in figure 2.3. Gathering the Kansei words that describe the product, starting with the desired domain. In the second stage, the word count is reduced to something more practical. The product's properties are similar to the semantic space. The steps in product properties can be seen in Figure 2.3. The step for this part is to collect all the attributes representing the selected domain, select the attributes that seem to have the greatest impact on the user's Kansei, and select the products that represent the selected product attributes before compiling the data for the next synthesis step.

Conventional survey-based approaches provide high-quality sentiment data that has been extensively used in many sentiments design studies. Early Kansei knowledge extraction was mainly based on psychological experiments collected using complex methods such as questionnaires, interviews combined with oral protocol analysis, semantic discrimination methods, conceptual sketching, and image scaling methods. However, most of the existing research is done on a relatively small scale because it depends on the fact that users will actively participate in the research. For example (Chou, 2016) 7 users participated in the Kansei for 10 product evaluations. (Jiang et al., 2015) recruited 4 users to rate 10 products (Guo et al., 2016) studied 36 people in 16 designs. Moreover, traditional survey questions are designed based on expert thinking rather than consumer's perspective (Hsiao et al., 2017). Respondents can only passively answer questions developed by experts. Also, respondents may not be consumers of the target product. Additionally, this method is time-consuming caused by distribution, and the collection of questionnaires makes a survey-based approach inadequate to engage too many users and products. Also, it is not suitable for real-time. However, due to the high competition in the industry and the trend of product customization, more and more products are coming to market in a very short time. Therefore, to effectively analyse consumers' feedback, it is necessary to develop with technology approach.

Traditional survey-based methods produce high-quality sentiment data, which has been widely used in numerous sentiment design research. Early Kansei knowledge extraction relied heavily on complicated procedures including surveys, interviews paired with oral protocol analysis, semantic discrimination methods, conceptual sketching, and picture scaling methods. The majority of existing research, however, is conducted on a limited scale because it relies on users' active participation in the study. For example,

seven users took part in the Kansei for ten product evaluations (Chou, 2016). (Jiang et al., 2015) enlisted the help of four people to appraise ten goods, whereas (Guo et al., 2016) looked at 36 people in 16 different designs. Furthermore, traditional survey questions are created based on expert opinion rather than the viewpoint of the consumer (Hsiao et al., 2017). Respondents can only answer questions designed by professionals in a passive manner. Furthermore, responders may or may not be users of the desired product. Furthermore, due to the time-consuming nature of distribution and the collecting of questionnaires, a survey-based approach is insufficient to engage a large number of users and products. It is also unsuitable for use in real-time. However, as a result of the industry's intense competitiveness and the growing trend of product customisation, an increasing number of items are being introduced to the market in a very short period of time.



CHAPTER III

RESEARCH METHODOLOGY

3.1 Research Object

The focus of this research will be on the development of the procedure to conduct Kansei engineering using aspect-based sentiment analysis. This research was conducted by using customer review data. The data were gathered by using a scraping bot to gather Amazon.com customer reviews. The specific product data is Samsung Galaxy S9.

3.2 Literature Review

The literature review is conducted in order to enhance the research. Literature review includes the previous studies that discussed topics are related to this research (inductive study), which includes machine learning method, sentiment analysis, and Kansei engineering. The deductive study is a theory that supports this research.

3.3 Data Collection

Data collection for this research is using primary data. The primary data are defined as the data obtained from the customer review on Amazon.com. In this chapter, the method to gather user review data from Amazon.com is using a scraping bot.

3.4 Data Processing

Data processing will briefly explain how to build aspect-based sentiment analysis using spacy library. Starts from data cleaning and pre-processing, until building the pipeline for spacy. There are software that used to perform data processing such as Python, and some python libraries such as Spacy, NLTK, Pandas, Bs4, JSON.

3.5 Discussion

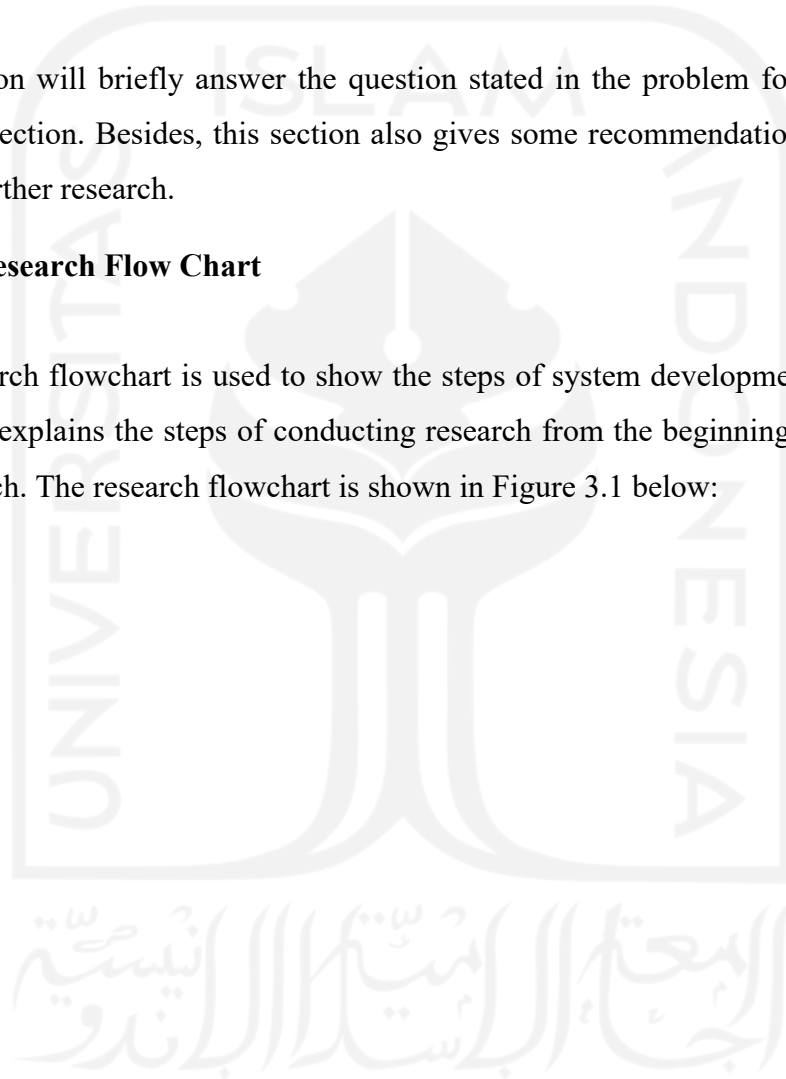
After the system is built, then discussion will be conducted to discuss the problem formulation and the result of data processing. The interpretation of outcome data is explained through discussion.

3.6 Conclusion and Recommendation

This section will briefly answer the question stated in the problem formulation in the previous section. Besides, this section also gives some recommendations that might be used to further research.

3.7 Research Flow Chart

This research flowchart is used to show the steps of system development. The research flowchart explains the steps of conducting research from the beginning until the end of the research. The research flowchart is shown in Figure 3.1 below:



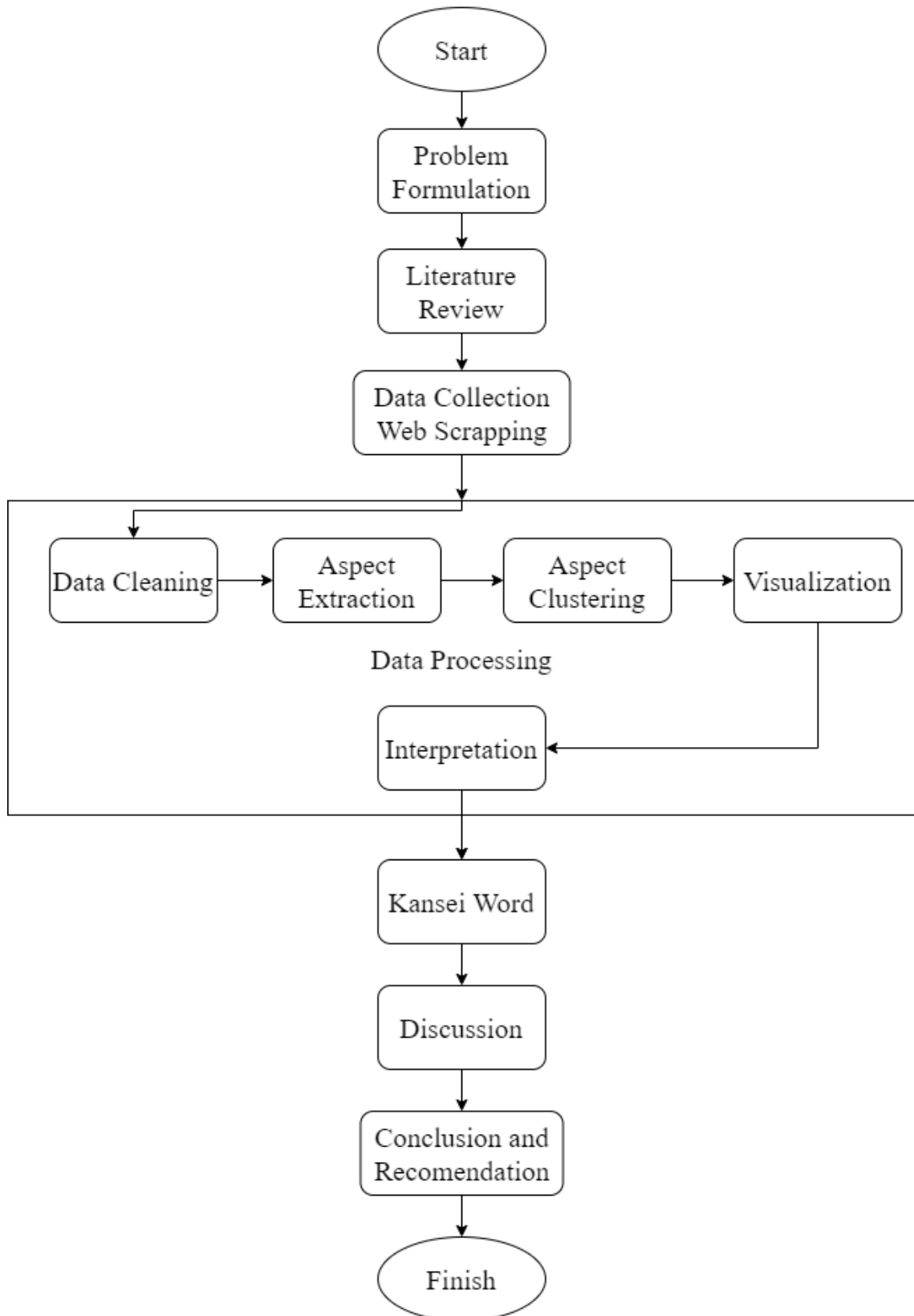


Figure 3. 1 Research Flowchart

CHAPTER IV

DATA COLLECTION AND PROCESSING

4.1 Framework Development

There are numerous differences between conventional Kansei engineering and Kansei engineering using aspect-based sentiment analysis. In conventional Kansei engineering, normally the data to create Kansei guidelines is typically obtained through a survey or questionnaire. The disadvantages of conventional Kansei engineering, it takes longer time to collect the data, and only a small scale data is obtained. Compared to Kansei engineering using aspect-based sentiment analysis, the possibility of using online review is possible. Nowadays, user-generated content on internet is already abundant. It is easy to find the online reviews in the online marketplace as in social media. It provides the opportunity to collect a larger amount of data on the product and the customer experience. The challenge when using conventional Kansei engineering is data and time. Motivated by the problem occurred in above, this paper proposes framework to conduct Kansei engineering using aspect-based sentiment analysis. As we can see below in figure 4.1.

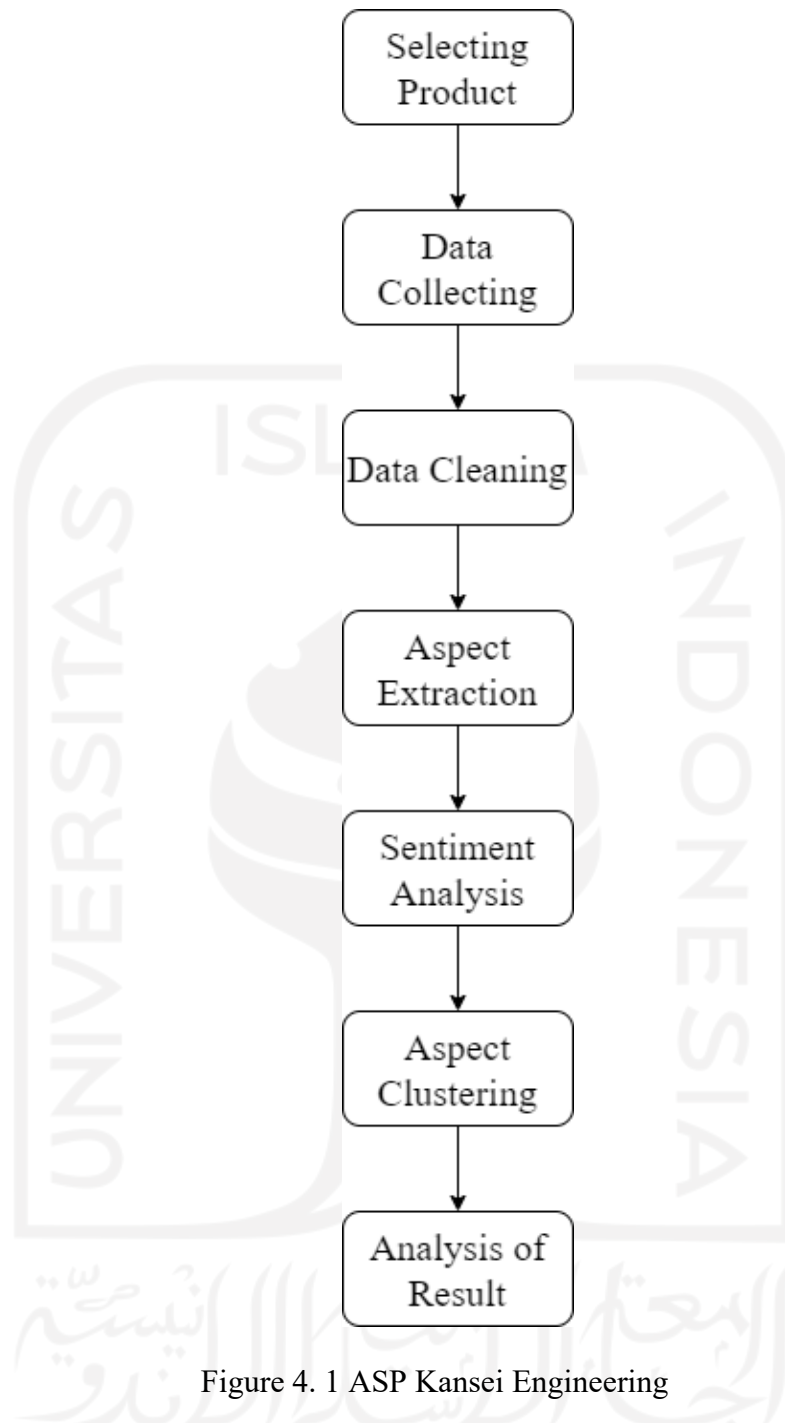


Figure 4. 1 ASP Kansei Engineering

Kansei engineering using aspect-based sentiment analysis is carried out through a few steps. The initial stage is to choose the service domain and gather data. In this paper, Samsung Galaxy s9 was selected as the target product. A huge amount of user review is available online about the product experience, which can serve as an input for Kansei engineering after aspect-based sentiment analysis. Online review is an excellent platform to gather user reviews about any product. One of the online reviews is provided by Amazon.com. Amazon has more than 150 billion active users, it provides a platform for

users to express what they experienced after buying a product on Amazon.com. therefore, python was selected to develop a scraper bot to collect user's reviews from Amazon.com. the reviews are saved directly on json file. Each review has a username, rating, date, time, header, and review text. The third phase of this work is data cleaning. The review after collected from the web is in an unstructured form therefore, data cleaning is carried out. This step is included removing HTML tag, punctuation, unused character, and facial emotion. These steps are necessary to ensure the data already in a structured format before proceeding with aspect extraction. The next step is aspect extraction and sentiment analysis. Aspect is the phrase that users regularly use to express their sentiment about a product. Aspect is mainly composed of a noun and an adjective, with the adjective functioning as a modifier for the noun. There are numerous rules for extracting nouns, adjectives, and verbs, which were previously discussed. The sentiment analysis process for this work is done with the help of Vader sentiment. When extracting aspect, the sentiment process is carried out at the same time. Next step for this work is aspect clustering. Each review will have several opinions, resulting in a large number of noun-modifier pairs. The clustering is done using K-means algorithm in Scikit-Learn. For this work, K-Means provide an optimum result with 5 clusters. The most frequent word in each cluster will be used to label the cluster. The final step is a synthesis of the result. Frequency analysis is utilized for data visualization as well as interpretation for each category. In this section, data collection, how to do data processing will be explained below.

