Focusen: Real Time Sentiment Analysis to Understand Consumer Behavior

By

ALAVI BEEN AZAM

TP041230

UC3F2005IS

A project submitted in partial fulfillment of the requirements of Asia Pacific University of Technology and Innovation for the degree of

BSc (Hons) in Intelligent Systems

Supervised by Mr. RAHEEM MAFAS

2nd Marker: Dr. WADDAH WAHEED HASSAN SAEED

December-2020

Acknowledgement

The researcher would like to express my greatest appreciation and sincere gratitude towards all of those who supported me in different ways to complete this report. First and foremost, the researcher would like to thank his supervisor Mr. Raheem Mafas for the constant guidance, support and encouragement for completing this report. The completion of this report would have been difficult without his suggestions and feedback.

The researcher would also like to extend his deepest gratitude towards his friends and peers who provided guidance and emotional support throughout the entire duration of the project. The researcher is grateful for their constant motivation, feedback and constructive criticism whenever required.

Also, most importantly, the researcher would like to thank his family and relatives for their love, constant support, motivation and guidance. None of this would have been possible without their hard work and sacrifices to provide us with the best of everything.

The researcher would also like to extend his humble gratitude towards Mr. Ashrique Thevendran and his team for their precious time, mentorship and guidance. I would sincerely like to thank all of you for sharing your valuable knowledge and experiences with the researcher and helping him in every step of the way to complete this project successfully.

Sincerely,
Alavi Been Azam

Table of Contents

1.0 Introduction to the Study

1.1 Background of the Project

It is increasingly becoming more important for businesses to understand their consumers in order to stay relevant in the market and generate profits by understanding their behavior to provide the right products and services according to their wants and needs. According to a salesforce report from 2019, 73% of consumers expect companies to understand their needs and expectations while 84% say that the experience a company provides is as important as its products and services (Donegan, 2019). This directly translates to the fact that if a consumer isn't happy with a brand, they will move to a competitor alternative. Most modern successful companies take this into account and make their product development and marketing decisions based on insights generated from data collected from their consumers through various mediums. This is backed by the stat provided by Qualtrics that shows organizations that lead in customer experience outperformed laggards on the S&P 500 index by 80%. Consumer behaviour is described as a psychologically-study of how individuals make purchase decisions, it is understanding of what motivates individuals to purchase a specific product or service (DJ Team, 2020). Several key characteristics are classified under the study of consumer behaviour such as how a consumer feels about certain brands, products and services, what motivates a consumer to choose one product over the others and why, what factors in a consumer's everyday environment affect their buying decisions or brand perception and how consumers make decisions in groups or when they are alone. Multiple factors are needed to be taken into consideration to understand the characteristics and determine consumer behaviour such as social factors, psychological factors, economic factors and even simple personal traits.

Traditional approaches for studying and understanding consumer behaviour such as focus groups and marketing surveys and social-media monitoring have been popular for a long time but require a lot of time and resources to generate actionable insights. Several organizations including market research and analytics companies are now conducting surveys and focus groups online and also trying to developing tools and service to monitor social media constantly in order to generate consumer insights and understand their behavioural traits. Even though moving things online and technology have made the process of gathering data relatively faster with significantly less resources but the actual process of combing through the data to clean,

organize and analyse it to generate actionable insights has still remain a tedious process which requires a high level of statistical and technical expertise along with domain knowledge.

Several advancements in Artificial Intelligence such as machine learning has provided individuals and organizations with several tools and techniques in order to make faster and more accurate computations to generate insights automatically with minimal to no human supervision. One of these techniques is known as sentiment analysis. Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a topic, product, etc. is positive, negative, or neutral. To identify the underlying tone of the expression, sentiment analysis uses natural language processing (NLP). Different types of sentiment analysis use different methods to identify the undertone of a phrase, it is mainly classified into two major types, subjective/objective identification which classifies a sentence or a fragment of text into one of the two categories and feature/aspect-based sentiment analysis which allows for the determination of different opinions or features relating to different aspects of an entity which provided a more defined overview of opinions and feelings (Sims, 2015). The rise of social media sites like Facebook, twitter and Instagram and every growing popularity of online shopping platform with ratings, reviews and recommendations systems, companies are becoming increasingly interested in the sentiment analysis to extract valuable information from the vast amount of unstructured data available online. Organizations globally are using different sentiment analysis techniques to better understand consumer behaviour by extract valuable insights from various communications channels to take more effective and customer focused actions. This project aims to use sentiment analysis in order to generate reliable and actionable insight from online focus group conversations, marketing surveys and social-media platforms with quick turnaround time and minimal human input.

1.2 Problem Context

The problem area includes building an ETL (extract, transform & load) data pipeline which will take text data from online focus group conversations, surveys and social-media as input. Next the data will be passed through a sentiment analysis-based machine learning model which will evaluate the data to determine the sentiment of the text, its polarity and other text based analytic to generate a report which will easily be available on a web-based dashboard to provide insight into the consumer sentiment and help drive more accurate data-driven decisions. The

purpose of this research is also to develop a sentiment analysis model and insights dashboard that can be used to tag and generate insights from online focus group conversations, survey responses and social-media in real time to provide users with a detailed report based on the text input fed into the model in a web-based dashboard. Some of the problem areas that the researcher will tackle in this research has been highlighted below:

1. **It is resource intensive and time consuming to manually analyze the data and extract insights from text data in online focus groups, open-ended survey responses and social media data sources.**

The process of analyzing data manually to generate insights after qualitative text data has been generated from online focus group, survey responses and social media scrapping is heavily labor intensive and time consuming. The process often takes days to weeks and even months for researchers in the organization to manually comb through the data to discover and report relevant insights. For research companies who are constantly conducting focus groups and survey this process can be highly tedious and in-efficient (Vase, 2020).

2. **Ruled-based systems lack optimal accuracy and are limited.**

In some organizations, often some rule-based classifiers are used to classify the qualitative text data to generate insights faster, but these rule-based systems have limited capabilities and can produce inconsistent results with low accuracy due to its hard coded nature and inability to learn from historical data, making the insights generated from the data unreliable (Vase, 2020).

3. **Data analyzed manually can be in-consistent and be prone to human bias.**

Humans often get tired, disagree with one another and can have different ways of interpreting things based on personality and other psychological factors. So, it highly difficult for individuals and teams to tag text responses generated from surveys and focus groups consistent without fluctuations or biases which can result in inconsistent and biased insights may not be totally accurate towards understanding consumer behavior (Vase, 2020).

4. **Real-time analysis and insights are not possible.**

It is not possible for businesses to receive instant real-time feedback from their consumers using manual analysis and classification which can lead to missed opportunities and quick

testing of ideas and hypothesis which may generate actionable insights for swift business decisions (Vase, 2020).

5. **Lack of scalability.**

Market research firms and other research entities often produce thousands of text-based responses from surveys and focus groups every week, it is highly unrealistic for researchers to analyze such high volumes of data manually to generate reports on consumer behavior. Organizations need to dedicate extensive human resources to tackle this issue which can be dedicated to other human-dependent task if the process is automated which translates to lower operational costs making market research significantly cheaper and more accessible for companies (Vase, 2020).

1.3 Rationale

All of the issues highlighted above in the project background shows the importance of consumer sentiment for brands and why it required to generate reliable insights fast which can influence critical products, services and business decisions for organizations. Therefore, marketing research teams and companies need to find a way to generate reliable and actionable insights fast instead of relying on human experts who has to manually comb through the data to classify responses to produce reports or rule-based systems which have limited capabilities and can portray a significant amount of bias in the insight generated. Also, it is resource incentive to find experts who have strong knowledge in several domains and it also impossible to build rule-based system from every use-case since these are not intelligent and adaptive system.

By implementing an AI based sentiment analysis model to tag and categories online-focus group conversations, survey responses and data scrapped from social media, trained with a diverse dataset to account for expert knowledge in several domain areas, organisations can easily generate insights in a shorter timeframe with higher accuracy and less bias while also saving a huge amount of resources which can be deployed in other business areas or help them provide their services at a more affordable cost to get more clients . It will also help individual researchers, marketing teams and market research companies to generate more insights and take on more clients and help them make better data driven decision based on reliable insights

generated using sentiment analysis and ML (machine learning) techniques by automating the entire ETL to insight pipeline.

Also, using sentiment analysis techniques provides several other benefits such as improved customer service for companies, developing quality products, discovering new marketing strategies, improving media perception, increasing sales revenue and improving crisis management (Anon, 2019). Companies can use sentiment analysis to track key messages in consumer communications which help organizations to understand their opinion and thoughts towards a brand, this helps customer service divisions to be aware of any issues with their products and services and get ahead of them by taking swift actions based on customer feedback which can improve consumer relations. Also, keeping consumers happy and loyal towards a certain brand is a taxing job which requires huge amount of commitment for brands. Sentiment analysis techniques help brands thoroughly understand what customers truly want and can help them make better product development decisions accordingly. Brands can also process the abundance of consumer data available online to improves their marketing channels and strategies to test the sentiment towards their marketing campaigns by gaining insights from customer conversations in different medium which can also help segment and target potential customers. Another key benefit is that sentiment analysis can be used to track the understanding of the journalists, writers, columnists, market analysts, media researchers or independent contributors towards the company, be it the product, service, company values, human resources etc. This is crucial as any misinterpretation or negative connotation can lead to negative key messages which forms an undesirable perception (Anon, 2019). This shows that the scope of sentiment analysis to understand consumer behaviour lies way beyond focus-groups, survey responses and social media monitoring which are the researchers primary focus for this project.

1.4 Potential Benefits

Potential benefits of a project can be divided into two distinct categories, tangible and intangible benefits. Tangible benefits are quantifiable and can be measured while intangible benefits are harder to measure and are highly subjective (Capozzi, 2017). This section below highlights the potential tangible and intangible benefits of this project.