4.2 Data Collection

The data was taken from amazon customers' reviews. Asin stands for amazon standard identification number. It is unique identification that developed by Amazon, which is specifically talking about the experience of customers' usage on the Samsung S9 smartphone. It is scraped from 3 different asins which are B07C5ZZXDG, B07VYTLC6Q, and B079H6RLKQ, all asin contain Samsung Galaxy S9 from a different retailer. There are a lot of retailers that sell Samsung galaxy s9 on amazon, but these three retailers already had good reputation and the seller already served for more than 1000 customer. The data is in the form of textual data that has been posted since August 2018 until the scrapping process is conducted.

SORT BY **FILTER BY**

Top rated | All reviewers | All stars | All formats | Text, image, video

Showing 1-10 of 2,313 reviews

Patricia Morrison
★☆☆☆☆ **No support from Amazon for defective phone.**
August 11, 2018
Color: Lilac Purple | Size: 64 GB | Style: S9 | Pattern: Single | **Verified Purchase**
Received the phone. It worked fine for the first 6 weeks, then I lost battery power I turned the phone off at night while charging, did not help. Then every else started to fail. Was sent to "Youbrakeifi" by Samsung tech support. Diagnostic showed a flawed motherboard. I am leaving for overseas next week. Amazon declined refund, so I have no phone. This the only way I can communicate, no other phone in house. I have to spend an other \$700 today to buy a phone. Will never ever purchase a phone from Amazon. No support.
1,488 people found this helpful
Helpful | 30 comments | Report abuse

Kathy A Baker
★☆☆☆☆ **Not new**
August 6, 2018
Color: Lilac Purple | Size: 64 GB | Style: S9+ | Pattern: Single | **Verified Purchase**
Repackaged with sticky cellophane that left adhesive on back of phone and camera lenses that can't be removed. Finger sensor appears dirty. And nothing was in box except phone. No papers, no chargers, no headset. Nothing.
909 people found this helpful
Helpful | 10 comments | Report abuse

Daniel Deklotz
★☆☆☆☆ **No phone in the box!**
September 16, 2018
Color: Black | Size: 64 GB | Style: S9 | Pattern: Single | **Verified Purchase**
The box I received didn't have a phone in it!!! instead, it had a cheap plastic phone case. Also, the box had no seal on it so it could have been tampered with at any point. I watched "unboxing" videos for this phone, and there's supposed to be a sticker seal you have to break to open the box. From other reviews, it sounds like other customers have received empty boxes or dummy phones.

Figure 4. 2 Amazon Customer Review Website

As shown in figure 4.2, the review data is text. The text contains reviews from the experiences of users that have purchased a product of Samsung Galaxy S9. These reviews are raw and require to be processed until it is readable for computing in the data processing. Below is a flowchart for scraping bot. the bot uses bs4 to collect html data from amazon.com then analyze the html tag to gather the data required.

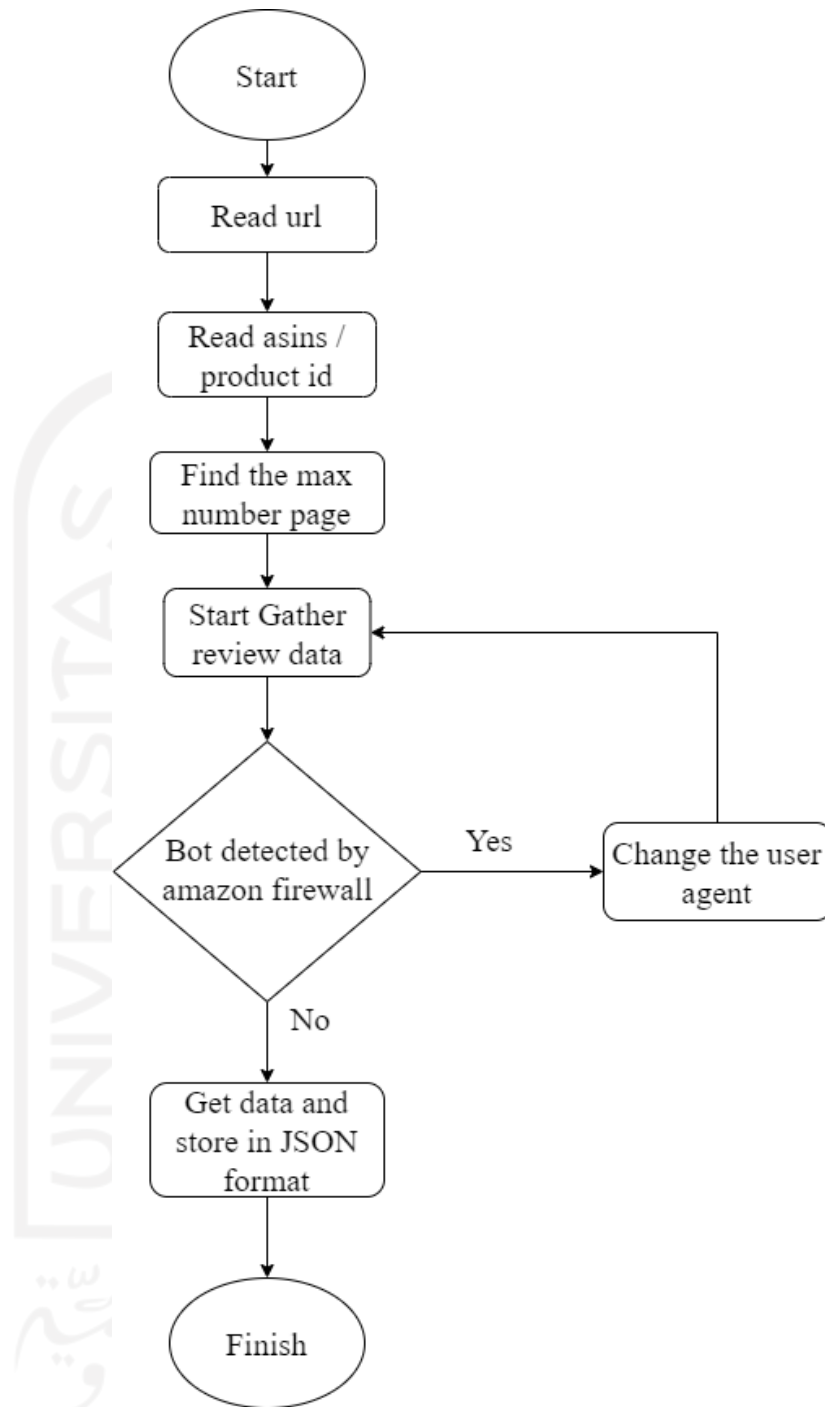


Figure 4. 3 Scrapping bot flow process

Figure 4.3 is the flow process of scrapping bot to gather data from amazon reviews. First, we need to define the amazon url and asins of the product, after that we can start the process to gather the data from the amazon web. The python script will find the max page of the product review by getting the total review in a product, to be later calculated. In order to overcome the amazon firewall, we made a bypass function to change the

browser user agent when the amazon firewall blocks the connection. The bypass function is set in looping condition, every time the bot detects captcha in soup the function will change the user agent. We use a fake-user agent library to access the agent database and will randomly use the agent. The last step is storing the data that already gathered to json and saved the json file.

This research applies the method of scrapping via the BeautifulSoup/ bs4 module in python. Bs4 is a python library that can access and process items in HTML format by addressing the syntax to access the hierarchy of the preferred accessed system. In this research, the scrapping process is conducted using python. The detail of the scrapping process can be seen below.

4.2.1 Defining the structure of the website in python

By applying the Bs4 method to scrapping the amazon website, we can investigate the website structure via developer tools in Google Chrome. The customers' reviews on the website contain some divisions of classes, such as review author, review title, review ratings, date of review, and review text. Bs4 method tries to analyze the hierarchy of the system by defining the name of the classes in python (language of programming), which can be seen in the figure shown below, figure 4.3.

```
for review in reviews_list:
    try:
        rating = review.find(attrs={'data-hook': 'review-star-rating'}).attrs['class'][2].split('-')[1].strip()
        body = review.find(attrs={'data-hook': 'review-body'}).text.strip()
        title = review.find(attrs={'data-hook': 'review-title'}).text.strip()
        author_url = review.find(attrs={'data-hook': 'genome-widget'}).find('a', href=True)
        review_url = review.find(attrs={'data-hook': 'review-title'}).attrs['href']
        review_date = review.find(attrs={'data-hook': 'review-date'}).text.strip()
```

Figure 4. 4 Data-hook for Scrapping from Website

It can be seen, for example, the defined name for body variable is using find. find is submodule for search attribute in html tag. Then we define the findings of review-body in data-hook, later the data are constructed into the text and joint all text using strip python function. The element is inspected to define the html tag that will be scraped. In which, it can be seen from inspect elements menu in chrome designated for each variable in each

desired class on the website that are going to be scraped. The encompassing classes name on each sub-title for the customer reviews in this website can be seen in Figure 4.5, 4.6, 4.7, 4.8, 4.9, and 4.10.

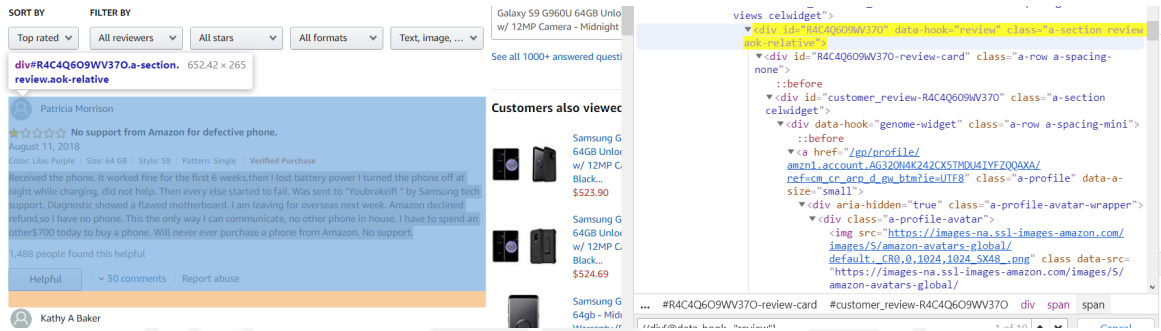


Figure 4. 5 The elements of review body on the website

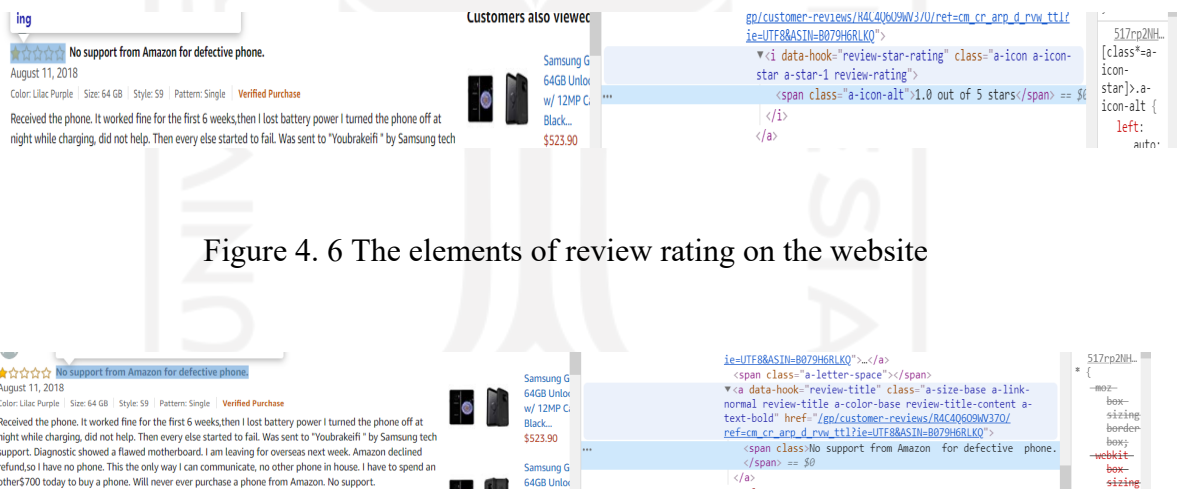


Figure 4. 6 The elements of review rating on the website

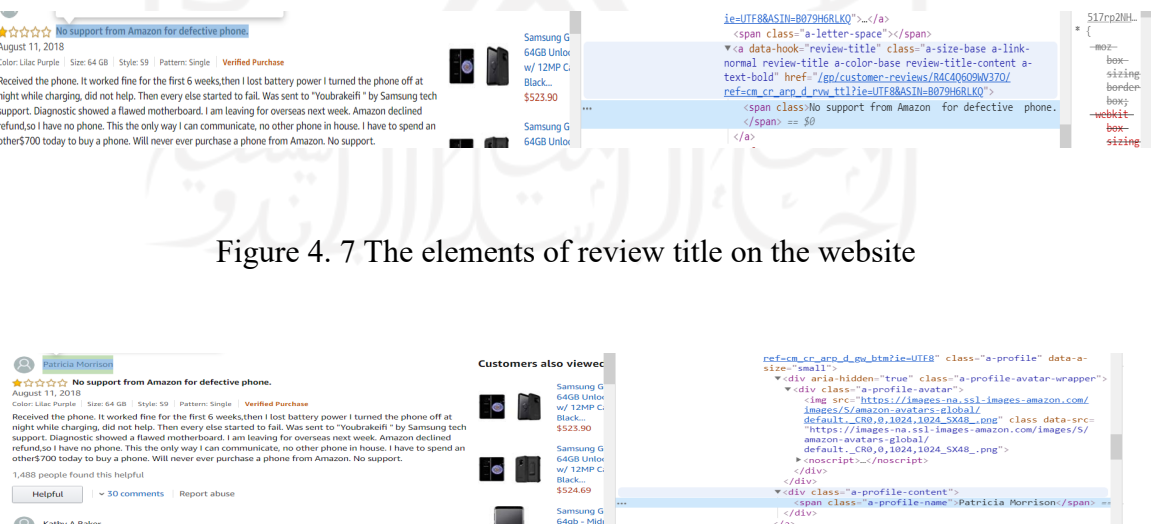


Figure 4. 7 The elements of review title on the website

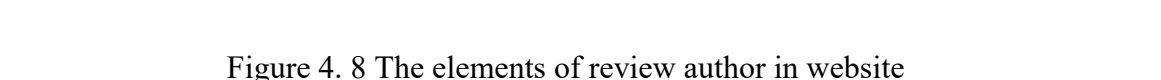


Figure 4. 8 The elements of review author in website



Figure 4. 9 The elements of review date on the website



Figure 4. 10 The elements of review helpful on the website

4.2.2 Website URL and website' ASIN input to python

In Bs4 method, The website's ASIN is used to get inside the hierarchy of the website structure specifically for amazons' websites. ASIN stands for Amazon Standard Identification Number. It is an alphanumeric unique identifier that is assigned by Amazon.com and its partners. It is used for product identification within Amazon.com organization. ASINs are only guaranteed to be unique within a marketplace. So, different national Amazon sites may use different ASINs for the same product.

```
def get_product_reviews_url(item_id, page_number=None):
    AMAZON_BASE_URL = 'https://www.amazon.com'
    if not page_number:
        page_number = 1
    return AMAZON_BASE_URL + '/product-reviews/{}/ref=' \
           'cm_cr_arp_d_paging_btm_1?ie=UTF8&reviewerType=all_reviews' \
           '&showViewpoints=1&sortBy=helpful&pageNumber={}'.format(
        item_id, page_number)

def get_comments_based_on_keyword(search): ...

def get_comments_with_product_id(product_id):
    reviews = list()
    if product_id is None:
        return reviews
    if not re.match('^[A-Z0-9]{10}$', product_id):
        return reviews

    product_reviews_link = get_product_reviews_url(product_id)
    so = get_soup(product_reviews_link)
```

Figure 4. 11 Function Define URL with Asin and Run Bs4

Figure 4.11 is functioned to run scrapper with defined URL by passing the item_id and page_number for running in beautiful soup, then the result is the whole html that will be decoded in next process.

4.2.3 The main process of scrapping

After we locate the elements of the website, afterward we need to define the function to navigate the web pages on the current websites. So, we define the function shown below.

```
def get_comments_with_product_id(product_id):
    reviews = list()
    if product_id is None:
        return reviews
    if not re.match('[A-Z0-9]{10}$', product_id):
        return reviews

    product_reviews_link = get_product_reviews_url(product_id)
    so = get_soup(product_reviews_link)

    max_page_number = so.find(attrs={'data-hook': 'total-review-count'})
    if max_page_number is None:
        return reviews
    # print(max_page_number.text)
    max_page_number = ''.join([el for el in max_page_number.text if el.isdigit()])
    # print(max_page_number)
    max_page_number = int(max_page_number) if max_page_number else 1

    max_page_number *= 0.1 # displaying 10 results per page. So if 663 results then ~66 pages.
    max_page_number = math.ceil(max_page_number)

    rev_count = 0
    for page_number in range(1, max_page_number + 1):
        if page_number > 1:
            product_reviews_link = get_product_reviews_url(product_id, page_number)
            so = get_soup(product_reviews_link)
```

Figure 4. 12 Web page navigation

The function in Figure 4.12 is objected to know how many pages the reviews, by finding total-review-count then divided by 10 so we can get the number of page reviews. For example, the review page has 1560 reviews so it will be 156 pages of review. After navigating the web page, we continue to define the json data structure that is used to hold the scrapped values using the json library in python. For Json, we need to name the column inside the structure that is used to store the data, and the codes and function of those processes mentioned above can be seen in figure 4.13.

```
reviews.append({'title': title,
               'rating': rating,
               'body': body,
               'product_id': product_id,
               'author_url': author_url,
               'review_url': review_url,
               'review_date': review_date,
               })
rev_count += 1
if (rev_count == CACHE_CHECK):
    with open(CACHE_FILE, 'w', encoding='utf-8') as fp:
        json.dump(reviews, fp, sort_keys=True, indent=4,
                  ensure_ascii=False)
    rev_count = 0
return reviews
```

Figure 4. 13 Naming the data-frame & Store in Json

Scrapper is often got blocked by Amazon. So, we need to find the solution for this problem, after finding some solution, we get into a fake user agent generator to bypass the amazon security check. This will slower the bot to take the review comment, but its works well. The code to generate a user agent can be seen below.

```
def get_soup_retry(url):
    from fake_useragent import UserAgent
    ua = UserAgent()
    UserAGR = ua.random
    if AMAZON_BASE_URL not in url:
        url = AMAZON_BASE_URL + url
    nap_time_sec = 1
    logging.debug(['Script is going to sleep for {} (Amazon throttling)'.format(nap_time_sec)])
    sleep(nap_time_sec)

    header = {
        'User-Agent': UserAGR
    }
    logging.debug('-> to Amazon : {}'.format(url))
    isCaptcha = True
    try_cnt = 0
    while isCaptcha is True:
        out = requests.get(url, headers=header)
        assert out.status_code == 200
        soup = BeautifulSoup(out.content, 'lxml')
        if try_cnt >= MAX_BAN_RETRY:
            return soup

        if 'captcha' in str(soup):
            UserAGR = ua.random
            print('Bot has been detected... retrying ... use new identity: ', UserAGR)
            isCaptcha = True
        else:
            UserAGR = ua.random
            print('Bot bypassed')
            isCaptcha = False
            return soup
        try_cnt += 1

def get_soup(url):
    soup = get_soup_retry(url)
    return soup
```

Figure 4. 14 Function to bypass Amazon security check

After running the scrapper, the scrapped data will be stored in json structure. Scrapping process itself takes at least 30 minutes to complete because the Amazon security check is block our bot. The result of scrapped data can be seen below.

```
{
  "author_url": "/gp/profile/amzn1.account.AG9YS8PH72V73KQME3ULSV0QA",
  "body": "Warning these are not Factory Unlocked 59+.They are SIM unlocked. Meaning they only work on certain US Carriers that they were originally purchased from.Verizon, AT&T, & Sprint Currently refuse to",
  "product_id": "B07CSZ2X0G",
  "rating": "1",
  "review_date": "Reviewed in the United States on September 13, 2018",
  "review_url": "/gp/customer-reviews/R28R8146W91Z7QAS1M-B07CSZ2X0G",
  "title": "not a SIMONE FACTORY UNLOCKED PHONE"
},
{
  "author_url": "/gp/profile/amzn1.account.AGR020X0X0V4M3NS5F7C0X10R0Q",
  "body": "I got this phone just in time to switch carriers as my current billing cycle ends tomorrow. I went into Verizon to setup this phone and its NOT UNLOCKED despite the \"LTE Unlocked\" sticker on the b",
  "product_id": "B07CSZ2X0G",
  "rating": "1",
  "review_date": "Reviewed in the United States on August 30, 2018",
  "review_url": "/gp/customer-reviews/R332L8N8C188F7AS1M-B07CSZ2X0G",
  "title": "NOT UNLOCKED!"
},
{
  "author_url": "/gp/profile/amzn1.account.B4821N32A2365XN9Y9R4LCEP19A",
  "body": "I received the phone today, well packaged (NOT in it's original box, but the packaging was appropriate). Upon opening the package and inspection, I was pleasantly surprised to see the original charg",
  "product_id": "B07CSZ2X0G",
  "rating": "5",
  "review_date": "Reviewed in the United States on March 29, 2019",
  "review_url": "/gp/customer-reviews/R1193UCAASUQA7AS1M-B07CSZ2X0G",
  "title": "Pay close attention - not quite as described!!!"
}
}
```

Figure 4. 15 Scrapped data result

4.3 Data Processing

For data processing, the process of data processing is started from pre-processing before being inputted to the main process. We need to clean and normalize the data to meet the standard dataset. The next process is the main processing, which is the implementation of aspect based sentiment analysis using the spacy library. The data is already scrapped from amazon using bot, which shows a specific review for Samsung Galaxy S9.

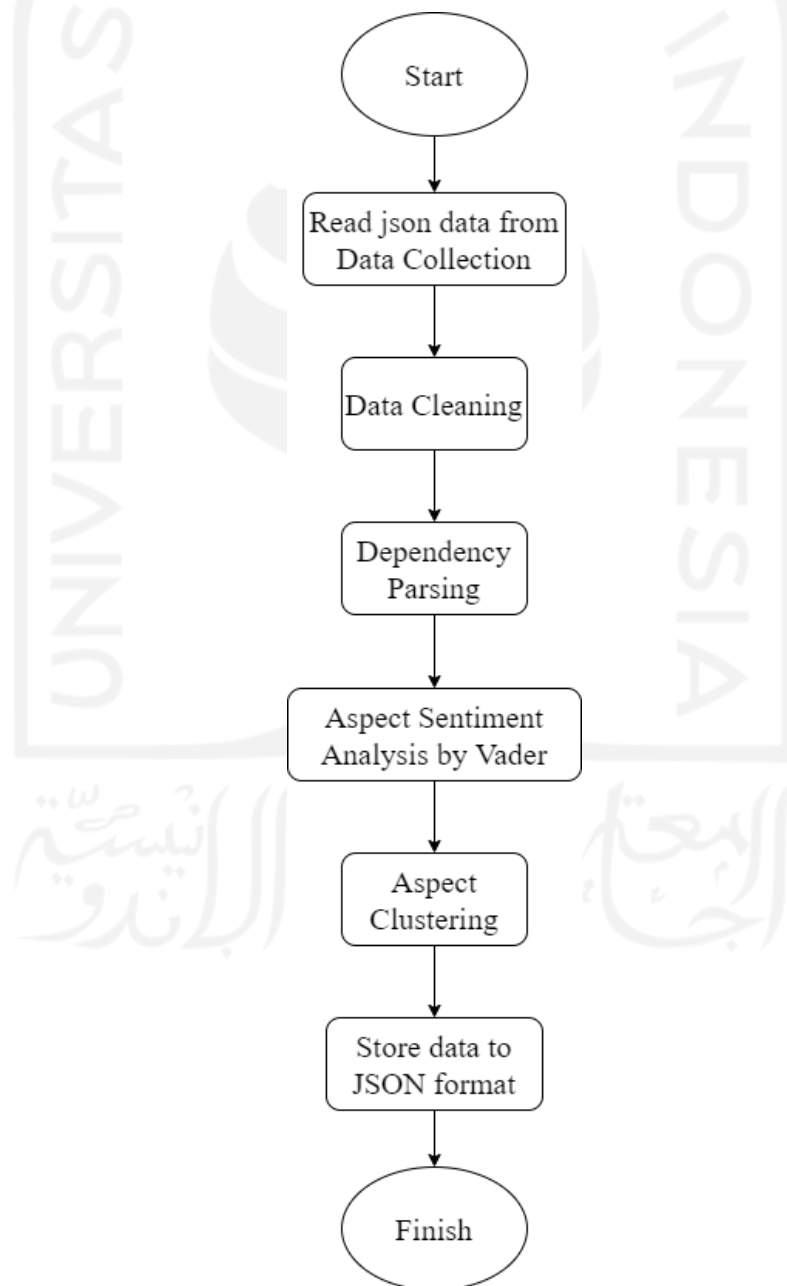


Figure 4. 16 Data Processing Flowchart

4.3.1 Pre-Processing

Pre-processing step is a crucial step for this research, a correct pre-processing will deliver good insight into the data used. Sometimes the data that we gather from the internet is not in ideal condition, so we need to make sure the data have no missing value, double data, or the format is not supported to our system. This research will clean any unnecessary string in text data so it will be read by the system that we build. The function to clean data is shown below.

```
def clean_data(df):
    pd.options.mode.chained_assignment = None

    print("*****Cleaning Started*****")

    print(f'Shape of df before cleaning : {df.shape}')
    #df['review_date'] = pd.to_datetime(df['review_date'])
    df = df[df['review_body'].notna()]
    df['review_body'] = df['review_body'].str.replace("<br />", " ")
    df['review_body'] = df['review_body'].str.replace("[?\.+?]?\\", " ")
    df['review_body'] = df['review_body'].str.replace("\\{3,}", " ")
    df['review_body'] = df['review_body'].str.replace("&\\#.+\\&\\#\\d+?;", " ")
    df['review_body'] = df['review_body'].str.replace("\\d+\\&\\#\\d+?;", " ")
    df['review_body'] = df['review_body'].str.replace("\\&\\#\\d+?;", " ")

    #facial expressions
    df['review_body'] = df['review_body'].str.replace("\\:\\|", "")
    df['review_body'] = df['review_body'].str.replace("\\:\\)", "")
    df['review_body'] = df['review_body'].str.replace("\\:\\(", "")
    df['review_body'] = df['review_body'].str.replace("\\:\\V", "")

    #replace multiple spaces with single space
    df['review_body'] = df['review_body'].str.replace("\\s{2,}", " ")

    df['review_body'] = df['review_body'].str.lower()
    print(f'Shape of df after cleaning : {df.shape}')
    print("*****Cleaning Ended*****")

    return(df)
```

Figure 4. 17 Clean data function

For cleaning the dataset, we use pandas library and some regex python functions to remove a character that will make noise in the dataset. For example, html tag (
), punctuation (?.), unused character, and facial emoticon. If the dataset does not clean from the unnecessary string will make an error on the next step. So, this step is very important before can go to the main process.

4.3.2 Dependency Parsing

Dependency parsing is a method that functions to analyze the grammatical structure of a sentence based on the dependencies between words. Dependency parsing can be used to find nouns and adjectives to complete this research for creating Kansei guidelines on Samsung galaxy s9.

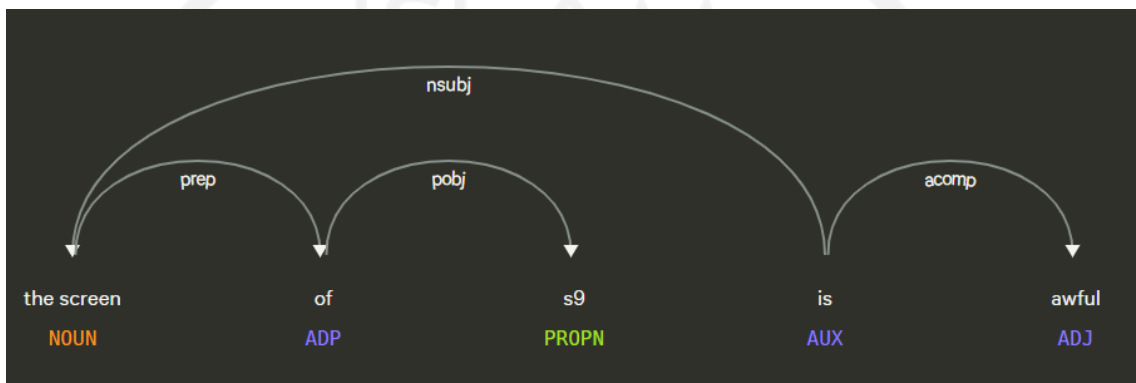


Figure 4. 18 Noun Adjective Visualizer

In Figure 4.18, it is shown the relationship between the word, there is noun, adjective, and else. There is also a relationship between ‘the screen’ and ‘is’ it is nominal subject. There is the next relationship with adjective complement in the word ‘awful’. There is a dictionary concerning word dependence and part of speech in the spacy library ([Glossary](#)).