1.4.1 Tangible Benefits

Firstly, analytics on the data collected from various sources such as online focus-groups, surveys and social media can be carried out significantly faster by automating the entire process. The system will have an ETL pipeline which will take clean data as an input and all the pre-processing steps while be carried out automatically using python scripts without any human supervision. The pre-processed data will then be fed into a sentiment analysis-based machine learning model which will generate a detailed analytics report in a few minutes in contrast to what will take researchers days or even weeks to process manually based on the size of the input data. This will help researcher and marketing teams conduct more studies as the turnaround time from data to insights will be cut down significantly. Secondly, using a ML-based sentiment analysis model will help generate more accurate insights which will help organizations make better data-driven decisions faster helping them maximize profits and consumer satisfaction. The capability to generate insights in real time will provide researchers and companies to make time sensitive decision faster which can really help organizations cut down losses if a negative sentiment is detected from their consumers and swift action can be taken to ensure customer satisfaction and maintain profit margins. Also, having an automated ML (machine learning) model makes the system more scalable and adaptive as the model can be re-trained whenever required by with the correct dataset to enhance its knowledge domain and improve the accuracy of the system which can help generate better insights with broader domain expertise compared to human researcher who can only have a certain amount of domain knowledge. Most importantly, having an automated system to generate insights from consumer data will help significantly reduce operational expenditure for individual researchers, market research companies and teams as less manpower is required to process the data collected which will help companies conduct more research operations at significantly lower costs as they do not need to hire as many researchers to comb the data and produce reports manually.

1.4.2 Intangible Benefits

Apart from the direct tangible benefits of the project, there are several subjective intangible benefits as well. Firstly, having an automated sentiment analysis based system to generate insights from data collected will ensure that the reports and insights are delivered to the clients/stakeholders way faster which will ensure a high level of client satisfaction for market research companies and also for other stakeholders and decision makers involved in the

process. Also, having an automated system will provide more consistent results and reduce biases influenced by human researchers as it is very common for human to have certain biases which can negatively impact the quality of the insight generated manually.

1.5 Target Users

The primary target users for this system would be market research and validation companies who need to collect and rely on data from consumers using various sources such as surveys, focus-groups and social media scrapping on a regular basis, this system will be able to assist them in generating reports faster and providing more accurate insights in almost real-time which will help them improve their workflow by using sentiment analysis techniques to analyse, tag and classify text data responses from structured online focus group conversations and open-ended survey responses. This will allow them to significantly lower operational costs and help them take on more clients to maximize profits. The secondary target users for this system would be individual researchers and marketing teams within organizations who also want to conduct similar research on consumer behaviour but do not necessarily have the amount of resources required to run such operations, this system will help them easily generate reports from their data to gain insights faster with significantly lower cost and resources, which will help them make better product and business decisions with more accurate and reliable insights generated using Focusen to ensure high level of consumer satisfaction.

1.6 Scope and Objectives

The scope of this project is to develop an end-to-end cloud-based sentiment analysis system (SaaS) which will ingest clean text data from various sources such as online focus group conversations, survey responses and social media scrappers in real-time as input using an ETL (extract, transform, load) data pipeline the data will then be pre-processed, cleaned and structured. Then the structured data will be feed into a sentiment analysis based pre-trained machine learning (ML) model which will generate a predictive analytics report on the input text data containing details about its sentiment classification, polarity score, keyword identification, geographic segmentation and other relevant insights in real-time, which will help researchers better understand the overall consumer sentiment from the consumer data collected. To demonstrate this the researcher will be using the Twitter API to scrap tweets in real-time, store it in a relational database, pre-process the data using the appropriate libraries

in python and use a pre-trained Textblob sentiment analysis model to generate insight and display them on a web-based dashboard.

The application will primarily be deployed on local-host to save cost but has the potential setup to deployed on cloud services such as Amazon Web Services (AWS) and Heroku. It will provide a detailed analytics dashboard with data visualizations and an interactive graphical interface to make it interactive and user-friendly for end users to better understand and interact with the data.

1.6.1 Aim

The aim of this research is to design, develop and deploy an end-to-end web-based sentiment analysis application (Focusen) that will take clean text data from various sources such a social-media, online survey responses and focus group conversations as input in real-time into a pre-trained machine learning model and generate a predictive analytics report on web-based dashboard to help researchers, marketers and organizations to understand consumer behaviour towards a certain brand or a topic in real-time.

1.6.2 Objectives

The objectives for this project are:

- Implement a real-time data ingestion pipeline using the Twitter API as an example to feed data in to the ML model.
- Develop to ETL pipeline to process the data and feed it into a pre-trained sentiment analysis-based ML (machine learning) model to generate insights in real-time.
- To understand the different aspects of text-based analytics to provide an insightful sentiment analysis report generated from the input data to understand the different factor affecting consumer behaviour.
- To develop and deploy a web-based dashboard that will display full predictive analytics report of the input data with interactive data visualizations for end-users as a packaged end-to-end sentiment analysis system (SaaS).

1.6.3 Deliverables

The deliverables of this project will be a full end to end web-based sentiment analysis application which will include but are not limited to a data ingestion pipeline to process the

input data, a ML model to evaluate the data to generate insights and a web-based dashboard providing a detailed report on the output of the model and data analytics in real-time showing a detailed breakdown on the sentiment report, data visualizations to make the report more interactive and easier to use. The researcher will work in conjunction with a popular market research company to understand the requirements better and develop a working solution for the problem stated.

- A data ingestion pipeline to ingest tweets (Twitter data) in real-time using the Twitter API according to the keyword provided as input.
- An ETL data pipeline to pre-process and structure the data before being feed to the ML model in real-time.
- A machine learning (ML) model capable of predicting the sentiment of the input text to determine consumer sentiment.
- A web-based dashboard to display the results from the model in the form of an interactive predictive analysis report in real-time with visualizations to understand the sentiment from the input data.
- An end-to-end web-based sentiment analysis system (Focusen) to gain actionable insights from various data sources such as social media, online focus groups and surveys in real-time.

1.6.4 Nature of the Challenge

The nature of the challenge is to develop a web-based sentiment analysis system which can adapt to data from several different industries and domain areas since most data analytics and market research companies work with a diverse set of clients who require insights to be generated in a variety of different fields. Another critical challenges would be the accuracy of the system, as the insights are generated in real-time from Tweets which are also posted and collected in real-time, the researcher cannot hyper-tune the parameters of model to provide the most accurate insights based on the keywords defined. Similarly, another massive challenge would be the localization of the system, since the ML model is trained to only handle a handful of languages, any data the system ingest which are outside the language parameters of the model will corrupt the insights generated.

1.7 Overview of the Report

This report is divided in several sections. The first part of the report provides an introduction to the research project, it contains the background of the project which clearly highlights the importance for companies to understand consumer behaviour and how data driven decisions can make product and services a success by maximizing profit by listening to customers and understanding their sentiments. The section also draws on several technological advancements such machine learning and sentiment analysis can be used to tackle the resource intensive task of carrying out consumer research and generating insights from the vast amount of data. The background is followed by the problem context section which discusses the key problem the researcher intends to solve through this project. In the next part of the report, the researcher discusses about the rationale of the project along with its tangible and intangible benefits. The introduction section also contains details about the target users of the project, its scope, objectives, deliverables and the nature of challenges the researcher will most likely face while continuing with this project.

The next section of the report focuses on the literature review being carried out for the project. This section investigates domain areas of consumer behaviour, machine learning, sentiment analysis and cloud computing by drawing upon previous work conducted by other researcher in the concerned fields. The researcher reviews extensive literature through online peer review of journals and other relevant sources to determine the importance of consumer research, how machine learning and sentiment analysis can be used to generate more accurate and actionable insights faster, the best frameworks and tools available and other relevant resources that will assist in completing this project successfully. In the later part of this section the researcher all looks at similar system or solutions that are currently available in the market or have been proposed by peers in the field.

The third section of the report technical research that will be carried out for this research. The component contains details about the tools, frameworks and libraries that will used in this project and will compare the pros and cons of the different options available. It will provide a detailed look at the technical tools that will be used in this project.

The next section of the report will discuss the system development methodology that will be used for his project according to the design and implementation requirements and will justify

with reasoning why it has been chosen for this project. The component will also compare the chosen methodologies for each component of the project regarding software engineering and data analytics with other suitable candidates to draw down on the advantages and disadvantages of the options available.

In the following section the researcher will discuss about the research methods that has been used to gather requirements for this project and justify the chosen method followed by the analysis of data collected to validate and justify the requirements.

In the final section of the report, the researcher will thoroughly discuss the architecture of the system including the abstract architecture, database design and also the interface design. Following the system architecture section will be the project plan which highlight the release plan of the project according to its set functional and non-functional requirements and also discuss in details about the test plan that will be used to validate the functionalities of the of the proposed system. In the next component, the paper will provide a detailed walk-through on the implementation of the proposed system along with screenshots and descriptions. The section will also contain details about the code written to implemented some of the mission critical components of the application. For the last few chapters of the report, the paper will provide a detailed description and results of the system validation conducted to validate the functionalities and requirements of the application followed a conclusion and some reflections by the author in regards to this project.

1.8 Project Plan

This section will briefly discuss the overall timeline and plan for this project. A detailed breakdown of all the tasks that been carried out throughout the entire duration of this research project along with a project timeline in the form of a Gantt chart has been attached in the appendix section.

2.0 Literature Review

2.1 Introduction

This section will contain the details of the literature review that has been carried out for this project. The researcher has peer-reviewed several journals available online and other resources in fields of market research techniques, the importance of consumer behaviour, machine learning, sentiment analysis, sentiment analysis to understand consumer behaviour, data pipelines and cloud computing. The later section will contain research in similar system available in the market regarding the project and discuss the pros and cons of each option available.

2.2 Domain Research

2.2.1 Market Research Techniques & Importance

According to the MYMG Team (2011) market research is described as a continuous process of collecting, investigating and interpreting information about a particular market a company operates in or a product/service the company offers for selling in that market, and also about potential and existing competitors and the past, present and potential customers who purchase and consume the offered product/service. Conducting a market research translates to the effort of researching and analysing all the information in regards to a product/service, customers and competitors to gain insights on how a company can establish a hold in the market and appeal to its target users and gain a competitive advantage. As also mentioned by the MYMG team (2011) market research help business achieves increased sale, better customer management and business growth.

An article by Brandwatch (2019) discusses the essential methods for market research methods. There are mainly two different kinds of market research primary (field) research and secondary (desk) research. Primary research is the type of research that has been carried out directly by an individual researcher or an organization in regards to the concerned topic, whereas secondary research is the kinds where the researcher leverages on research carried out by other researcher an organizations in similar topics. The most essentials market research methods are focus groups, surveys, social media listening, interviews, experiments and field trails, observation, competitive analysis, exploring public domain data and buying research.