From the explanation above, several rules are developed to extract nouns and adjectives. There are several rules to catch the noun and adjective in review by using rule matching function from spacy. From paper titled Concept, methods, and tools in Kansei engineering by (Schütte et al., 2004), stated a Kansei word is a term that refers to a product's domain. Adjectives are frequently used, but other grammatical forms are also conceivable. For example, when describing the domain 'fork-lift truck,' adjectives such as effective, sturdy, rapid, and so on can be used, as well as verbs and nouns (acceleration). Shown below for the function of extracting the adjective, verb, adverb and noun in review.

```

for token in doc:
    A = "999999"
    M = "999999"
    if token.dep_ == "amod" and not token.is_stop:
        M = token.text
        A = token.head.text

        # add adverbial modifier of adjective (e.g. 'most comfortable headphones')
        M_children = token.children
        for child_m in M_children:
            if(child_m.dep_ == "advmod"):
                M_hash = child_m.text
                M = M_hash + " " + M
                break

        # negation in adjective, the "no" keyword is a 'det' of the noun (e.g. no interesting characters)
        A_children = token.head.children
        for child_a in A_children:
            if(child_a.dep_ == "det" and child_a.text == 'no'):
                neg_prefix = 'not'
                M = neg_prefix + " " + M
                break

    if(A != "999999" and M != "999999"):
        if A in prod_pronouns :
            A = "product"
        dict1 = {"noun" : A, "adj" : M, "rule" : 1, "polarity" : sid.polarity_scores(token.text)['compound']}
        rule1_pairs.append(dict1)

```

Figure 4. 19 Rule 1

For rule 1, as shown in Figure 4.19, this function is purposed to extract adjectives and nouns with adverbial modifier relationships. In the function we declare that if the token has dependency 'advmod' then will be stored in dataframe, if the A / noun exists in prod_pronouns which is 'it', 'this', 'they', and 'these' it will be replaced with 'product' noun. For the polarity is using nltk vader sentiment by passing adjective to the function and will return with polarity compound.

```
children = token.children
A = "999999"
M = "999999"
add_neg_pfx = False
for child in children :
    if(child.dep_ == "nsubj" and not child.is_stop):
        A = child.text
        # check_spelling(child.text)

    if((child.dep_ == "dobj" and child.pos_ == "ADJ") and not child.is_stop):
        M = child.text
        #check_spelling(child.text)

    if(child.dep_ == "neg"):
        neg_prefix = child.text
        add_neg_pfx = True

if (add_neg_pfx and M != "999999"):
    M = neg_prefix + " " + M

if(A != "999999" and M != "999999"):
    if A in prod_pronouns :
        A = "product"
    dict2 = {"noun" : A, "adj" : M, "rule" : 2, "polarity" : sid.polarity_scores(token.text)['compound']}
    rule2_pairs.append(dict2)
```

Figure 4. 20 Rule 2

For rule 2, the function is to extract nouns and adjectives from review with noun has nominal subject and adjective that has direct object relationship. For example, 'Toyota defeated Honda' there is nsubj relationship between toyota and defeated. The direct object is like 'samsung make me dissapointed' there is a direct object relationship.

```
children = token.children
A = "999999"
M = "999999"
add_neg_pfx = False
for child in children :
    if(child.dep_ == "nsubj" and not child.is_stop):
        A = child.text
        # check_spelling(child.text)

    if(child.dep_ == "acom" and not child.is_stop):
        M = child.text

    # example - 'this could have been better' -> (this, not better)
    if(child.dep_ == "aux" and child.tag_ == "MD"):
        neg_prefix = "not"
        add_neg_pfx = True

    if(child.dep_ == "neg"):
        neg_prefix = child.text
        add_neg_pfx = True

if (add_neg_pfx and M != "999999"):
    M = neg_prefix + " " + M
    #check_spelling(child.text)

if(A != "999999" and M != "999999"):
    if A in prod_pronouns :
        A = "product"
    dict3 = {"noun" : A, "adj" : M, "rule" : 3, "polarity" : sid.polarity_scores(token.text)['compound']}
    rule3_pairs.append(dict3)
    #rule2_pairs.append((A, M, sid.polarity_scores(M)['compound'], 2))
```

Figure 4. 21 Rule 3

Rule 3 is for extracting aspect from noun with relation with nsubj and adjective with acomp relation, for example, ‘The sound of the speakers would be better.’. So, the noun is ‘speaker’ with ‘would’ as nsubj, adjective is ‘better’ with ‘be’ as acomp.

```
children = token.children
A = "999999"
M = "999999"
add_neg_pfx = False
for child in children :
    if((child.dep_ == "nsubjpass" or child.dep_ == "nsubj") and not child.is_stop):
        A = child.text
        # check_spelling(child.text)

    if(child.dep_ == "advmod" and not child.is_stop):
        M = child.text
        M_children = child.children
        for child_m in M_children:
            if(child_m.dep_ == "advmod"):
                M_hash = child_m.text
                M = M_hash + " " + child.text
                break
        #check_spelling(child.text)

    if(child.dep_ == "neg"):
        neg_prefix = child.text
        add_neg_pfx = True

if (add_neg_pfx and M != "999999"):
    M = neg_prefix + " " + M

if(A != "999999" and M != "999999"):
    if A in prod_pronouns :
        A = "product"
    dict4 = {"noun" : A, "adj" : M, "rule" : 4, "polarity" : sid.polarity_scores(token.text)['compound']}
    #print (dict4)
    rule4_pairs.append(dict4)
```

Figure 4. 22 Rule 4

Rule 4 is for extract aspect with noun has relation with nominal subject pass and adjective with relation to the adverbial modifier. For example, ‘The headphone is most comfortable headphones ever.’ The headphone is a noun with nsubjpass to ‘is’, for the adjective is ‘comfortable’ with adverbial modifier ‘most’.

```
children = token.children
A = "999999"
M = "999999"
add_neg_pfx = False
for child in children :
    if(child.dep_ == "nsubj" and not child.is_stop):
        A = child.text
        # check_spelling(child.text)

    if((child.dep_ == "attr") and not child.is_stop):
        M = child.text
        #check_spelling(child.text)

    if(child.dep_ == "neg"):
        neg_prefix = child.text
        add_neg_pfx = True

if (add_neg_pfx and M != "999999"):
    M = neg_prefix + " " + M

if(A != "999999" and M != "999999"):
    if A in prod_pronouns :
        A = "product"
    dict7 = {"noun" : A, "adj" : M, "rule" : 7, "polarity" : sid.polarity_scores(M)['compound']}
    #print (dict7)
    rule7_pairs.append(dict7)
```

Figure 4. 23 Rule 5

Rule 5 is to extract noun and adjective with an attribute such as great, seem, appear, and is. For example, “it’s a great misrepresentation that is omitted when purchasing.” From review data. ‘great’ in a sentence is an attribute that connects between words, and the result is (it’s, misrepresentation). “this’ in the sentence will be replaced with ‘product’ because it exists in pronouns and the adjective is a misrepresentation.

```
text = row["review"]
doc=nlp(text)
rule = []
#displacy.render(doc, style="dep", jupyter=True, options={'distance': 100})
for token in doc:
    if (token.head.pos_ == "VERB" and token.dep_ == "dobj"):
        adj = token.head.text
        #adj = token.text
        children = token.children
        for child in children:
            if (child.dep_ == "compound" or child.dep_ == "amod" or child.dep_ == "advmod") :
                aspect = child.text
                dict_6 = {"noun" : aspect, "adj" : adj, "polarity" : sid.polarity_scores(adj)['compound']}
                rule.append(dict_6)
                print (rule)
```

Figure 4. 24 Rule 6

Rule 6 is designated to extract verbs on the review. A verb is defined by the word that has ‘VERB’ pos and has a “direct object” dependency. For the noun is a word that has one of the compound, amod, or advmod.

```
aspects = []  
  
aspects = rule1_pairs + rule2_pairs + rule3_pairs + rule4_pairs + rule5_pairs + rule6_pairs + rule7_pairs  
  
dic = {"product_id" : product_id , 'title' : title, 'rating' : rating , "review_body" : review_body, "aspect_pairs" : aspects}  
return dic
```

Figure 4. 25 Store Json Data

After defining all the rules that needed to extract aspect-pairs, we need to combine the result to the new json to store the data that will be used in aspect clustering.

4.3.3 Aspect Clustering

Aspect clustering is needed to see the group of the aspect/noun. So, we can get better insight into the product review. Since every product has multiple reviews and has a lot of aspect and sentiment pairs. We also found that different word is used for a very similar aspect of the product. So, we can get a better understanding by seeing the cluster. The clustering in this research will use the build-in spacy library for the vectorization and will use sklearn to create the cluster of 4. For the explanation, the code can be seen below.

```
def main():  
    time1 = time()  
    nlp = init_spacy()  
    time2 = time()  
    print("-----**-----")  
    print("Unloading aspect pairs file")  
    with open("/content/drive/MyDrive/Colab Notebooks/new dataset/reviews_aspect_mapping.json", 'r') as fobj:  
        reviews_data = json.load(fobj)  
    print("Finished loading aspect pairs!\n")  
    print("-----**-----")  
    time3 = time()  
    update_reviews_data(reviews_data, nlp)  
    aspect_json_encoding("/content/drive/MyDrive/Colab Notebooks/new dataset/model_resultsx.json", "/content/drive/MyDrive/Colab Notebooks/new dataset/model_results_encodingx.json")  
    time4 = time()  
    print("Time for loading spacy: {:.2}s".format(time2-time1))  
    print("Time for loading aspects json file: {:.2}s".format(time3-time2))  
    print("Time for running aspect clustering: {:.2}s".format(time4-time3))
```

Figure 4. 26 Main Clustering Function

```
def update_reviews_data(reviews_data, nlp):
    updated_reviews = []
    ctr = 0
    print("Total number of unique products in this category: {}".format(len(reviews_data)))

    for i,product in enumerate(reviews_data):
        for prod_id, this_product_reviews in product.items():
            print (prod_id)
            this_product_upd_reviews = add_clusters_to_reviews(this_product_reviews, prod_id, nlp)
            updated_reviews.append(this_product_upd_reviews)

        if ((i%100 == 0) ):
            ctr += 1
            print("\n-----**-----")
            print("Updating results - batch {}".format(ctr))
            with open('/content/drive/MyDrive/Colab Notebooks/new dataset/model_resultsx.json', 'a') as f:
                json.dump(updated_reviews,f)
            updated_reviews = []
            print("Finished writing results to json!!")
            print("-----**-----")

    # Finally write to output file
    ctr += 1
    print("\n-----**-----")
    print("Updating results - batch {}".format(ctr))
    with open('/content/drive/MyDrive/Colab Notebooks/new dataset/model_resultsx.json', 'a') as f:
        json.dump(updated_reviews,f, indent=2)
    updated_reviews = []
    print("Finished writing results to json!!")
    print("-----**-----")
```

Figure 4. 27 Update review data with cluster

```
def get_aspects(reviews_data):
    aspects = []
    for review in reviews_data:
        aspect_pairs = review["aspect_pairs"]
        for map in aspect_pairs:
            aspects.append(map['noun'])
    return aspects
```

Figure 4. 28 Get Aspect Function

Those functions above designated is to run the Aspect clustering by passing the data from aspect extracting (dependency parsing). In Figure 4.28, it is functioned to get nouns in the json data then return the aspect data to next process.


```
def get_aspect_freq_map(aspects):  
    aspect_freq_map = defaultdict(int)  
    for asp in aspects:  
        aspect_freq_map[asp] += 1  
    return aspect_freq_map
```

Figure 4. 29 Get Aspect Frequency

The function in Figure 4.29 is to get the frequency of noun/aspect by calculating every single word exists in aspect, the aspect gets from get_aspect function. Then define in the loop to iterate over every word in aspects and will return aspect_freq_map as the result in an integer.

```
def get_word_vectors(unique_aspects, nlp):  
    asp_vectors = []  
    for aspect in unique_aspects:  
        token = nlp(aspect)  
        asp_vectors.append(token.vector)  
    return asp_vectors
```

Figure 4. 30 Get Word Vector

```
def get_word_clusters(unique_aspects, nlp):  
    # print("Found {} unique aspects for this product".format(len(unique_aspects)))  
    asp_vectors = get_word_vectors(unique_aspects, nlp)  
    print (unique_aspects)  
    print (len(unique_aspects))  
    if len(unique_aspects) <= NUM_CLUSTERS:  
        # print("Too few aspects ({} found. No clustering required...".format(len(unique_aspects)))  
        return list(range(len(unique_aspects)))  
  
    # print("Running k-means clustering...")  
    n_clusters = NUM_CLUSTERS  
    kmeans = cluster.KMeans(n_clusters=n_clusters)  
    kmeans.fit(asp_vectors)  
    labels = kmeans.labels_  
    print (labels)  
    return labels
```

Figure 4. 31 Get Word Clusters

In Figure 4.31, the function is to get cluster by using sklearn kmeans cluster. First, we must get word vectors because to calculate in kmeans cluster model need to convert first to array. Next is to do the fit process by passing asp_vectors variable, we also define the number of clusters of 4. After the fit process is done, we will take out the kmeans labels that will be used to show the cluster mapping.

```
def get_cluster_names_map(asp_to_cluster_map, aspect_freq_map):  
    cluster_id_to_name_map = defaultdict()  
    clusters = set(asp_to_cluster_map.values())  
    for i in clusters:  
        this_cluster_asp = [k for k,v in asp_to_cluster_map.items() if v == i]  
        filt_freq_map = {k:v for k,v in aspect_freq_map.items() if k in this_cluster_asp}  
        filt_freq_map = sorted(filt_freq_map.items(), key = lambda x: x[1], reverse = True)  
        cluster_id_to_name_map[i] = filt_freq_map[0][0]  
    return cluster_id_to_name_map
```

Figure 4. 32 Get Cluster Names Map

In Figure 4.32, the function is objected to map the aspect to the cluster by iterating through the cluster then returning the id to identify the cluster. This will be used to map the noun and the cluster in the next process.

```
def add_clusters_to_reviews(reviews_data, prod_id, nlp):
    product_aspects = get_aspects(reviews_data)
    aspect_freq_map = get_aspect_freq_map(product_aspects)
    unique_aspects = aspect_freq_map.keys()

    aspect_labels = get_word_clusters(unique_aspects, nlp)
    asp_to_cluster_map = dict(zip(unique_aspects, aspect_labels))
    cluster_names_map = get_cluster_names_map(asp_to_cluster_map, aspect_freq_map)
    updated_reviews = []

    for review in reviews_data:
        aspect_pairs_upd = []
        aspect_pairs = review["aspect_pairs"]
        for map in aspect_pairs:
            noun = map['noun']
            cluster_label_id = asp_to_cluster_map[noun]
            cluster_label_name = cluster_names_map[cluster_label_id]
            map['cluster'] = cluster_label_name
            aspect_pairs_upd.append(map)

        review['aspect_pairs'] = aspect_pairs_upd
        updated_reviews.append(review)
    result = {prod_id:updated_reviews}
    return result
```

Figure 4. 33 Add Cluster to Review

In Figure 4.33, the function is used to add a cluster column to the data frame. The set function in `add_label_to_reviews` can be seen in figure 4.28 until figure 4.32. then for update/add cluster to the data frame uses iteration method to slice every data on `reviews_data`. Then will add cluster in every aspect pair by passing the noun on the data frame to the clustering function, then will return the label to be inputted in the data frame.

4.4 Data Visualization

Data visualization is one of the steps of data processing, in order to get a better sight of the result data by showing chart, map, or graph. The goals of data visualization are to make the reader to easier to identify patterns, trends, or outliers in large data sets. Below will show the function that is used in this process.

```
with open ("/content/drive/MyDrive/Colab Notebooks/new dataset/model_results_encoding.json", "r") as f:
    data = json.load(f)

def json2Df (data) :
    num = range(len(data))
    results = []
    counting = 0
    #num = range(1) ## Unit test
    for n in num:
        for item in data[n]:
            print (item)
            product_id = item
            for x in data[n][item]:
                review_id = counting
                counting += 1
                for y in x['aspect_pairs']:
                    record = {
                        'review_id' : review_id,
                        'product_id' : x['product_id'],
                        'title' : x['title'],
                        'rating' : x['rating'],
                        'review' : x['review_body'],
                        'noun' : y['noun'],
                        'adj' : y['adj'],
                        'polarity' : y['polarity'],
                        'cluster' : y['cluster']}
                    results.append(record)
    return results
```

Figure 4. 34 Convert JSON to Dataframe

To make it simpler to process in the following phase, a new function is written to convert json, which will then be loaded into a new data frame. In Figure 4.34, the function is an iteration process that reads every single dictionary and key in json, then stores them in a dictionary, las pass the dictionary to a new data frame.

```
def get_wordcloud(dataframeAndColumn,wordcloudname,stp,saveornot,showornot):
    stop = stp
    text = " ".join(review for review in dataframeAndColumn.astype(str))
    print ("{} Word in {}".format(len(text)) + wordcloudname)
    if stop != "":
        wordcloud = WordCloud(stopwords= stop, width=800, height=600, background_color='black',collocations=False).generate(text)
    else :
        wordcloud = WordCloud(width=800, height=600, background_color='black',collocations=False).generate(text)
    if showornot == True:
        plt.figure(figsize=(15, 15))
        plt.imshow(wordcloud, interpolation='bilinear')
        plt.axis("off")
        plt.show()
    if saveornot == True:
        wordcloud.to_file(path_picture + wordcloudname + ".png")
```

Figure 4. 35 Get Wordcloud Function

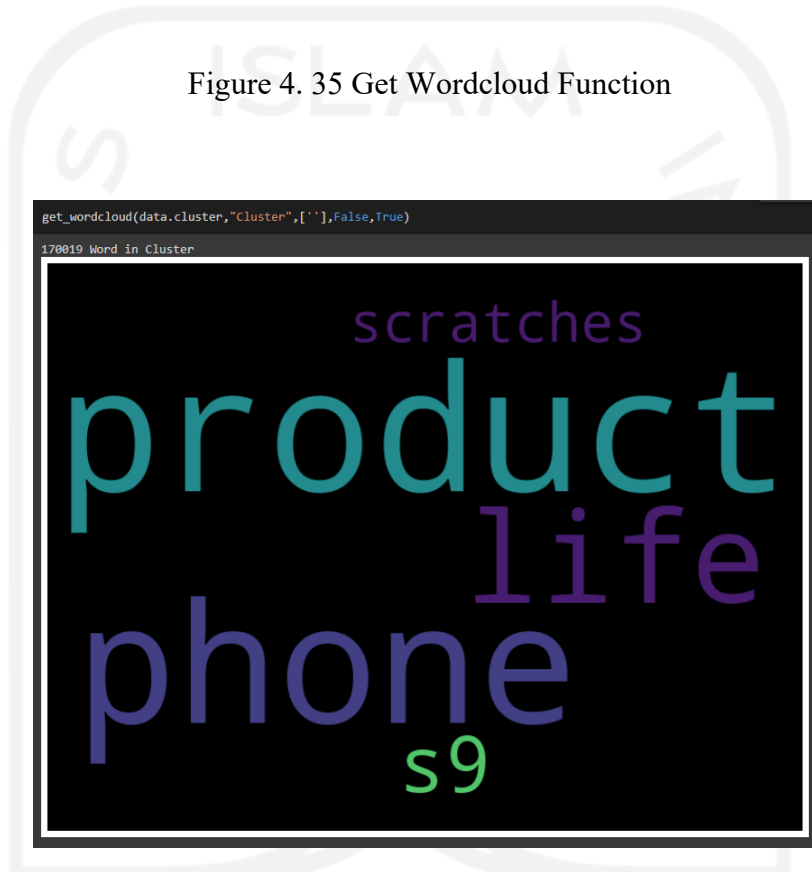


Figure 4. 36 Wordcloud Result

A function to make a word cloud is shown in Figure 4.35. A way to visualize data, particularly text data, is using a word cloud. The procedure is straightforward: count each word and display the results in a word cloud. The more frequently a term appears in the data, the thicker the word cloud becomes. Figure 4.36 depicts the outcome of the word cloud.

```
def get_top_n_words(corpus, stp, n=None):  
    if stp != "":  
        vec=CountVectorizer(stop_words=stp).fit(corpus)  
    else :  
        vec=CountVectorizer().fit(corpus)  
    bag_of_words = vec.transform(corpus)  
    sum_words = bag_of_words.sum(axis=0)  
    words_freq = [(word, sum_words[0, idx]) for word, idx in vec.vocabulary_.items()]  
    words_freq =sorted(words_freq, key = lambda x: x[1], reverse=True)  
    return words_freq[:n]
```

Figure 4. 37 Get Top Words Function

```
def DrawChart(titlename, isidataframebro, by="count", x_axis="adjective", y_axis="count", color="red", savename=None):  
    fig, ax = plt.subplots(figsize=(8, 8))  
  
    isidataframebro.sort_values(by=by).plot.barh(x=x_axis,  
                                                y=y_axis,  
                                                ax=ax,  
                                                color=color)  
    ax.set_title(titlename)  
    if savename != None:  
        plt.savefig(path_picture + savename + ".png")  
    return plt.show()
```

Figure 4. 38 Draw Chart Function

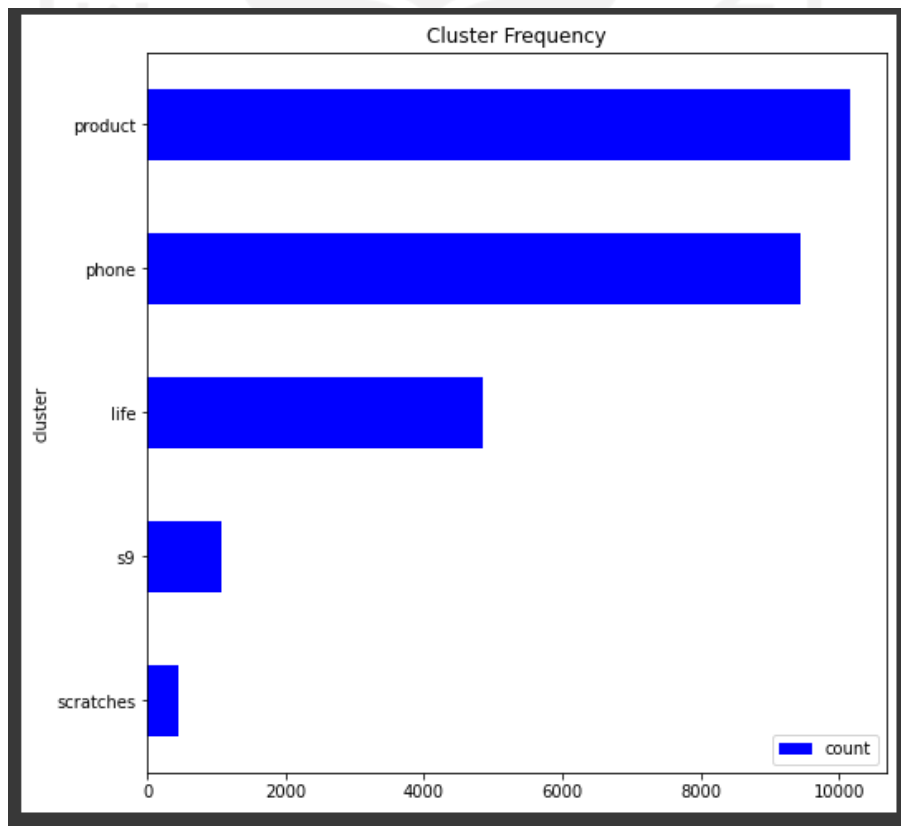


Figure 4. 39 Frequency Chart

We construct a function to determine the top word in the review in Figure 4.37. The text is converted to tokens using sklearn countvectorizer, and then a vocabulary of known words is built. The vocabulary is then encoded into new documents and stored in a new data frame. Figure 4.38 is a matplotlib method for drawing a frequency chart. The data required to create a frequency chart will be provided by the get top word function. Figure 4.39 shows the outcome of the draw function.

That all the process for data preprocessing the last process for data processing is data visualization. The next process of this research is the discussion of the result from those functions that will lead us to creating Kansei product guidelines.



Chapter V

RESULTS AND DISCUSSION

5.1 Kansei words identification

A Kansei word refers to a term that describes the domain of a product. Adjectives are the most common type, although there are other grammatical forms as well. Verbs and nouns, on the other hand, can be used in the same way (Schütte et al., 2004). Finding the relationship between modifier and noun is the first step in identifying Kansei word. The researcher extracts the aspect and aspect modifier for every aspect in this context. The objective is to know what the user says in every aspect, because the noun will represent the aspect, and the adjective is what the user says about the product. After doing aspect clustering, this is the retrieved result. Because there are so many elements to the noun, the researcher clusters it in order to classify it. This allows us to see the groups of nouns that may occur in the outcome. The result can see below in figure 5.1.

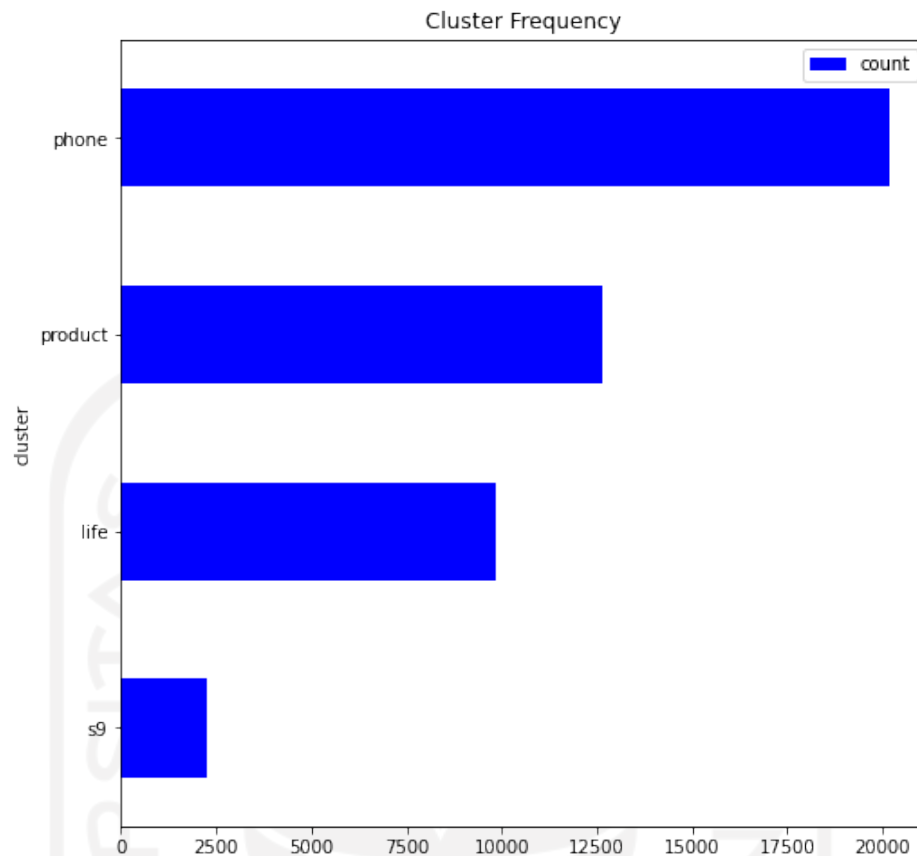


Figure 5. 1 Most Frequent Word in Cluster

Aspect clustering, as shown in Figure 5.1, produced four clusters: phone, product, life, and s9. The ‘product’ cluster is a group of overall review the on Amazon seller service not the Samsung s9. The noun in ‘product’ cluster, such as condition, price, quality, deal, seller, service, shipping, etc. The noun and adjective that have a high link with the Samsung galaxy s9 are represented by the phone and s9 cluster. Phone, screen, camera, charger, battery, gadget, card, and so on are nouns in the cluster. Because the life cluster has no association with the Samsung Galaxy S9, the researchers will disregard it.

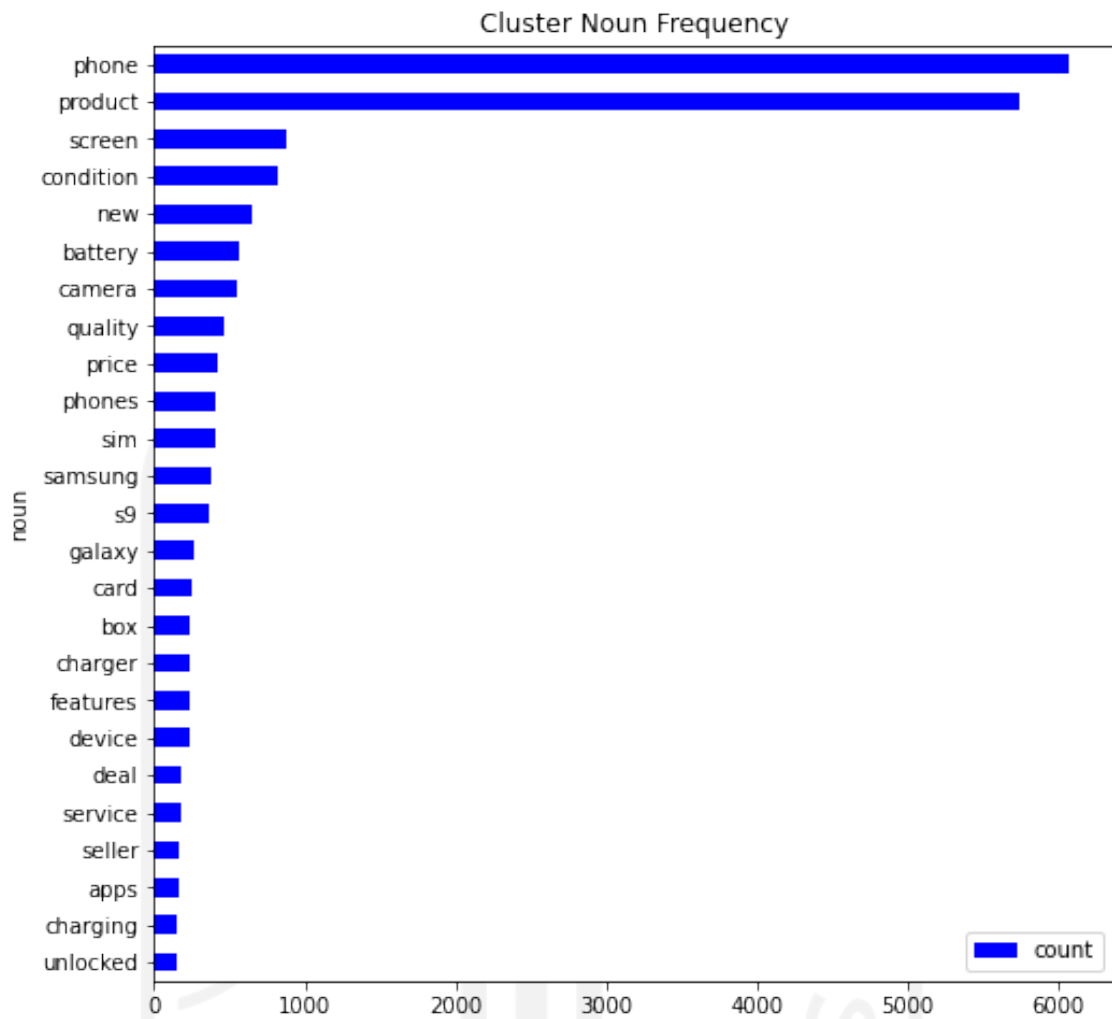


Figure 5. 2 Cluster Noun Frequency

From Figure 5.2, it can be inferred that the most 25 frequent noun words in the review. In the frequency chart, the ‘product’ represents the overall of the Samsung galaxy s9. However, not all the words represent the smartphone aspect, so the researcher decided to look up aspects that correlate with smartphone features. Referring to previous research by (Gupta et al., 2019) in Aspect based sentiment analysis of mobile review. In the (Gupta et al., 2019) research, they do observations on various smartphone specifications from GSMarena. They identified aspect categories as shown in Table 5.1.