A research conducted by Witell (2019) shows that co-creation marketing research techniques has strong relation with profit margins. Another research by Javalgi et al., (2006) explores and discuss the importance of market research and market orientation regarding customer relationship management. The paper mentions that marketing research is a key mechanism through which companies can understand their current as well as potential customers. Through several anecdotal and case examples Javalgi et al., (2006) has illustrated the essential linkage between market research, market orientation and customer relationship management (CRM).

2.2.2 The importance of understanding consumer behaviour

According to a paper by Anderson et al. (1994) consumer satisfaction has long been recognized in marketing thought and practice as a central concept as well as am important goal for all business activities. Realization of its importance has resulted in a proliferation of research on consumer satisfaction over the past few decades (see, for example, the vast references provided by Yi, 1990). “Even a causal glance at business journals and business sections of daily newspapers reveals that the subject of customer satisfaction is receiving extraordinary attentions” (Anton, 1997). High consumer satisfaction has many benefits for the firm, such as increased consumer loyalty, enhanced firm reputation, reduced price elasticities, lower costs of future transactions, and higher employee efficiency (Anderson et al., 1994; Fornell, 1992; Swanson and Kelley, 2001). This shows how important it is for business to properly understand consumer behaviour. This drives organisation to conduct focus groups and surveys to understand customer satisfaction levels and to determine pricing and marketing strategies along with other critical business decision based on data collected.

In another journal article Statt, n.d. (2013) discuss on how consumer behaviour is at the heart of effective marketing. The primary goal for marketing is to satisfy customer needs, to marketers and businesses must have a solid understanding on what their customers want and need, this can be achieved by understand the overall consumer behaviour of customers.

2.2.3 Machine learning and Sentiment Analysis

Advancement in technology, specifically in the field of artificial intelligence has led to machine learning becoming one of the top contenders for solving problem in several domains. Machine learning tools and techniques have become hugely popular for nature language processing (NLP) and text analysis. In a research, Khan et al., (2016) mentions that machine learning (ML)

techniques are frequently used for application in natural language processing and has definite patterns, ML techniques adapt to domain specific solution at high accuracy depending upon the feature set used. In the same research Khan et al. (2016) also explores several sentiment analysis techniques and highlights the need to address domain specific nature language processing (NLP) open challenges to make further advancements in this space to solve more complex NLP problems.

According to Pang and Lillian (2008) sentiment analysis is a type of text classification that deals with subject statements. The process is also known as opinion mining, since it processes opinions to learn about public perception. Sentiment analysis uses NLP techniques to collect, examine and classify the sentiment behind words, phrases and sentences. SA is also explained as identifying the sentiment of people about a topic and its features (Pang & Lee, 2008). Khan et al. (2016) also mentions that business is constantly interested to know which features of their products and services are more popular among customers in order to make profitable business decisions.

Nowadays with the popular of the internet there is a huge repository of content-based content available are various online sources such as blogs, forums, review website, social media and etc. The repository keeps growing as more and more opinion-based content is added continuously. It is, therefore, beyond the control of manual techniques to analyse millions of reviews and to aggregate them towards a rapid and efficient decision. Sentiment analysis techniques perform this task through automated processes with minimal or no user support (Khan et al., 2016). Khan et al. (2016) also mentions that online datasets may also contain objective statements which do not contribute effectively towards sentiment analysis, so such statements are required to be filtered out during pre-processing of the data. The research also mentions that opinion mining deals with identifying opinion patterns and presenting them in a easy to understand manner. The outcome from sentiment analysis techniques can be binary classifications, such as categorizing opinions within the boundaries of recommended and can be considered a multi-class classification problem on a given scale of likeness. In similar research Cambria et al. (2013) used common sense knowledge to improve the results of sentiment analysis through which results can be presented in the form of a short summary generated from the overall analysis. Sentiment analysis contain various sub-genres which includes emotion analysis, trend analysis, bias analysis and etc (Khan et al., 2016).

SA techniques have also been used in a wide array of problems such as using sentiment analysis in emails for gender identification through emotional analysis (Mohammad and Yang, 2011) and emotion being applied to fairy tales to draw interesting patterns. (Mohammad, 2011). Pang and Lee (2008) in their research also mentions that there are several applications for sentiment analysis such as application to review-related websites, applications as sub-component technology, application in business and government intelligence and application in several other problem domains.

There are several type of machine learning (ML) techniques such as supervised, semi-supervised and unsupervised machine learning. Supervised machine learning techniques uses labelled data, while semi-supervised techniques require manual tuning from domain experts to get consistent and accurate results. On the other hand, unsupervised machine learning techniques make use of statistical analysis on large volume of data to provide output. ML techniques has a large feature set using Bag-of-words (BOW) and results are improved by pruning repetitive and low-quality features. The opinion words are extracted to identify the polarity of opinion expressed for a feature. The performance of a classifier is measured through its effectiveness at the cost of efficiency. Effectiveness is calculated as precision/recall and F-measure, which are measurements of relevance (Khan et al., 2016). Sentiment analysis is also considered to be a complex network which contains nodes and edges joining each other. There are several machine learning classifiers that can be used for sentiment analysis. There are 3 level of analysis that can be applied to textual data, document level, sentence level and word level. Sentiment analysis can widely perform at the sentence and word level analysis to obtain the best results. ML techniques suits sentiment analysis as the data is in abundance and there is obvious presence of patterns (Schouten & Frasincar, 2016). The classifiers are trained on label dataset having samples representing all classes. A test dataset is used to evaluate the performance of the classifiers for the given task. In their research Khan et al. (2016) analysed some of the most popular classifiers and techniques used in sentiment analysis, these are naïve bayes, nearest neighbour, centroid based, support vector machine, unsupervised techniques, lexicon-based techniques and corpus-based techniques. They also highlighted the several complex challenges faced document level, sentence level, feature level and lexicon level during sentiment analysis operations.

Further into their research Pang and Lee (2008) also mentioned the several challenges faced with the application which highlighted the challenges of contrasts with standard fact-based

textual analysis and factors that make opinion mining difficult. Khan et. Al. (2016) concluded that sentiment analysis has gained huge popularity in both academic and commercial applications. A clear analysis showed that even though ML techniques are a popular choice for sentiment analysis and nature language problem (NLP) problems there are several limitations and open issues that cannot be controlled yet with the current methods in practise. Because of its close relevance to NLP, sentiment analysis also faces similar challenges such as co-reference resolution, negation handling, word sense disambiguation etc. However, it is also useful to note the SA is a highly restricted NLP problem and researcher do not need to understand the semantic if each word within a sentence to make acceptable classifications. Domain-based datasets and complex network analysis has shown promising results in the problem space.

2.2.4 Sentiment Analysis to understand consumer behaviour

Research conducted by Gunter et al. (2014) highlights that rapid spread for online news and online chatter in blogs, micro-blogs, social media and several other online platforms have created a potentially rich source of public opinion. Public are constantly expressing their opinions and feelings online through these platforms which has led businesses, advertisers, governments and policy makers take a notice and have woken up to the fact that this universe of self-perpetuating human sentiment could represent a valuable source to guide business and political decisions (Gunter et al., 2014). It is near impossible to manually analysis this massive repository of data so computer sciences are constantly looking into way using different tools and techniques that can apply linguistic rules to provide electronic readings of meanings and emotions (Gunter et al., 2014).

In similar research, Troisi et al. (2018) has developed a tool called “TalkWalker: which is a big data analytics tool that uses an algorithm developed with the context of social data intelligence allowing users to understand sentiment of a group of people regarding a specific theme. The tool was used to conduct a case study on a student population to determine what influences their choice of university. The research findings showed that the factors affecting the choice were trainings offered, physical structure, work opportunities, prestige, affordability, communication, organisation and environmental sustainability (Troisi et al., 2018). This shows how such tools and system can be used for sentiment analysis to mine people opinions and understand the emotions and sentiments behind their decisions. Similar work has also been

done by Laksono et al. (2019) were the researcher used sentiment analysis techniques for restaurant customer reviews on TripAdvisor using Naïve Bayes. The research also highlights why it is important for business to understand the behaviour of their customer to ensure customer satisfaction. The research also concluded that using Naïve Bayes increased the accuracy of the model by 2.9% compared to TextBlob (Laksono et al., 2019).

Since one of the main challenges the researcher expects to face during this project is localisation as the target market for the project is Malaysia which has several different languages, the researcher also review some research conducted by Almuqren & Cristea (2016) which analysed the challenges face and developed a sentiment analysis framework for mining Arabic tweets to measure customer satisfaction toward telecom companies in Saudi Arabia. Some of the challenges the researchers faced with while carrying out semantic sentiment analysis using the Arabic language were diacritization, negation and spelling errors (Almuqren & Cristea, 2016). The peer-review conducted for similar work will act as a guidance for the researcher and help tackle some of the challenges that the project might face.

2.3 Similar Systems

This section will analyze and compare different sentiment analysis tools and systems that are already available in the market and has similar functionalities as the system proposed by the author in the research. In an online article for SocialMediaToday, Barysevich (2020) described sentiment analysis tools to be tools that interpret a text in a way that aims to reveal the intent and tone behind it, using various techniques the tool processes the data to generate a comprehensive report with charts and graphs which business can use to monitor their overall sentiment of their consumer towards their brands, products and services. For the purpose of this research the researcher has decided to compare three different commercially available tools that share similarities with the proposed system in this research. The three tools and systems that have been taken into consideration are RapidMiner, Lexalytics and drive research. The researcher has also peer reviewed other online journals that have proposed similar system and solutions for sentiment analysis.

RapidMiner is a cloud-based data mining tool which uses machine learning for sentiment analysis and an array of data science related operations. The platform can be used by data scientist, marketers and researcher to analyze consumer data and generate insights with relative

ease. The tool has been a popular choice for business at different scale because it provides an easy-to-use automated machine learning and other industry best practice to build predictive models easily to analyze all kinds of textual data to generate insights. (Anon, n.d. e) The tool also provides several other advanced features for ML experts and data scientist to carry out more advanced research using the platform. The tool is popular among business to understand consumer behavior through mining data from social medias, review website, blogs and other online sources.

The next tool that has been considered as a similar system to the one proposed by the researcher is Lexalytics. It is a cloud-based business intelligence solution which is used for text analytics through analyzing different kinds of texts. Lexalytics works with social comments, surveys, reviews and any other type of text document (Barysevich, 2020). Apart from sentiment analysis the tool has other features such as theme extraction and intention detection that makes it easier for users to see expanded context and achieve a greater understanding of consumer sentiment. Lexalytics is similarly also popular amongst large corporations and brands looking to understand their customers better.

The third commercially available solution that the researcher has considered is drive research. Drive research is a market research company that actively uses sentiment analysis to understand the sentiment behind open-ended online survey responses. In an online article by Gell (2019) he describes the process of how sentiment analysis is used in open-ended text response to interpret whether it is positive, neutral or negative in nature. The tool provides a proprietary online survey tool that can be used to ask almost any type of question where the answer boxes contain a sentiment analysis feature that predicts the underlying sentiment behind the response which brands can use to better understand sentiment of their consumers. The alternative approach the solution takes is to run a sentiment analysis on the combined response gather to develop an analytics report. The survey tool leverages the Valance Aware Dictionary and sEntiment Reasoner (VADER) Sentiment Analysis Package, which is part of the Natural Language Toolkit (NLTK) (Gell, 2019). Text is fed through a dictionary that has predetermined sentiment values for words, acronyms, slang terms, and even emoticons. Everything has a value from negative 4 to positive 4, with 0 being neutral. The values from every piece of the response are then added together, and the score is normalized to fall between -1 and +1. If a respondent’s answer is multiple sentences, the sentences' scores are averaged to determine the overall sentiment score (Gell, 2019).

Lastly, the final system the researcher consider is a proposed system by Ali et al. (2017) which proposes to provide sentiment analysis as a service. The research proposes the idea of providing sentiment analysis as a service (SAaaS) framework to abstract sentiment from social information services to analyse and transform the information into insights using machine learning techniques. SAaaS uses this classification to dynamically compose services for noise removal, geo-tagging (e.g., location extraction) and sentiment extraction. Finally, the results are presented in various formats, i.e., maps, charts using social media health surveillance as a motivational scenario. SAaaS uses a generic information composition approach to compose the social sensors' data as a service from multiple sources for sentiment analysis. Traditional approaches do not consider different types and characteristics of social information services for sentiment analysis. On the contrary, SAaaS considers the different properties such as data size, type, etc., and dynamically composes appropriate services for sentiment analysis specifically focusing on disease surveillance (Ali et al., 2017).

The table below provides a brief comparison and analysis of similar system that has been taken into consideration by the researcher.

Attributes	**RapidMiner**	**Lexalytics**	**Drive Research**	**Focusen**
Core Features	Automatic ML models, Data mining	Business intelligence (Text analytics, theme extraction, intention detection.)	Sentiment analysis from surveys.	Data pipeline and sentiment analysis for open-ended text
Target User	Market Researcher & Non-technical researcher.	Business Analyst & Marketing teams	Marketing teams.	Market Research companies and Individual Researchers.
Cost	Paid/Subscription based.	Paid/Subscription based.	Paid	Open-Source
Mode	SaaS	SaaS	Service	SaaS

2.4 Summary

In the literature review component of this report the researcher provided a thorough review of the domain space of the project by looking into the importance of market research, importance of understanding consumer behaviour, the usage of machine learning and sentiment analysis techniques to understand consumer behaviour and deploying solutions in the cloud. The review highlighted the importance of market research for business to achieve profitability and consumer satisfaction which has been backed by several research findings. Similarly, the review has also investigated the importance of consumer behaviour and why understanding consumer behaviour is at the heart for marketing to satisfy existing customers and gain new one as highlighted by several researches. Next, the review discusses the use of machine learning and sentiment analysis to understand consumer behaviour and discuss why machine learning has been a popular candidate for several nature language processing and sentiment analysis operations. The section highlights the different tools, techniques and classifiers available to tackle the different levels of text analysis and sentiment analysis problem but also discuss on some of the challenges and drawback faced by several researchers in the domain space.

3.0 Technical Research

This section provides the details of the technical research that have been conducted to develop the proposed system in this research. This will include the programming languages analysed and chosen, the integrated development environment (IDEs) to be used, libraries and tools chosen, databases, operating system and browser required to successful develop the proposed system.

3.1 Programming Language

The table below shows a brief analysis and comparison of the programming languages that has been taken into consideration for the proposed project.

Attributes	Python 3	R	MATLAB
Main Purpose	General Programming	Statistical Computing	Commercial Numerical Computing
License	Open Source	Open Source	Proprietary
Cost	Free	Free	Paid (free version available for academic use)
Ease of Use	Easiest	Medium	Difficult
Commercial Acceptance	Highly in demand	Popular for data science, statistics & data mining	Most Academic usage

According to research conducted by Ozgur & University (2017) investigated most suitable programming language options available for students to learn operational research, statistics and data science in an academic setting. The researcher chooses the top three candidates to be Python, R programming language and MATLAB and looked into the overall pros and cons of each language and its advantage, disadvantages and features in depth for a mostly academic setting and concluded that Python was the best choice because of the attributes highlighted in the table above. The research also concluded that R also a great option for students since it is

widely accepted in the industry for data mining operations and having the knowledge of R will provide a significant competitive advantage for students within the field of data science.

Based on the research conducted above by Ozgur & University (2017), personal research conducted by the researcher and taking into consideration several other factors the researcher has decided to use python as the primary programming language for this project. The decision was backed by the fact that there are several libraries, tools and frameworks available in python for data science and machine learning operations which the researcher can use in this project. Also, since the primary focus of this project would be sentiment analysis and machine learning, the researcher has found that python has extensive number of resources available for this use-case compared to R and MATLAB and is relatively easier to use. Python offer several machine learning frameworks and libraries, web-frameworks and other data science tools and libraries for free that the researcher can leverage on for this project making Python the most suitable choice.

The developer will also use SQL (Structured Query Language) which a programming language used to create, modify and extract data from relational databases and is the standard language for relational database management systems. Since the project will implement the use of a RDMS solution for it database storage (further discussed in the database section), SQL will be used to interact with the database and run quires in the application. The project will also use HTML, CSS and a bit of JavaScript web programming languages to develop the frontend dashboard for the application in conjunction with Dash (also discussed in detailed in the following sections).

3.2 Integrated Development Environment (IDE)

An integrated development Environment (IDE) is a software application that provides comprehensive facilities to computer programmer to develop software. An IDE usually consists of at least a source code editor, building automation tools and a debugger (Anon, 2020). For this project the developer will primarily use two different IDEs, the first one being Jupyter Notebooks, which will be used for data science and machine learning components of the project and Visual Studio Code (VS code) which will be used for the web development portion of the project. The researcher has decided to use to different IDEs because each specializes in different components of the project. Jupyter Notebooks, being a leading open-

source IDE solution for data scientists and VS code being a free open source for software development. As a secondary IDEs for the data science component of the project, the researcher will also use Google Colaboratory (Google Colab).

3.2.1 Jupyter Notebook

Jupyter Notebooks is a web-based IDE that supports multiple languages and is used to create research documents with code, equations and narrative text (Project Jupyter, 2020). Jupyter provides excellent support for Python which the primary programming language chosen for this project and provides an interactive and user-friendly interface making It suitable for data scientists and machine learning practitioners. Jupyter notebook will primary be used in this project for designing and developing the data pipelines, exploratory data analysis, pre-processing data, experimenting and fine-tuning sentiment analysis based machine learning models and testing the outputs from the model using interactive charts and visualizations. Jupyter notebooks are the primary choice for most data science professional because it provides and interactive layout where the user can type in code into the code block and simultaneously check the output from each code block in the attached output window. Users can also add text blocks to write down notes and documentation related to the code making it and excellent choice for coding and documenting progress for data science and analytics projects. Also, the IDE is free and open source which also contributes to its high adaption amongst the industry, academia and individual researchers for research projects.

3.2.2 Visual Studio Code

The next IDE that the researcher will use for the web development component is visual studio code (VS code). This IDE will be used to develop the front-end web application and interactive dashboard for the project. Visual Studio Code is a lightweight but powerful source code editor which runs on your desktop and is available for Windows, macOS and Linux. It comes with built-in support for JavaScript, TypeScript and Node.js and has a rich ecosystem of extensions for other languages (such as C++, C#, Java, Python, PHP, Go) and runtimes (such as .NET and Unity) (Visual Studio Code, 2020). The IDE has been developed and maintained by Microsoft. Even though Jupyter notebook provides an excellent platform for the data science and machine learning component of this project, it does not necessarily have the functionalities and flexibility of a software development IDE such as VS code. Since the developer plans to use a JavaScript framework called Vue.js for the front-end dashboard of the application, VS code is

the idle choice as it provides extensions for different libraries and frameworks with auto completion, linting and debug functionalities which help massively assist the developer during the project.

3.2.3 Google Colaboratory

The secondary IDE the researcher will use for the data science and machine learning component of this project is Google Colab. Colab provides similar functionalities as Jupyter notebook but is hosted by Google and provides cloud based computational resources for free which can be a massive advantage for researcher who do not have a powerful machine. Colab also provides some useful feature such as graphical processing unit (GPU) and tensor processing unit (TPU) support which can come in handy while training large scale machine learning models.

3.3 Libraries and Tools

Python is an excellent programming language which includes thousands of great third-party packages and libraries. To complete this project successfully, the researcher will use several third-party libraries along with the standard libraries provided with python for data pre-processing and machine learning model development. For the dashboard component of this project, the developer will also use popular python front-end and backend libraries for constructing the data pipeline, visualizations and deployment of the application.

3.3.1 SciPy

The first library package that will be used for this project is SciPy. It is not a specific library but a group of python packages including Pandas and NumPy and Matplotlib that are used for scientific computing (SciPy.org, 2020). Libraries included in this package such NumPy, Pandas and Matplotlib will be used for data analysis, manipulation, visualization and to perform other scientific operations to get a better understanding of the dataset that will be used in this project and to pre-process the data before it is feed into the ML (machine learning) model.

3.3.2 Natural Language Toolkit (NLTK)

NLTK is a platform for developing python programs with natural language processing (NLP). NLTK provides an easy-to-use interface and several corpora and lexical resources such as WordNet along with several text processing libraries for classification, tokenization, stemming, tagging, parsing and semantic reasoning which are integral parts of this project and will be used for text analysis and pre-processing.