Table 5. 1 Table Aspect Vector Gupta Research, adapted from (Gupta et al., 2019)

Aspect Category	Aspect Terms
Network	Network call quality, connectivity, network signals, call recording
Body	Dimension, body, design, weight, build quality
Display	Size, resolution, multitouch, glass, screen, display, touchscreen, touch screen, led, LCD, touchpad, touch pad, display quality, touch, screen size, ui, screen display quality, screen quality.
Platform	Chipset, os, operating system, CPU, gpu, ios, processing, android.
Performance	Processor, clock speed, cache, startup, bootup, boot up, start up, boots up, starts up, performing, operating system, performance, run, runs, perform, performs, speed, respond, responds, response, keyboard, navigate, navigation, battery performance, phone performance, processor speed.
Memory	Card slot, internal memory, external memory, ram, rom, microsd support
Camera	Primary camera, secondary camera, video, image, photo, flash, hdr, panorama, front camera, rear camera, night mode, camera quality, photo quality, pictures, dual camera, video recording quality, recording quality, video quality, slow-motion recording, front camera quality, rear camera quality, picture quality, mobile camera, camera features, portrait mode.
Sound	Loudspeaker, speaker, speakers, audio jack, vibration, stereo, mic, headphone, headphones, audio, sound, voice, microphone, sound quality, audio quality, phone mic, speakers quality, loudspeaker sound, sound effect.
Communication	Wlan, Bluetooth, wifi, hotspot, airdrop, 3g,4g, lte, volte, gsm, 2g, wcdma, nfc, radio, usb, gps, hotspot.
Features	Sensors, messaging, browser, dual sim, fingerprint sensor, facial unlock, digital compass, ambient light sensor, accelerometer, status indicator, email, features, fingerprint, sensor, fingerprint sensor, fingerprint sensor.
Battery	Standby, stand-by, backup, backup, power, charger, battery, battery life, adapter, battery backup, battery quality, fast charging, battery charges.

Gupta said in the research that aspect categories such as Network, Body, Display, Platform, Performance, Memory, Camera, Sound, Comms, Features, Battery. Compared to figure 5.2. The most frequent word in the cluster, there is the similarity between the previous research such as screen/display, camera, charger, battery, speaker/sound. So, the researcher decided to explore more in these features which is screen/display, camera, charger, battery, and speaker/sound.

5.1.1 Screen Feature

The screen is one of the characteristics that has been picked. Customers have given the screen the highest rating on the chart. In the chart, however, the terms screen and display have quite similar meanings. So, in this scenario, we'll combine the charts and look for the most common adjectives.

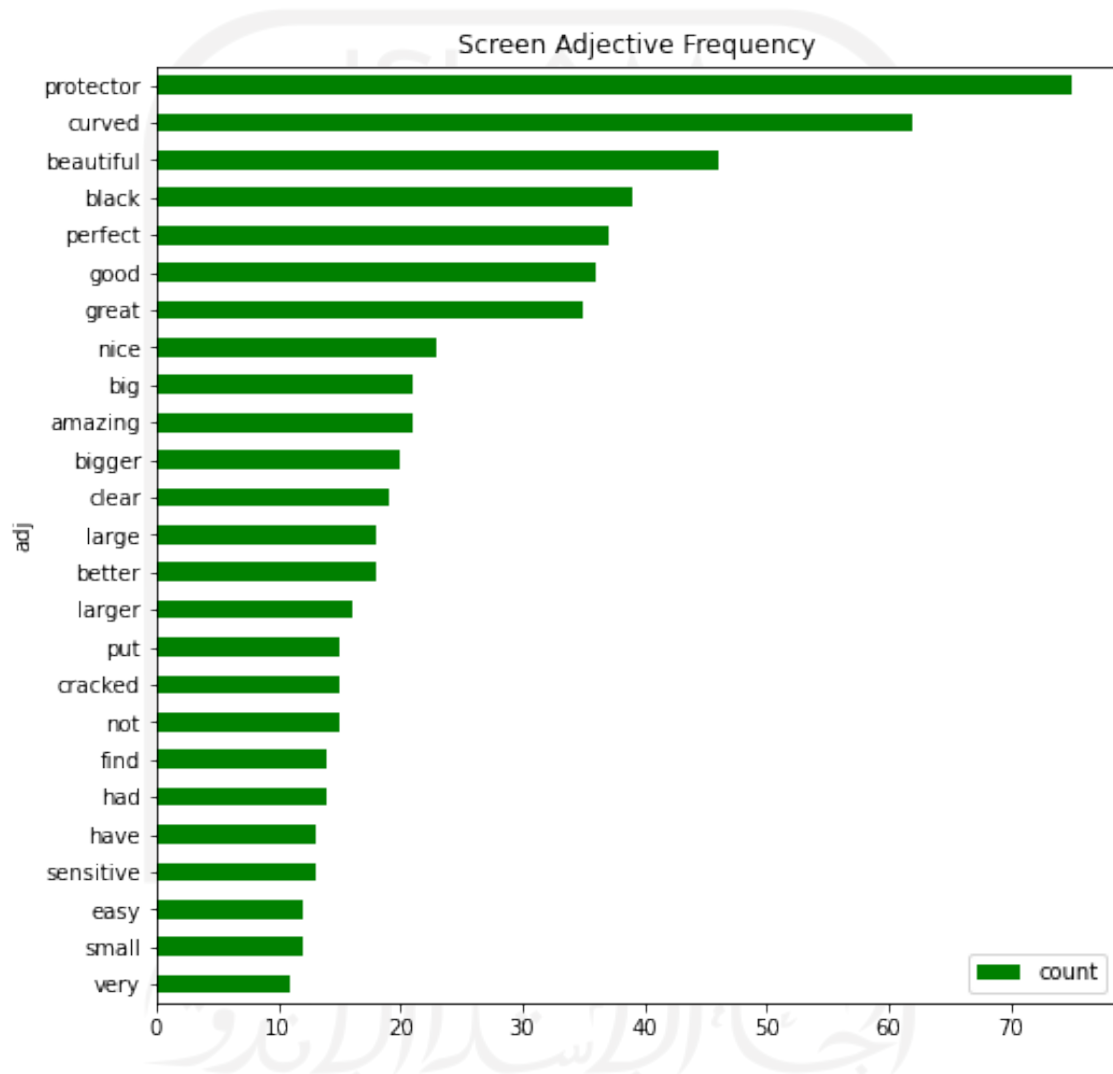


Figure 5. 3 Most Frequent Adjective in Screen Cluster

From Figure 5.3, it is shown that the screen in the product seems to get positive sentiment. The chart shows that most users say that the screen is a protector, curved, beautiful, great, and good. However, to identify the problem that is needed for product improvement, the researcher needs to find the negative word from the review can be seen below.

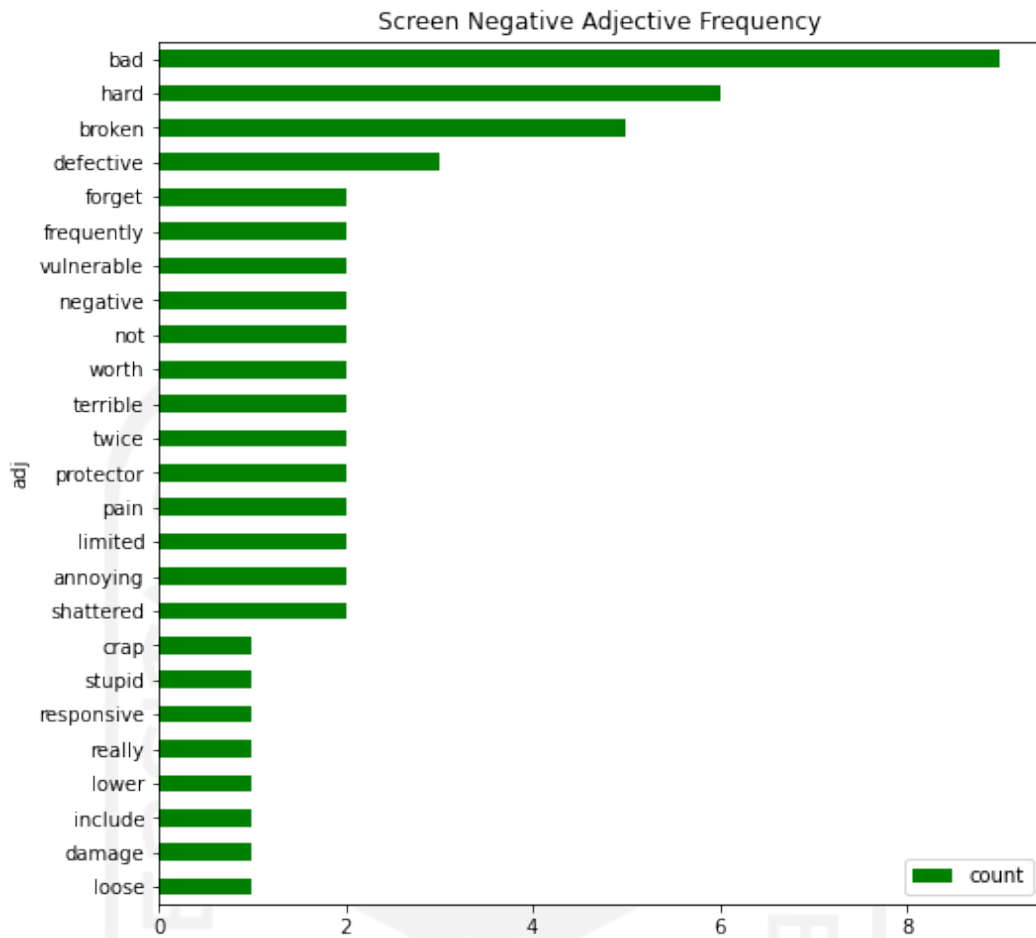


Figure 5. 4 Negative Adjective in Screen Feature

In this step, the researcher tries to explore more from the negative adjective word in the screen feature. The negative adjective obtained by filter polarity with less than 0. As for the negative frequent adjective words from customers. From figure 5.4, we found that the customer has complained about the bad screen quality, the hard screen has bad sentiment from some user, the screen is considered broken and defective. There are several recommendations to improve the next series products. The manufacturer still can keep several features to maintain the customer's satisfaction, yet some improvements should be made for several features such as the quality and durability of the screen.

5.1.2 Camera Feature

The second highest frequency noun is the camera with 529 words mentioned in the review. The camera is very often to use in smartphones also many manufacturers aggressively implement higher quality for their smartphones so, this will reveal a lot of information which is good for product improvement.

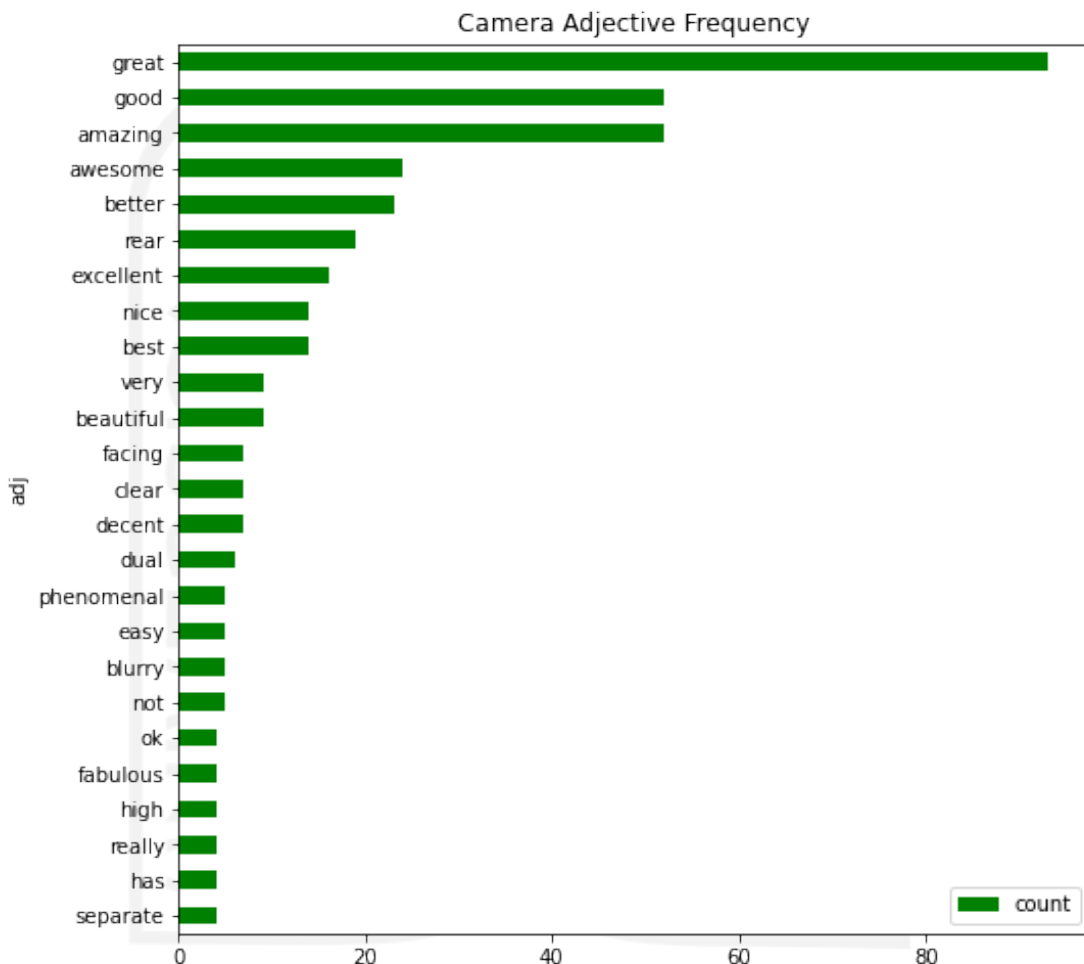


Figure 5. 5 Camera Adjective Frequency Chart

From Figure 5.5, we can infer for the camera feature, most user mentioned that the camera is great, amazing, good, and awesome. For example, the review stated that it delivers a great color with a good quality. That shows positive sentiment for the camera feature.