3.3.3 TextBlob

The next crucial library that will be used for the development of this project is TextBlob. TextBlob is a Python (2 and 3) library for processing textual data. The library provides a simple API for common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more (TextBlob, 2020), which are the core requirements for this project. The application will use the pre-trained sentiment analysis model provide by TextBlob for sentiment classification of real-time tweets. Under the hood the sentiment analysis modules use a Naïve Bayes Classifier to calculate average polarity and subjectivity scores of each tweets through keywords identification which will be ideal of the real-time processing requirements of this project.

3.3.4 Tweepy

Another library that is mission critical for this project is Tweepy, which is an easy-to-use python library for accessing the Twitter API (Tweepy, 2020). The project will use the Tweepy library to process the real-time data collected using the Twitter API before storing it in a relational database for further processing. The libarary makes it easy to use twitter streaming as it handle authentication, connection, sessions and also read incoming tweets in real-time.

3.3.5 Plotly

The next python library, that will be used for this project is Plotly, which is an open-source graphing library for python (Plotly, 2020). The graphing library in Plotly allows users to develop interactive graphs, plots and other visualization charts which will be used in the front-end dashboard of the Focusen application.

3.3.6 Python Flask

Flask is a python-based server-side web application framework. Flask will be used to develop the backend of the Focusen sentiment analysis tool. Once the model has been trained and evaluated Flask will be used to deploy and serve to model to the frontend dashboard in conjunction with Dash for the application. Flask is chosen because its simplicity, ease of use and since it is developed in python, it can easily be integrated with the rest of the project (Pallets Projects, 2020).

3.3.7 Dash

Dash is a python framework for developing frontend web-applications. The framework is built on top of Flask, Plotly.JS and React.JS and allows developers to build front-end dashboards using pure python (Plotly, 2020). The flexible and core structure of Plotly makes it ideal for this project because the Focusen application relies on Plotly for visualizations and the front-end dashboard can be easily developed utilizing the Dash library.

3.3.8 Psycopg2

As the developer has opted to use PostgreSQL database as the relational database of choice for this project due to its flexibility and robustness in deployment scenarios, the project will use Psycop2 as the connector of choice to connect the database to the python application. Psycopg2 is the most popular PostgreSQL database adapter for python programming language due to its complete implementation of the Python DB API specifications and safety standards (Gregorio, n.d.).

3.4 Database Management System (DBMS)

As for the demonstration of the proposed system, the developer has decided to use real-time Twitter data using the Twitter API, the data that is being fetched using the API needs to be stored in a database for further processing and recall for the front-end web application. Since the data is highly structured and can be stored in clear rows and columns, a Relation Database Management System (RDMS) will be used to fulfil the data storage requirements of the application. For the purpose of this project the developer has decided to use the PostgreSQL relational database for deployment and MySQL for running local testing of the application.

3.4.1 PostgreSQL

PostgreSQL is an open-source object-relational database system. The database is a popular choice for developers because of its reliability, feature robustness and performance (PostgreSQL Global Development, 2020) . The database will be used to store the data extracted using the Twitter API for the project in real-time before it is pre-processed and analysed by the model to be displayed on the frontend dashboard.

3.4.2 MySQL

For local testing of the application, the developer has opted to use MySQL as the database was readily available in the development environment used for this project. MySQL is another popular open-source Relational Database Management System (RDMS) and is fairly similar to PostgreSQL which has been used for the deployment of the proposed system.

PostgreSQL has been chosen over MySQL for deployment purpose because of its wider compatibility, robustness, scalability and better performance compared to MySQL in deployment scenarios. It is also an object-relational database while MySQL is a purely relational database and provide several other additional features such as full ACID (Atomicity, Consistency, Isolation, Durability) and SQL compliance (2ndQuadrant, n.d.). Also, an RDMS solution has chosen for this project instead of a No-SQL database such as MongoDB because the data is highly structured can be easily stored in rows and columns making relational database an ideal choice to fulfil the requirements of this application.

3.4.3 Beekeeper Studio – GUI SQL Editor

The developer will also use an open-source GUI based SQL editor named Beekeeper Studio to easily interact with the RDMS being used for this project. The Editor provides user a simple and user-friendly way to manage, query and interact with the relational database of choice. The editor fully supports PostgreSQL and MySQL which are the database being used to develop the proposed system making it an ideal choice.

3.5 Operating System

The researcher will be using Apple MacOS Catalina (Version 10.15.5) or newer for the local development and testing of the application. The application is planned to be deployed on

localhost (MacOS) for demonstration purposes but has the full setup the be deployed on a server or public cloud service such as Heroku or AWS in which case the machine will be configured to run a server-side distribution of Ubuntu as it is a stable and robust OS and also a popular choice for servers.

3.5.1 Desktop Usage for Development

Since the researcher uses an Apple Macbook Pro as the primary computer for development. MacOS Catalina (Version 10.15.5) or newer will used for the development and testing of the application. MacOS provides built in support for python programming language and several other UNIX features making it a suitable choice for developers and data scientists.

3.6 Browser

The primary browser that will be used for this project is Google Chrome (Version 83.0.4103) or newer. The researcher will use this to run locally hosted Jupyter instances and to run Google Colab during the data science and machine learning phase of this project. Chrome will also be used for research and other web-based activity. The browser will also be used for development and testing of the web application for this project and also be used to demonstrate the application after it has been deployed locally. As a secondary browser, the researcher will use Apple Safari Web browser which is also a popular choice for MacOS users.

3.7 Summary

Different tools and frameworks have been discussed in the above section which will be used to successfully complete this project. However, several other tools may be added in the latter half of the project to accommodate changes in requirement. The tools and frameworks chosen will provide excellent and stable environment for the research environment throughout this project. The table below provides a summary of all the technical requirements for this project according to the research conducted above.

Category	Choice	Remarks
Programming Language	Python 3.8	3.8 is the current stable release of python
IDE	1. Jupyter Notebook 2. Visual Studio Code 3. Google Colab 4. Beekeeper Studio	Different IDEs will be used in different components of the project.
Tools/Libraries	1. SciPY 2. NLTK 3. Python Flask 4. Dash 5. TextBlob 6. Plotly 7. Tweepy 8. Psycopg2	The latest stable version able during the development phase has been used.
Storage	1. PostgreSQL 2. MySQL	PostgreSQL has been chosen as the RDMS of choice for deployment while MySQL has been used during the test phase of the project.
Operating System	1. MacOS X Catalina (Version 10.15.5) 2. Linux Ubuntu Distro (Server)	MacOS latest version for local development, testing and deployment. Latest stable Ubuntu release for server/cloud deployments.
Hardware	1. Apple Macbook Pro 2019 (i9, 16GB, 512GB SSD) for development and local deployment.	Desktop for local development, testing and deployment demonstration.

	2. Cloud instance for public deployment (TBC)	

4.0 Methodology

This section discusses the methodology that has been chosen for the project. The methodology used for the project plays a vital role in deciding whether the project will be a success or not, so it very important to analyze the different options available and choose a methodology that will properly suite the project and can be properly executed by the researcher. Since the project has two main components, the first being the data science and analytics component and the second being the web development component it is hard to choose a single methodology that will accommodate the project entirely. Therefore, the researcher has decided to use a mixture of a software development methodology and data science methodology. The software development methodology will focus on the software development life cycle (SDLC) for the web-application of the project while the data science methodology will focus on the analytics and machine learning model development component of the project. For the primary component of the project which is data analytics and model development component the researcher has chosen CRISP-DM (cross-industry standard process for data mining) and SCRUM to be the leading candidates as the primary methodology of this project. The researcher as also looked in SEMMA as a close runner up candidate for the project. In the section below the research will analyze and compare the two different methodologies shortlisted for this project and chosen a suitable candidate to move forward with for this project. For the second component which is the web development and deployment component of the project, the researcher has chosen RAD (Rapid Application Development) software development methodology as the most suitable option for this project.

4.1 Comparison of Methodologies: CRISP-DM vs SCRUM vs SEMMA

4.1.1 CRISP-DM

The CRISP-DM (Cross Industry Standard Process for Data Mining) was proposed as a comprehensive process model for carrying out data mining projects. The process model is independent of both the industry sector and the technology used (Wirth & Hipp, 2000). The methodology is described as a hierarchical process model that consist a set of tasks described at four level of abstraction (from general to specific): phase, generic task, specialized task and process instance (Chapman et al., 2000). The methodology contains six different iterative phases which each contain their own defined tasks and deliverables such as documentation and reports. The six-phase included in CRISP-Dm are: 1. Business Understanding 2. Data

Understanding 3. Data Preparation 4. Modelling 5. Evaluation and 6. Deployment. Even though CRISP-DM was initial proposed as a methodology for data mining operations, the methodology has gained popularity for all kinds of data science and analytics projects. A poll conducted by KDnuggets asking "What main methodology are you using for analytics, data mining, or data science projects" from 2007 till 2014 showed the most popular methodology amongst respondents was CRISP-DM with 43% of the total votes (KDnuggets, 2014). As Saltz (2015) has mentioned in this research, CRISP-DM is possibly the first step towards defining a data science methodology. Even though the methodology has made a significant impact with the data science community it is still far from being perfect. Even though it has several benefits such as flexibility, adoptability, generalization and flexibility but since it was first proposed as primarily a methodology for data mining operations, it is heavily task-focused and fails to address team and communication issues making not ideal for large scale project with big teams. It is also heavy on documentation, old-school and does not take a project management approach. But the methodology is popular amongst individual researcher and small teams making it a suitable candidate for this project.

4.1.2 SCRUM

SCRUM is an agile software development methodology co-founded by Jeff Sutherland and Ken Schwaber in the 1990s. It was proposed as a flexible software development methodology which assumed that the processes in software development are unpredictable, complicated and can only be roughly described as an overall progression. SCRUM defines the software development process as a loose set of activities that combines known, workable tools and techniques with the best that a development team can devise to build a system (Schwaber, 1997). Even though the methodology was mainly proposed as a software development methodology, it is popular across main different industries and a wide variety of companies. The SCRUM guide is a definitive guide to Scrum which recommends a team consisting of 3 to 9 development members and defines the roles as product owner, Scrum Master and Development team (The Scrum Guide, 2017). Even though it is mostly popular for software development projects scrum has also gained popular amongst data science and analytics projects. Some benefits of the scrum are its focus towards customers, autonomy, improvement through inspection, accountability and the sense of urgency. Even though scrum is a very popular methodology even for data analytics projects it is mostly suitable for teams and not individual researchers and developers.

4.1.3 SEMMA

A few years before CRISP-DM was proposed to the world, SEMMA (Sample, Explore, Modify, Model, Assess) was independently developed by the SAS institute as a data mining and analytical operations methodology. The methodology has 5 stages in its process: 1. Sample – sampling the data by extracting it from a large dataset. 2. Explore – exploration of the data to search for trends and anomalies 3. Modify – Modifying the data by creating, selecting and transforming the variable for model selection. 4. Model – Modelling the data to allow the software to search for patterns. and 5. Assess – assessing and evaluating the results. Although SEMMA process is independent of the tool chosen it is heavily guiding users towards SAS Enterprise Miner software due to its nature of instruction and heavy linkage with the DM tool (Azevedo & Santos, 2008). This is somewhat considered a limiting factor for the methodology and might limited its flexibility and usage for this project.

4.1.4 Summary of Comparison

In the above section the researcher has analysed and compared the most suitable methodologies that can be adopted for this project. After a detailed analysis and research taken into consideration several different factors and aspects of this project, the researcher has decided to chosen CRISP-DM as the primary methodology for the data science component of this project. In the next section the researcher will provide a detailed justification of the selection and a detailed analysis of the chosen methodology

4.2 Justification of Chosen Methodologies

As mentioned above the researcher has decided to select CRISP-DM (Cross-Industry standard process for data mining) as the primary methodology for the data analytics component of this project and decided to choose RAD (Rapid Application Development) for the web development component of this project.

The reason the researcher has decided to choose CRISP-DM is because the methodology has the best fit per the requirement of this project. The methodology has several advantages such as it a very general methodology suitable for all kinds of data science and analytics project even though it was first conceived as a methodology for mainly data mining. Next, the process is highly adaptable and flexible making is ideal choice for student researcher and student project such as this one. The method also provides a right start to the project through its

requirement of focus on business understanding which is crucial component for this project. The method also has a right ending with its deployment requirements which also perfectly aligns with the requirements of this project. Even though the methodology provides suggest emphasis on documentation, it would necessarily not be a problem since the project is a research-based project and need certain amount of documentation anyway. Also, the project is being carried out by a single researcher which also aligns with the communication and teamwork limitations of CRISP-DM. The different steps included in CRISP-DM which are business understanding, data understanding, data preparation, modelling, evaluation and deployment perfectly align with researcher expectations and timeline making it the ideal choice for this project considering all the factors and variables.

For the second part of the project which is the web application development component of the project, the researcher has decided to go with RAD (Rapid Application Development) methodology because firstly, the project need a suitable software development methodology because CRISP-DM even though it is the perfect fit for the data analytics component of the project does not necessarily provide the method for the successful development of a web application. Since the project is not a full-fledged software development project and only consist a small part of developing an interactive dashboard and a real-time data-pipeline using the Twitter API, RAD provided the right methods according to the fairly minimal requirements through its steps of requirement gathering, user design, construction and cutover. It allows developer to break the project in smaller components and manageable task and focuses on efficiency making it suitable for this project.

4.3 Details of chosen Methodologies

4.3.1 Cross Industry Standard Process for Data Mining (CRISP-DM)

CRISP-DM is a methodology suitable for data mining, data science and analytics projects. The methodology is broken down six major iterative phases with each containing it its defined tasks, set of deliverables such as reports and documentation. The six phases of CRISP-DM are business understanding, data understanding, data preparations, modelling, evaluation and deployment. The diagram below shows the cycle of a CRISP-DM project.

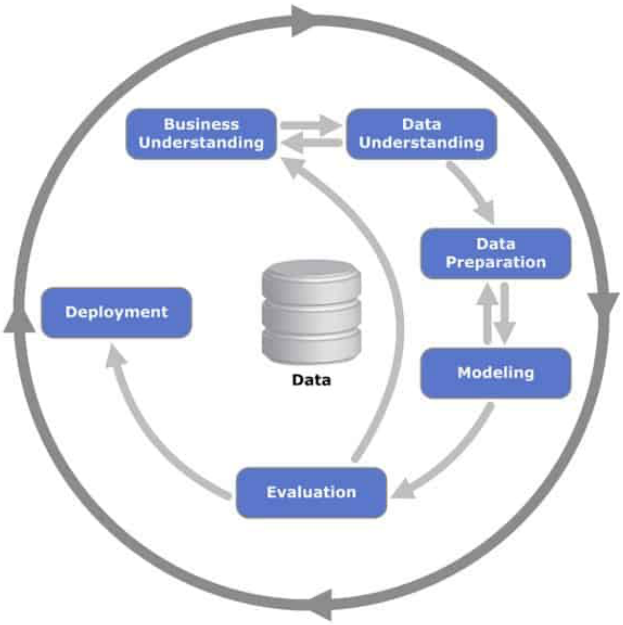


Figure 1 Phase of a CRISP-DM project (Wirth & Hipp, 2000).

1. Business Understanding

The first phase of the process focuses on understanding the project objectives and requirement of the project for a business perspective (Wirth & Hipp, 2000). Once the requirement and objectives has been understood thoroughly, the knowledge gained should be noted as a problem definition and objectives and a preliminary project plan has to be designed to meet the objectives.

2. Data Understanding

The next phase of the project begins with the collection of a suitable dataset that will be used for the project. Once a suitable dataset has been chosen, the researcher needs to analyse the data to assess the quality of the data available, identify irregularities and trends and conduct some exploratory data analysis to uncover some initial insights from the data and form hypothesis for further investigation.

3. Data Preparation

This phase consists of all the activities needed to prepare the final dataset such as cleaning the data, removing missing values, construction of new attributes and identify feature sets and

variables before the data is feed into the model for prediction. These tasks might be performed several times to achieve the best results from the model.

4. Modelling

In this phase, various machine learning and statistical models are chosen and applied to the dataset and the parameters are constantly tuned to achieve higher accuracy to get the optimal results from the model. This process can also be carried out multiple time and has a strong correlation with the data preparation phase as both strongly rely on each other.

5. Evaluation

In this stage of the process, the researcher can develop one or more models to compare the results in order to select the most optimal output for the analytics purpose defined. Before continuing to deployment, it is critical that the model is properly evaluated taking into considerations are the factors and limitations. It is also important to review and record the steps that has been taken the develop the model. It is crucial that the model output align properly with the business objectives and meet the requirements of the project. Once the model is thoroughly evaluated, the decision needs to be taken on whether the model is ready for deployment.

6. Deployment

Development of the model is generally not the completion of the project. The model needs to be deployed so the target users can use the model to fulfil the business requirements. The knowledge gained for the project and results need to be properly documented to presented to the key stakeholders. Depending on the requirements of the project the deployment phase can either be as simple as compiling and generating a report or deploying a full-fledged working system for use. In most scenarios the deployment is carried out by developers with domain knowledge in the specified filed rather than the data analyst to ensure a smooth deployment.

The diagram below shows the phases of a CRISP-DM project and also a detailed description of all the activities included in each phase.

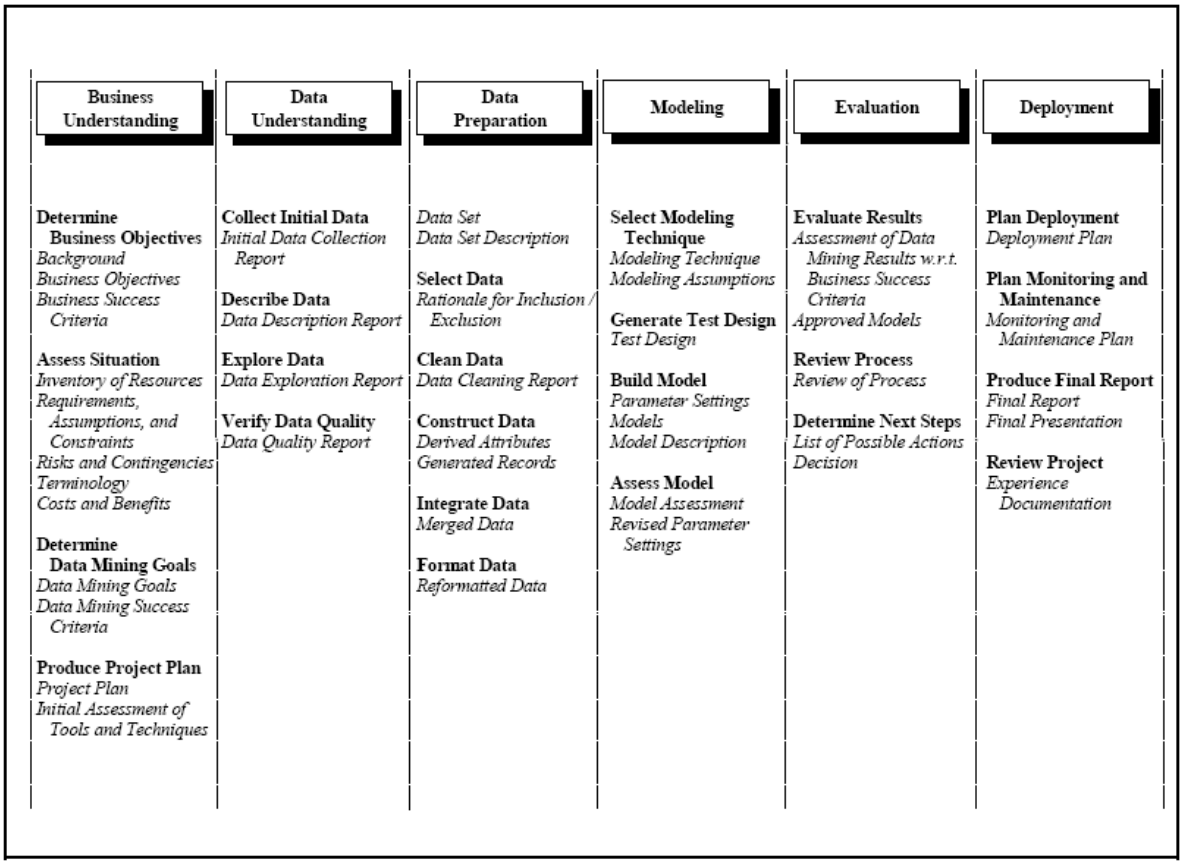


Figure 2 Phase and activities in a CRISP-DM project (Wirth & Hipp, 2000)

4.3.2 Rapid Application Development (RAD)

For the web development component of the project the developer has decided to choose RAD as the primary methodology. RAD is an agile software development methodology invented by James Martin in 1991 (Singh, 2019). It prioritizes rapid prototype release and quick iterations and heavily relies on user involvement in the process. It provides enhanced flexibility and adaptability for developers to make quick advancements during the development process (Singh, 2019). Because of its rapid nature it reduces development time and speeds up delivery. RAD also encourages the re-usage of code which means less manual coding. Fewer error and comparatively lower testing times. It also provides better risk management for stakeholders and far less surprises compared to other traditional methodologies. There are four phases in RAD which are: requirement planning, user design, construction and cutover.