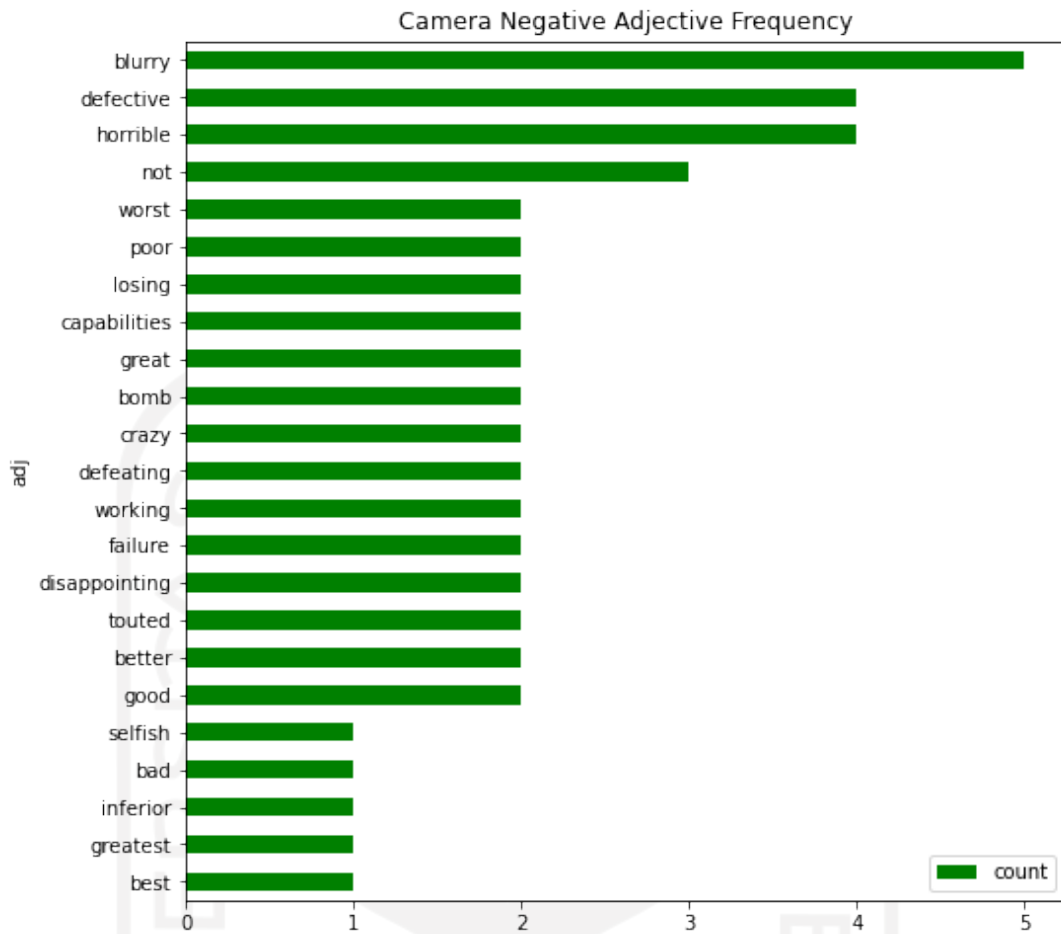


Figure 5. 6 Frequency Chart for Camera Negative Adjective

From Figure 5.6, it can be seen from the negative words identified in the camera feature, that the users complain about blurry result and defective camera, horrible camera quality. We may provide various recommendations to manufacturers based on our investigation to improve camera features in the future batch. The advice is to retain the camera function in excellent quality because the customer is satisfied with it; nevertheless, there are various things that the manufacturer should consider, including the fact that some cameras are of low quality, and some users claim that the camera is blurry.

5.1.3 Charger Feature

The third most frequent feature is the charger feature. The charger is an important feature since every phone has to charge up the battery. Around 226 words were found in the reviews, mostly the user like the charger feature such as wireless, original, and fast.

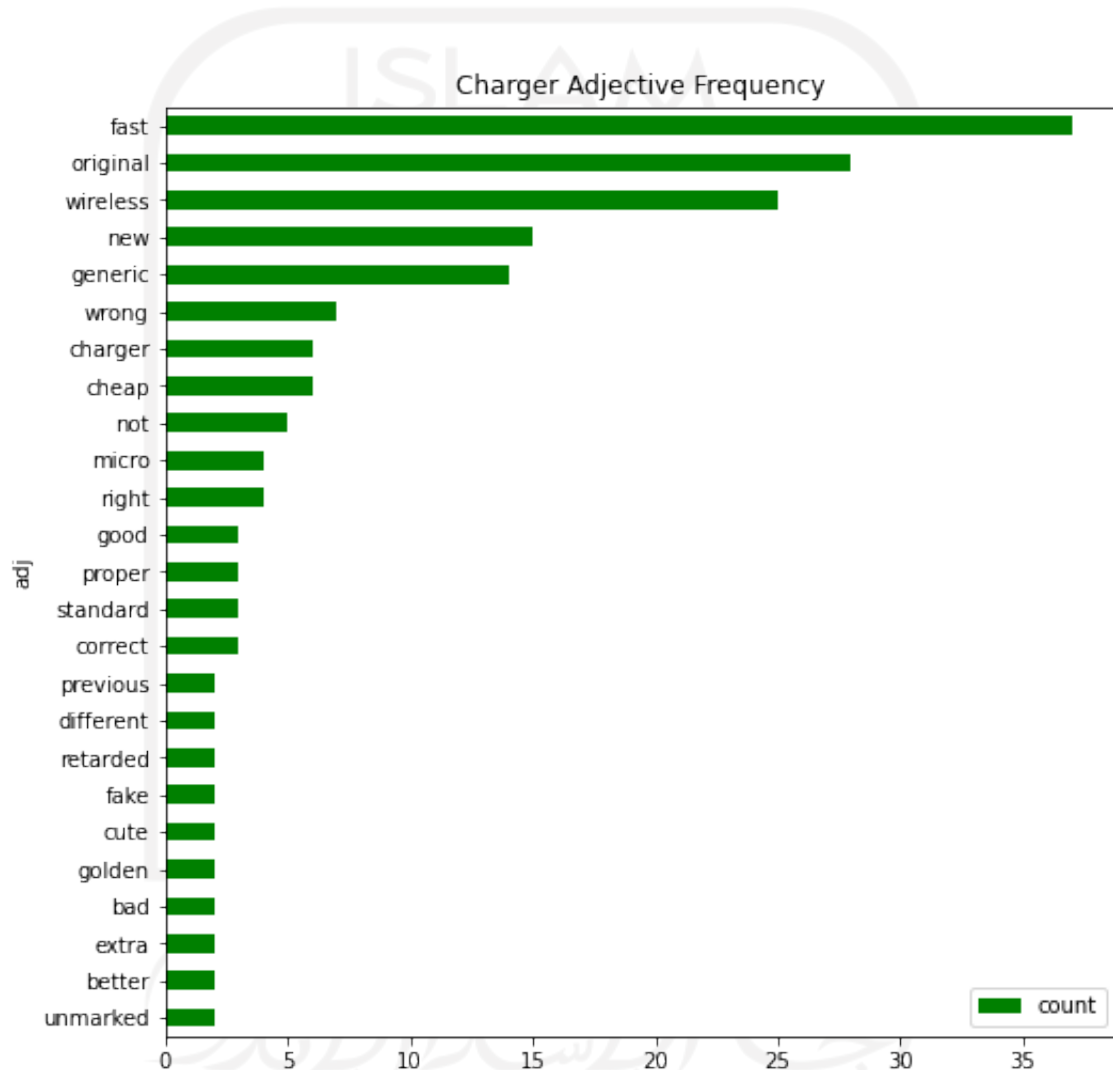


Figure 5. 7 Charger Adjective Frequency Chart

From Figure 5.7, we can conclude that the charger feature in galaxy s9 has good quality, fast charging, and have wireless charging ability. That is why most of the user give positive sentiments about the camera in galaxy s9. For more exploration, the researcher will find the negative word that can be found in the charger feature.

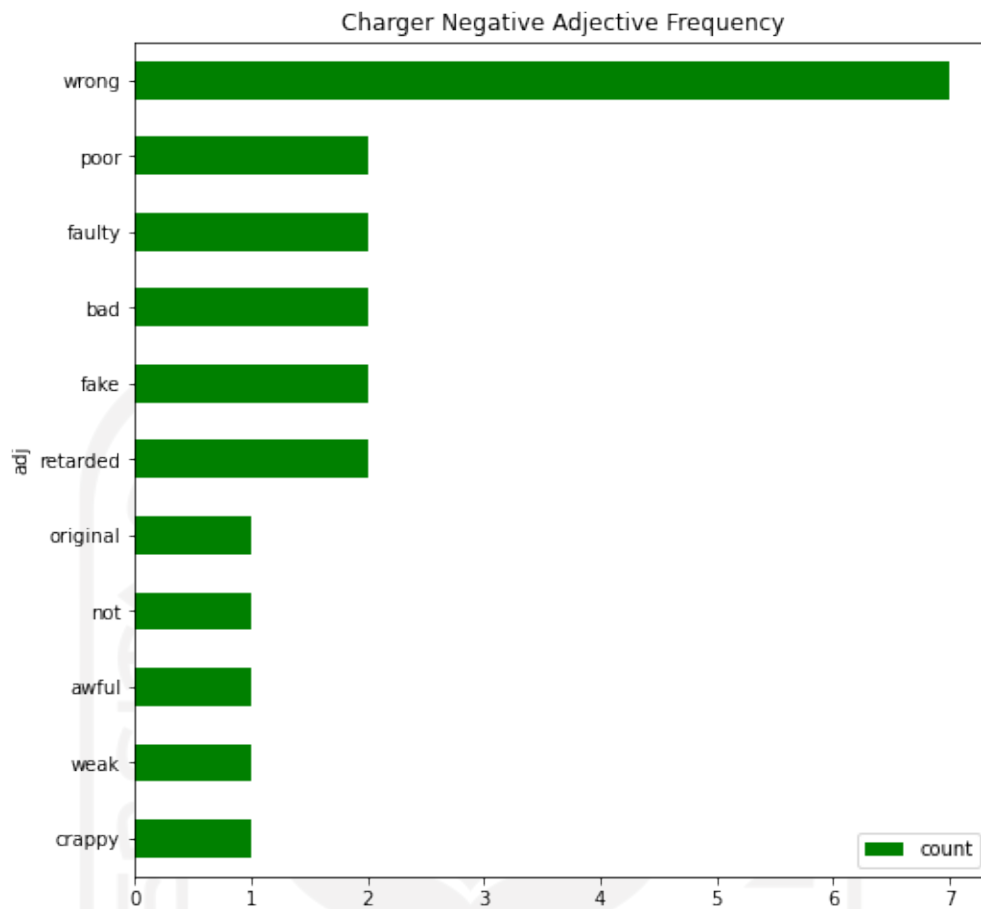


Figure 5. 8 Charger Negative Adjective Frequency

From Figure 5.8, we can conclude that some users complain about the charger, the users stated that the charger is wrong and faulty, and they get a poor quality charger for their smartphone. From those negative and positive adjectives, therecommendations can be resumed to the manufacturers, to keep this charger feature and fix some aspects that objected to user complaints. From negative adjectives, there are 7 users complaint related to the wrong charger. This may cause the retailer sells refurbished phones.

5.1.4 Battery Feature

The feature that the researcher decides to explore is the battery feature of Samsung galaxy s9. For the battery feature, we found around 294 words. From the chart below we can get insight that customer has a positive sentiment for the battery feature. That indicated from user stated good, long, fast, great, and quickly in review.

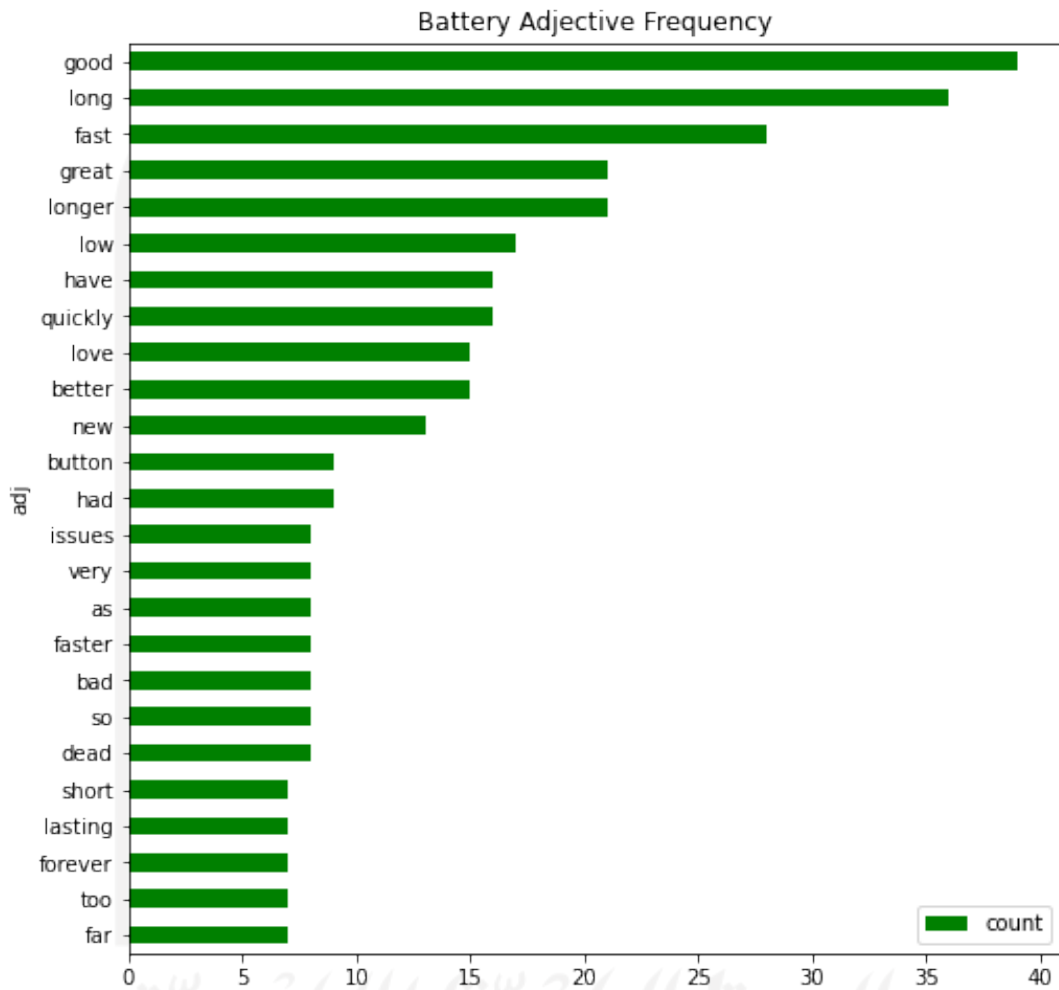


Figure 5. 9 Battery Adjective Frequency Chart

From Figure 5.9 above, we can get insight that the battery has good quality, long, fast and quickly charging this is related to the charger feature, long and longer indicated that the battery has good durability. That feature must keep existing in the next production because the user was happy about that. The product improvement is made by considering the negative review in the Samsung galaxy s9.

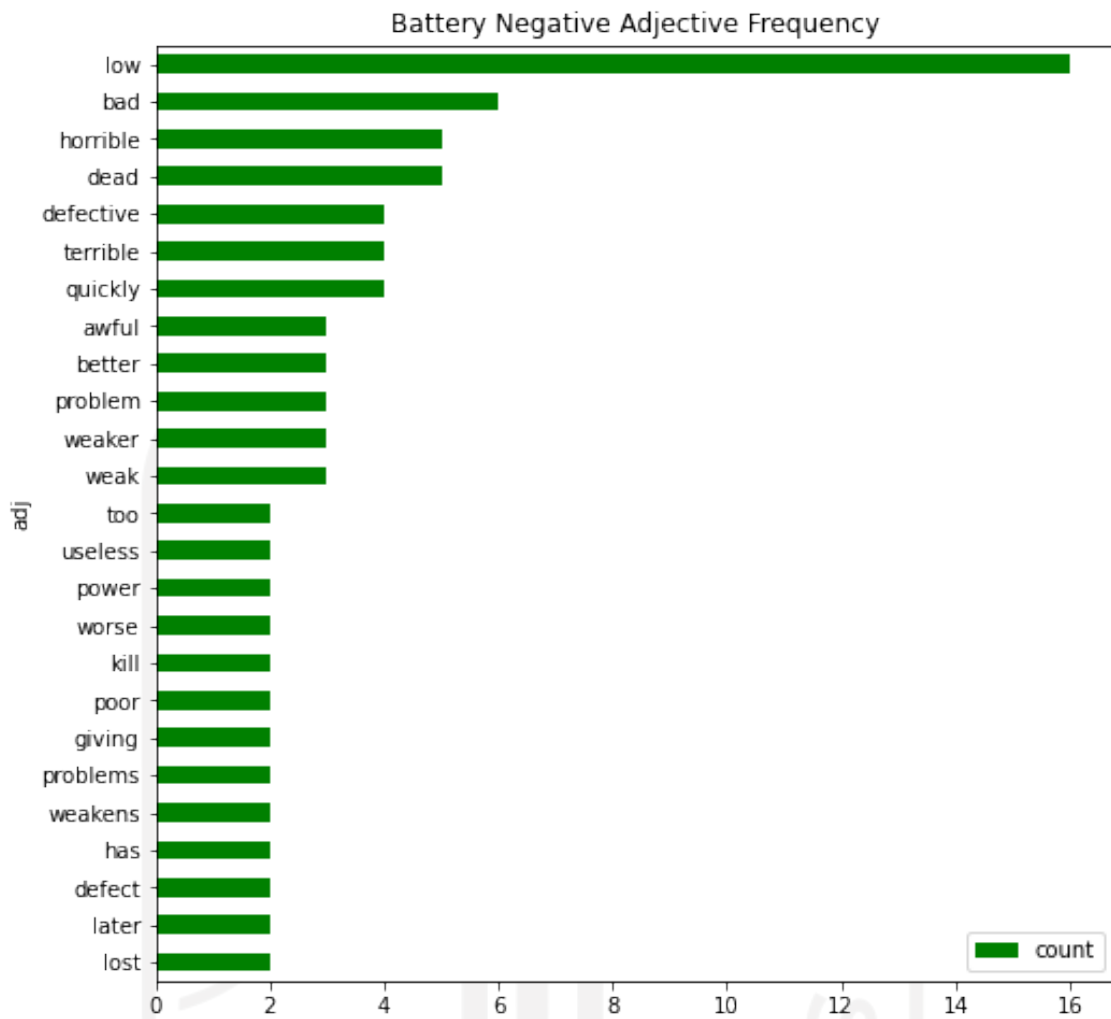


Figure 5. 10 Battery Negative Adjective Frequency

From Figure 5.10, we can conclude that most users complain about low battery quality, bad horrible dead battery quality, and defective battery. All these words indicate that the user get poor quality battery. The second word is quickly which translated as the user complains about the battery that quickly drained. So, after finding the positive and negative adjectives from review, we can draw the conclusion that several sub-feature must be kept and some sub-feature need to be improved. So, the researcher would suggest improving battery duration and must have efficient usage of battery. From the positive side, the researcher suggests keeping the fast charging feature.