1. Requirement Gathering

This is the first phase of the project and is the equivalent of the project scope meeting. Although the planning phase is vastly condensed compared to some other methodologies it is an important phase of the project where the problem is researched, the requirements of the project are defined and are finalized with consent from the project stakeholders.

2. User Design

Once the project has been scoped, in this phase the requirements are finalized, and the team immediately jumps in development to build out the user design through prototyping multiple iterations. This is the main phase of RAD as the process keep repeating until the requirements are meet. This method gives developers the opportunity to iterate the model as they go until they reach a satisfactory design.

3. Rapid Construction

In this phase the prototypes are taken from the design phase and are converted to fully working models. This phase usually includes preparation for rapid construction, program and application development, coding and unit, integration and system testing to ensure the system is functioning corrects and meets all the requirements.

4. Cutover

This is the implementation phase of the project where the finished product is launched and deployed to end users. This phase also includes data conversion, testing and changeover to new system if an older system exists as well as user training.

The diagram below shows the phases of a RAD Project.

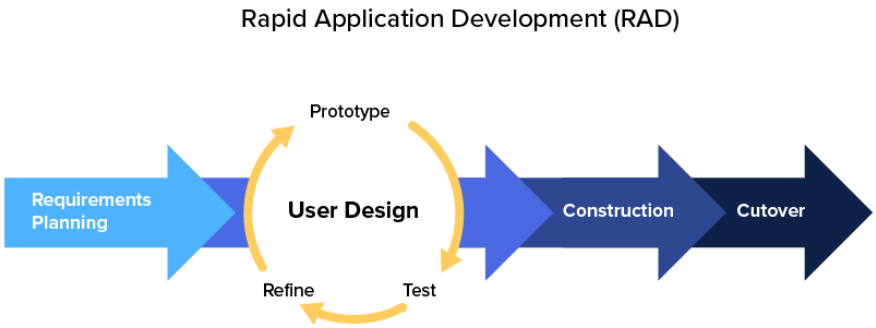


Figure 3 Phase of an RAD project.

4.4 Overview of Project Progression

This project is divided into two distinct components. The first part containing the data analytics component using the CRISP-DM methodology for data sconce and analytics and the second part containing the web development component using RAD software development methodology. This section the research will provide a detailed progression plan for the project regarding the methodologies chosen.

4.4.1 Progression of Data Analytics Component

1. Business Understanding

Tasks	Understanding business problems and defining main objectives for the project
Output	Project Objectives, Success criteria for project.

The first phase of the project will start with understanding the business requirements and setting up the main objectivities for the project. Since the project is mainly focused on developing a sentiment analysis tool for online focus group conversations and open-texted text responses from surveys to understand consumer behaviour for mainly market research companies and teams, the researcher will start the process by gathering requirements from subject matter experts working in the market research industry through interviews and questionnaires to discuss the problem and validate the requirements. For this project the researcher will collaborate with one of the leading marketing research companies based in Kuala Lumpur,

Malaysia. The data gathering techniques used and the data gathered will be discussed in the research and analysis section of this report.

Once the developer gather the data from the target users of the system and has defined the functional, non-functional requirements and objectives of the proposed system. The developer will move to the next part of the project which is data understanding, data pipeline development and exploration.

2. Data Understanding

Tasks	Exploratory analysis of dataset, data-pipeline development
Output	Data exploration report. Data ingestion pipeline.

Once the objectives have been set and the requirements have been validated the researcher will investigate different datasets that are available to find a dataset that fits the purpose of the project. As the tool is mainly target towards consumers, it would be best to choose a sentiment analysis-based dataset for the most optimal outcome.

Since the goal of the project is to develop a real-time sentiment analysis system, the developer has chosen to use a live data stream from social media using the Twitter API to stream tweets in real-time for analysis and sentiment classification. The developer has developed a real-time data ingestion pipeline using the API and Tweepy python library.

3. Data Preparation

Tasks	Clean and prepare chosen dataset for model. Select variable and features according to requirements.
Output	Data analysis report, Data pre-processing pipeline.

Once the data understanding phase has been completed the researcher will move into preparing the data for the machine learning model. The process will include several text analysis python-based tools and techniques to clean the data and prepare the raw dataset for input into the model. Not all variables present in the dataset can be used for model training, so the researcher clearly need to select the relevant variable to achieve the most optimal results. The data also needs to be normalized and cleaned to follow best practises and achieve the best results.

As the developer has decided to use a real-time data stream using the Twitter API, the pre-processing of the data needs to be simple, optimized and fast to achieve real-time performance and avoid delay and lags. For this the developer has chosen to use the python Regex library (RE) to clean and normalize the data and remove any unwanted characters from the tweets. Python NLTK will be used to for text tokenization and TextBlob will be used for further processing of the data before it is feed into the ML (machine learning) model.

4. Modelling

Tasks	Select model, build model, tune parameters to achieve optimal results.
Output	Models and techniques used.

The next phase of the project is the most critical component which is the Modelling phase. In this phase the training dataset would be used to train the model and check its accuracy. The optimal accuracy of the model will be determined by the selection evaluation matrix. If the optimal outcome is not meet the research may repeat the previous steps to fine tune the dataset to achieve better result. The entire process is based on trial and error and a lot of fine tuning and feature engineering required to achieve the best results. Once the researcher is satisfied with the results, he will proceed to the next stage which is further evaluation of the model. Further details of which models, frameworks and techniques and tools to be used will be discussed in the next section of the project. This phase will also be carried out using python programming language using a jupyter notebook.

Since the proposed system will need to process Tweets from the Twitter API and output sentiment classifications in real-time, the developer has decided to use a pre-trained ML model for this project. The project will use the popular python Textblob library and its simple sentiment analysis model to classify the tweets in real-time. The Textblob sentiment analysis model uses a pre-trained Naïve Bayes classifier to classify the tweets based on polarity and subjective of each word in the sentence as provided a high accuracy output in near real-time which is perfect for the requirements of this project. Once the data is ingested, pre-processed and classified using the ML (machine learning) model, the data and results will be stored in a SQL database which will be used to then output the results in the web-based dashboard.

5. Evaluation

Tasks	Evaluate model performance against business objectives. Determine next steps.
Output	Evaluation assessment, list of possible actions.

After the researcher is satisfied with the performance of the model, he will move into further evaluation to ensure of the business requirements and project objectives are being meet by the model. The researcher will also be looking into other model performance criteria such accuracy, root mean square error (RMSE) and check the result to validate if the model is being over fitted. Further on model evaluation will be discussed on the next part of this project.

6. Deployment

Tasks	Deploying planning.
Output	Deployment strategy.

Deploying a model is the phase where the model is made available to end users for use. In context to this project, the model will be deployed as part of a web application which will take input data from users, analyse the data and output a predictive analysis report based on the data provided in a web dashboard. However, in CRISP-DM methodology the deployment is not the end of the project, the model needs to be closely monitored and evaluated to ensure that it is fully functional and meeting the business requirements and objectives. A deployment plan will also be developed to ensure the process is carried out successfully. Further details about how the deployment will be carried out will be provided in the next section of the report under the progression of web development component of this project.

4.4.2 Progression of web development component

The second component of the project will be the development of the web application for the sentiment analysis model deployment and interactive dashboard. For this part of the project, the researcher will use adopt the RAD (Rapid Application Development) model.

1. The first phase of the project is requirement gathering. The requirements for this project are fairly simple and minimal as the researcher only want to develop an interactive web

dashboard with visualizations to display the predictive analysis report for the model based on the input data in real-time.

2. The next step is user design, this is the step where the researchers will keep iterating prototype designs until all the requirements for the project are meet. This is a cyclic process rather than a single step has the process keeps repeating until the requirements are meet. Once the design is confirmed the project will progress to the next step. The developer has chosen to add a real-time scatter plot which will display the sentiment classification on a time-series graph, a top keyword histogram which will display the top keywords from the tweets in a graph, a geographic distribution map which will show the location of the tweet users and all an overall summary of the streaming platform (Twitter) and a sentiment analysis percentage chart to summarize the classifications in real-time.

3. Once the user design phase has been completed, the next phase of the project is construction of the system according to the approved design. In this phase the developer will rapidly start the development of the web application. An IDE and web browser will be used to build and test the application according to the choice mentioned chapter 3. The model will also be connected to the web application using the backend framework mentioned.

4. The last phase of the project is cutover. In this phase the dashboard will be completed and deployed on localhost for demonstration and will be available for testing for the end users. The developer wanted to deploy the system on an Amazon EC2 instance to make it publicly available for the end user but during the time of development of this project, AWS does not allow data scrapping operations to run on EC2 instances. As this project uses the Twitter API to scrap data from Twitter the project can be deployed on AWS as therefor will be demonstrated and deployed on localhost.

5.0 Research Methods

5.1 Introduction

It is important for any researcher to validate the requirements of their project to know if the solution they propose will solve the business problem they are looking to solve. As mentioned in a study by Pandey et al. (2012) the development and gathering of good quality requirements is one of the most important activities for any developer to develop good quality software. Since the target users for the proposed system are market research companies and teams, the researcher approached experts working within the market research industry to validate the requirements of this project. For the purpose of this research, the researcher contacted one of the leading market research companies in Malaysia to collect data and analyse it to validate the functional requirements of the proposed system.