5.1.5 Speaker Feature

Speaker feature is an aspect that chosen from cluster, there is similarity word around speaker, sound, and voice, so for this feature, we decided to merge that into one data frame. From the reviews, most of the user says that the speaker of Samsung galaxy s9 is good and great in quality, has loud and clear sound, some users also emerge that the speaker is better than other phones.

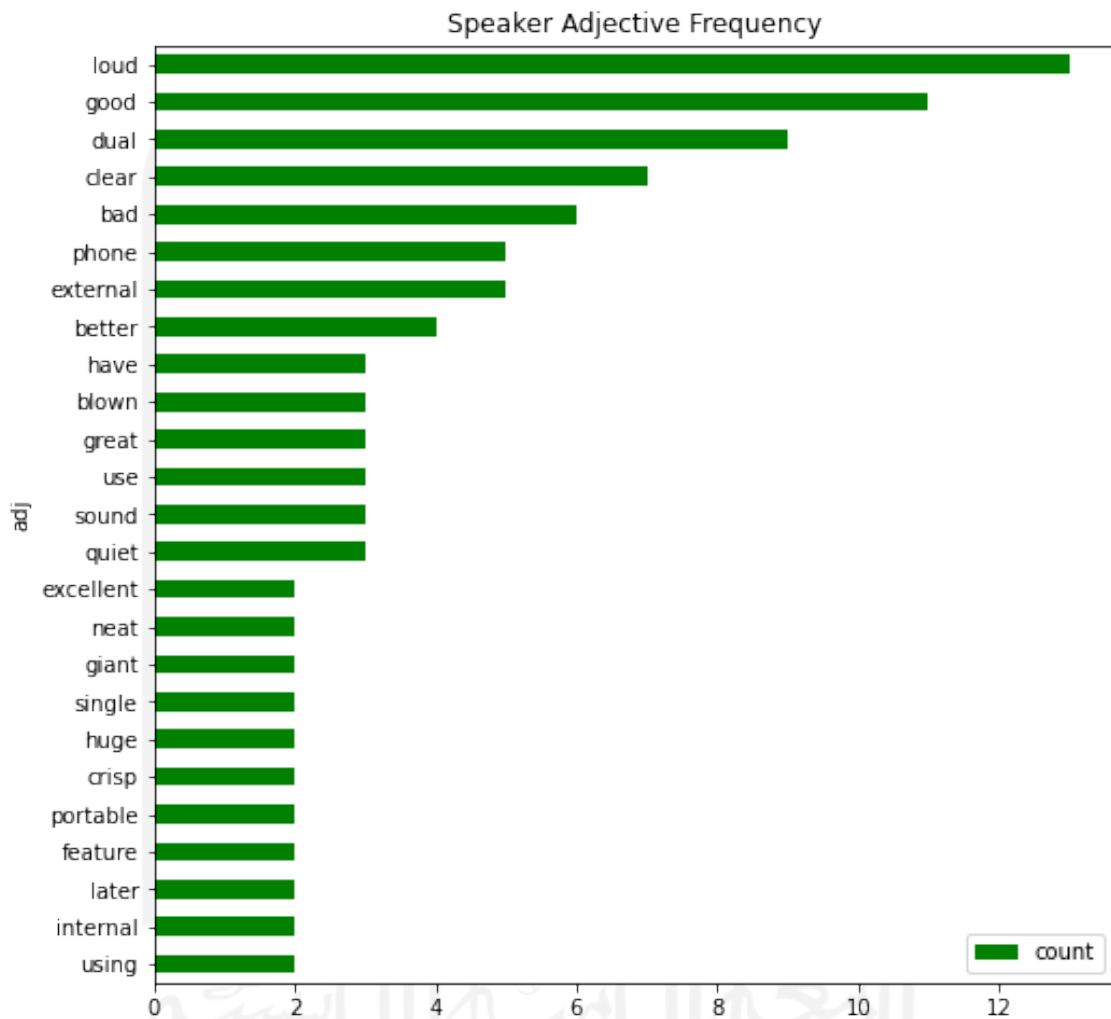


Figure 5. 11 Speaker Adjective Frequency

In process of extracting Kansei word, we must see both the positive and the negative words to get the Kansei word. From figure 5.11, we can conclude that the speaker has some adjectives that represent the speaker that generates loud and clear sound, and that characteristic makes the user happy in using the product.

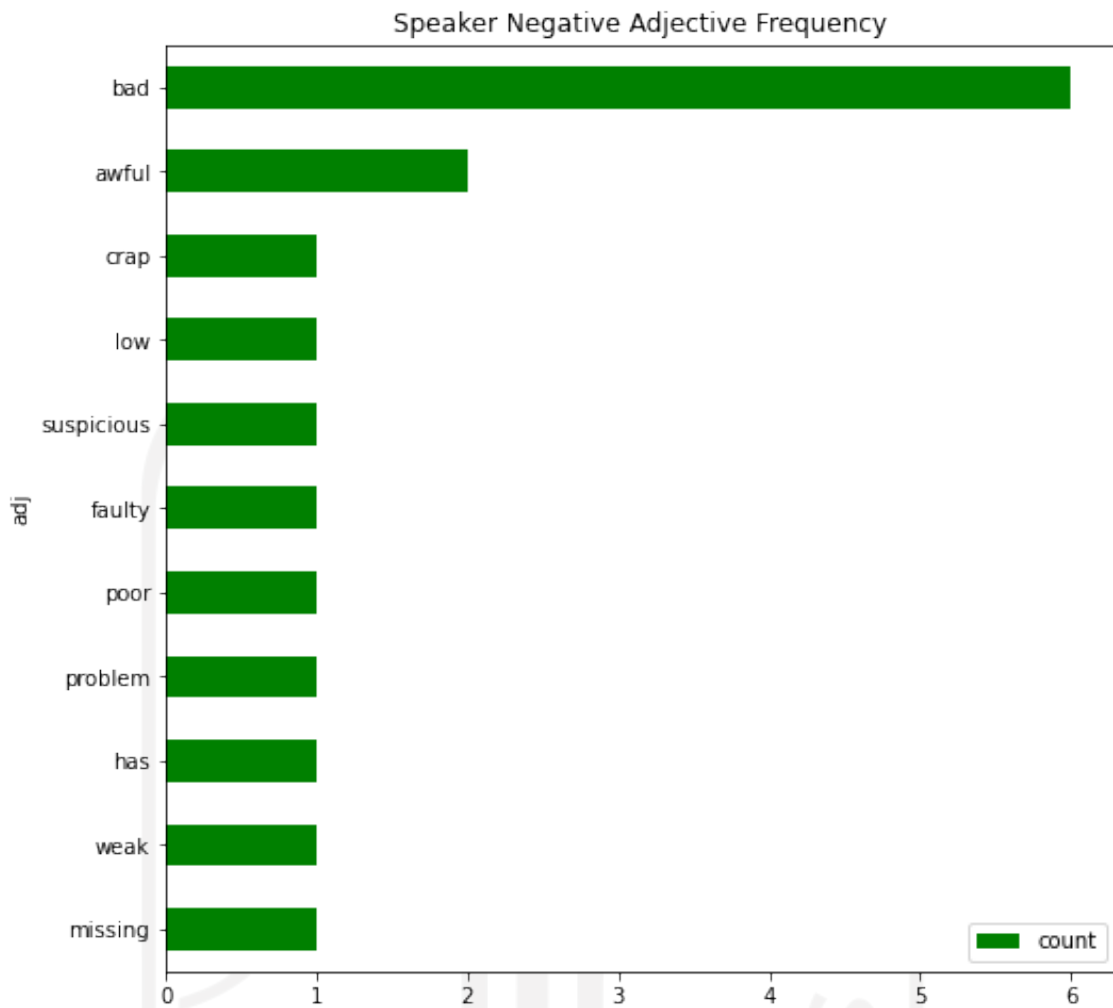


Figure 5. 12 Speaker Negative Adjective Frequency

From Figure 5.12, it can be revealed that several bad adjectives were founded for speaker, such as bad, awful, crap, low, etc. The bad and awful words in review related to the bad quality of speaker.

So, for sound feature, there are several recommendations that can be used in guidelines for product improvement. We recommend keeping the speaker clear and loud sound characteristic. Additionally, some users mention dual speaker is a good feature for smartphones. From bad reviews, we conclude that the manufacturer must take seriously about quality control of speaker because some users complains about the speaker in their Samsung galaxy s9 is in bad quality.

5.2 Product Improvement Guideline

This guideline will be resulted from the analysis feature above. The Samsung Galaxy S9 product improvement guidelines can be proposed and summarized.

Tabel 5. 2 Product Improvement Guidelines

No	Feature	Guidelines
1	Display	Curved Design and beautiful design. improve screen quality and durability.
2	Camera	The camera must have good colour sharpness, can take a video and picture with sharp and good quality, fix the blurry camera, and improve camera quality. Some users found the defective camera.
3	Charger	The charger must have fast charging ability, wireless charging ability. Increase quality control of the product, some users found the wrong charger, and charger faulty.
4	Battery	The battery must have a fast charging feature, increase battery duration, fix battery efficiency, and battery quality.
5	Speakers	Speaker must have a loud and clear sound characteristic, has a dual speaker. Improve speaker device quality

CHAPTER VI

Conclusion and Suggestion

6.1 Conclusion

After completion of the step in this research, a conclusion could be drawn. Based on the research results. There is the purpose of this paper to develop procedures to conduct aspect-based sentiment analysis for Kansei engineering. There are some advantages for using aspect-based sentiment analysis for conducting Kansei engineering, including the ability to analyse more data, increasing time efficiency. Compared to conventional Kansei engineering, this framework can process larger data and time-consuming operations by manpower. The purposed framework is started from selecting the product domain until the analysis of result. The spacy framework is employed to analyse the structure of phrases in this study. Part of Speech (POS) is used to determine the noun, adjective, and numerous modifiers in the phrases. For sentiment analysis, NLTK Vader Sentiment is used, and the compound result is shown to determine whether the sentiment is positive or negative. The majority of Samsung galaxy s9 users have a good sentiment based on the rating and polarity of the adjective word. The researcher obtained 9,894 user reviews from amazon using the python web scraping approach. The data was taken from three ASINs: B07C5ZZXDG, B07VYTLC6Q, and B079H6RLKQ. From the result in this research, the researcher extracts 5 features that come from the review. This research is compared with previous research that extracts features from GSMarena, and the result shows 5 similar features between this research and the previous research. So, the 5 features are display, camera, battery, charger, and speaker.

6.2 Suggestion

Based on the research result, the researcher provides several recommendations:

1. The next researcher can develop more specific rules to extract nouns and adjectives for more accuracy.
2. Future research can develop better data cleaning which can improve the data quality to get better accuracy.



REFERENCES

- Anil Zende, M., Bhaskar Tuplondhe, M., Baban Walunj, S., & Vasudev Parulekar, S. (2016). *TEXT MINING USING PYTHON* (Issue 3).
- Barry, J. (2017). Sentiment analysis of online reviews using bag-of-words and LSTM approaches. *CEUR Workshop Proceedings, 2086*, 272–274.
- Bello Garcés, S. (2018). Ideas previas y cambio conceptual. *Educación Química, 15*(3), 210. <https://doi.org/10.22201/fq.18708404e.2004.3.66178>
- Jiao, Y., & Qu, Q. X. (2019). A proposal for Kansei knowledge extraction method based on natural language processing technology and online product reviews. *Computers in Industry, 108*, 1–11. <https://doi.org/10.1016/j.compind.2019.02.011>
- Nagamachi, M. (2002). Kansei engineering as a powerful consumer-oriented technology for product development. In *Applied Ergonomics* (Vol. 33).
- Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow, 2nd Edition* (2nd ed.). Packt Publishing.
- Sammons, M., Christodoulopoulos, C., Kordjamshidi, P., Khashabi, D., Srikumar, V., Vijayakumar, P., Bokhari, M., Wu, X., & Roth, D. (2016). Feature Extraction for NLP, Simplified. *Proceedings of the 10th International Conference on Language Resources and Evaluation*, 4085–4092.
- Schütte, S. T. W., Eklund, J., Axelsson, J. R. C., & Nagamachi, M. (2004). Concepts, methods and tools in Kansei engineering. *Theoretical Issues in Ergonomics Science, 5*(3), 214–231. <https://doi.org/10.1080/1463922021000049980>
- Tan, W., Wang, X., & Xu, X. (n.d.). *Amazon Reviews for Sentiment Analysis | Kaggle*. 3–7.
- Zhang, L., Hua, K., Wang, H., Qian, G., & Zheng, L. (2014). Sentiment analysis on reviews of mobile users. *Procedia Computer Science, 34*, 458–465. <https://doi.org/10.1016/j.procs.2014.07.013>
- Al Amrani, Y., Lazaar, M., & El Kadirp, K. E. (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science, 127*, 511–520. <https://doi.org/10.1016/j.procs.2018.01.150>
- Gurini, D. F., Gasparetti, F., Micarelli, A., & Sansonetti, G. (2013). A sentiment-based approach to twitter user recommendation. *CEUR Workshop Proceedings, 1066*, 1–4.
- Zhang, Y., & Lin, Z. (2018). Predicting the helpfulness of online product reviews: A multilingual approach. *Electronic Commerce Research and Applications, 27*, 1–10. <https://doi.org/10.1016/j.elerap.2017.10.008>

Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, and Kuksa P. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 2011.

Python Deep Learning Second Edition, Ivan Vasilev, Daniel Slater, Gianmario Spacagna, Peter Roelants, Valentino Zocca. 2019.

Deep Learning with Keras. Antonio Gulli, Sujit Pal. 2017

C. Llinares, A.F. Page, Kano's model in Kansei Engineering to evaluate subjective real estate consumer preferences, *Int. J. Ind. Ergon.* 41 (3) (2011)