5.1.2 Importance of Data Gathering and Analysis

As mentioned above, validating the functional requirements for any software project with stakeholders is a crucial step to develop good software. The data gathered from subject matter experts will immensely help the researcher to identify the key objective and functional requirements of the proposed system. Since the purpose of the project is to develop a sentiment analysis model to understand consumer behaviour, it was important for the researcher to gather requirements from experts working in the market research industry to validate the project. The research reached out to one of the leading market research companies to discuss the details of this project with the Chief Technology Officer (CTO) of the company and his team which consists of data analysts, researchers and engineers whose daily tasks involve running surveys and focus groups for client from various different industries ranging from food and beverage to telecommunication and etc. The tasks of the team also include processing and analysing the huge amount of open-ended text data collected from survey response and focus group conversations to generate reports to identify trends and provide insights for better business decisions for their clients. This made them the perfect candid who has the knowledge and domain expertise to validate the feasibility of the objectives and the functional requirements set for this project. The data collected helped the researcher understand the business problem better and fine tune the objectives of this project to meet actual business requirements from the proposed target users of this system. In the following section the researcher will discuss the method chosen for data collection and justify the choices made.

5.1.3 Data Gathering methods chosen – Expert Interviews and Questionnaires

To validate the requirements of this project, the researcher decided to use two separate data gathering techniques which are, expert interview and questionnaires. The researcher decided to interview Mr. Asyrique Thevendran (CTO and technical co-founder of Vase Market Research Company) and use a questionnaire to collect data from his team as a form of written interview due to limitations of time and circumstances during the research period. Both techniques chosen falls under the paradigm of qualitative research since both the questionnaire and interview mostly contained open-ended questions.

5.1.4 Justification of Chosen Techniques

Interviews are considered as a humanistic method to collect the data based on the social reality, behaviour, perspective, beliefs, attitude as well as the experiences of an individual (Mohajan, 2018; Pathak, Jena & Kalra, 2013). While, there is no numerical data will be involved in the qualitative model, since it allows the participants to share their opinion as well as experiences during the research without any restriction (McLeod, 2019; Mohajan, 2018; Pathak, Jena & Kalra, 2013). The process of expert interview does not also require a large sample size, the number of interviews needed to be conducted can be determined based on the requirements of the researcher which in this case is very limited. Also, the researcher preferred to have a back-and-forth conversation with the expert to thoroughly discuss the concerned business problem and understand the challenges they are currently facing while analysing large sums of open-ended text data responses. Since the expert has considerable experience in founding and running a tech team in a market research company that strongly leverages on technology, the insights gather from this interview immensely helped the researcher to gain a better understanding of the business problems and the challenges that require solving and will be valuable in developing the proposed system.

The second method used for collecting data was a questionnaire, which was used as a means of a written interview to collect data from the team of data analyst, researchers, data engineers working at X Market Research Company. Since the team's primary task involves running online focus group and survey to collect data from their consumer panels and analyse the data to generate insight reports, they also domain knowledge and expertise to validate the requirements of a system proposing to use sentiment analysis for online focus group and

surveys to understand consumer behaviour. Questionnaires are a method of data collection which is completed by the respondent in written format. It is a popular research instrument to collect usable data in high volumes (Marshall, 2005). Previous research has also shown that it is better to first conduct a qualitative research in form of an interview or focus group to collect the initial data for the response and then move to a questionnaire to further validate the data collected from the quantitative methods. The researcher has also taken a similar approach for this project but with a significantly limited sample size. Also using a questionnaire was significantly less resources intensive and help save time for both the team at the company and the researcher as they could answer the questions at their own pace at any time the saw fit without any interference. Most of the questions included in the questionnaire were open-ended questions to let respondents fully express themselves through their responses so the researcher can better understand the challenges they are facing and set the research objectives and requirements accordingly.

5.2 Design

This section will discuss the design of the questionnaire and the questions designed for the interview and provide justification on why the specific questions were chosen and how they relate to the research topic.

5.2.1 Expert Interview

In this section the researcher will discuss, the questions that have been chosen for the expert interview with and provide justification for the choices made.

Question 1	As being a market research company, can you kindly reflect on the importance of market research for brands and organizations?

This question was chosen to get the expert opinion of the interviewee on why they think in general market research is important for brands and organizations. Being the first question of the interview, the researcher wanted to set a flow for the conversation as the research heavily focuses on the effort to understanding consumer behaviour using sentiment analysis.

Question 2	Can you kindly talk about how technology has changed the market research industry in the past 10 years?

This question was chosen by the researcher to understand the impact, usage and adaption of technology in the market research industry since the project propose the use of an artificial intelligence-based software system to analyse data.

Question 3	Can you briefly describe the workflow of your team from data collection to analysis while conducting online focus groups and surveys and the challenges faced?

This question was chosen to understand the overall workflow of the data analytics team and assess the different steps involved in the process of generating insights and building a report from raw data collected through focus group conversations and surveys. This helps the researcher understand where exactly the proposed solution will fit into the workflow of the target users.

Question 4	Can you kindly specify what is the average time frame for completion of an average analytics project from data collection to report generation?

This question was discussed to understand how much time is spent by the team on one analytics project from start to finish. This would give the researcher a rough idea who much time the target users are willing to spend using the proposed system to complete their tasks and how much time can be saved using the proposed system in comparison the current system in place.

Question 5	Can you specify what percentage of that time is spent by your team on analysing open-ended text responses from surveys and focus group conversations?

Since the focus for this project is using sentiment analysis for open-ended text responses in focus group conversations and survey responses, through this question the researcher wanted to know how much time specifically spent on analysing, tagging and classify open-ended responses to generate insights.

Question 6	What kind of tools and software are you guys using at X for analysing and tagging open ended text responses?

Through this question the researcher wants to know what kinds of tools and software systems the team currently uses (if any) for analysing open-ended text data. This will allow to research to setup a benchmark to compare the current processes with the proposed system and understand their limitations.

Question 7	Do you currently use any kind of machine learning or sentiment analysis techniques for analytics?

This question was chosen to understand if the company currently uses any machine learning and sentiment analysis models in their data analytics workflow.

Question 8	In your opinion, do you think the proposed system will be useful in improving the current workflow of open-ended text analysis?

Since the interview will be conducted with an expert, through this question the research wants to understand the general opinion and usefulness of the proposed system.

Question 9	In your opinion, do you think the business problem and objectives of the project have been defined properly?

The question was chosen the interviewee's opinion as a CTO of market research firm on whether the business problem defined by the researcher and the objectives set are feasible and realistic.

Question 10	In your opinion, can you highlight some of the limitations and challenges that this project may face?

The research chose to ask this question to discuss some of the limitations and challenges the project might face.

5.2.2 Questionnaire

This section will discuss and justify all the questions that was included in the written interview questionnaires developed for the X market research company analytics team.

Question 1	Can you describe your tasks on a daily basis?

This question was asked so that the team members can specifically describe what they exactly do in the team and some of their daily tasks.

Question 2	On average, what percentage of time do you spent on analysing open-ended text data?

This question was asked so that the researcher can understand how much time of the actual analytics process the analysts spend on analysing open-ended text data to get a rough estimate of how much time can be saved if the process can be automated using the proposed system.

Question 3	Describe the current process in place to analyse open ended text responses from surveys and online focus group conversations?

This question was included to understand the overall workflow of the analyst within the company to analyse open-ended responses, this will help the research understand the entire process of data collection to analytics for open-ended text data.

Question 4	Please list down the softwares you use to analyse open ended text data?

This question was added to understand what software packages the respondents currently use to analyse open-ended text data.

Question 5	If you listed any software above, please describe how you use them.

This question was asked by the researcher to better understand how the respondents uses the specified software in their workflow. This would help the research set a benchmark of how the

respondents were using the specified software in their workflow and whether the proposed system will help improve their workflow.

Question 6	Do you currently use any sentiment analysis in your workflow?

This question was asked to understand if the respondents use any sentiment analysis in their current workflow to analyse text data. This will help the researcher know if the respondents currently use any sort of sentiment analysis for analytics, so he understands whether having a sentiment analysis-based model will significantly change their workflow and help improve it.

Question 7	If you are using sentiment analysis, please describe how you currently use it?

This question was asked so that the researcher could further clarify on how the respondent use sentiment analysis in their workflow to draw a comparison with the functionalities of the proposed system.

6.0 Requirement Validation

In this section the researcher will analyse the data collected through the expert interview and questionnaire to draw up conclusions on whether the business problems highlighted in this project and the objective defined in line with actually business problem and challenges faced by the target users with the market research company. It will also help the researcher on gathering insight on whether the functional requirement set of the project are feasible and will actually help the issues and challenges faced by the target users. The first part of this section will analyse the responses collected from the expert interview followed by the response collected from the questionnaires distributed to the analytics teams at X market research company.

6.1 Analysis of data – Expert Interview

INTERVIEWER: ALAVI BEEN AZAM

INTERVIEWEE: MR. ASHRIQUE THEVENDRAN (CTO of X Market Research Company)

MODE OF INTERVIEW: ONLINE (GOOGLE MEET)

TOPIC: REQUIRMENT GATHERING AND VALIDATION FOR PROPOSED SYSTEM

Question 1	As being a market research company, can you kindly reflect on the importance of market research for brands and organizations?
Answer	Well very simply put, if you don't know who you are selling to you are not going to be able to sell. Business fundamentally optimizes for two which are reducing cost or increasing revenue, knowing your customer well will highly contribute to both things. Knowing the customer well through market research will ensure that you do not spend money on things which won't move the needle for you as a business, this helps business reduce cost because, each dollar is spent more efficiently and it also increases revenue because it helps you build things customers actually want.

According to the response from the interviewee, the value of market research for brand can be clearly identified and stated. As the expert works in the market research industry and works with several big and small brand and corporations, he is in a great position to highlight how important market research is to make business profitable and make better market decisions.

This clearly echoes the assumptions of the researcher on the importance of market research for companies and brands.

Question 2	Can you kindly talk about how technology has changed the market research industry in the last 10 years?
Answer	In general, old research techniques such as printed surveys and questionnaire and in-person focus group are still popular. But if you are taking about online techniques specifically, market research in the past 10 years haven't actually changed that much, the methodologies have not change, people still use the same tools such as surveys and focus groups but the major difference is that they are moving it from in-person to online basically, all other aspects in terms of the methodology and tools used still remain essentially the same. Also, currently the Covid19 pandemic has accelerated that process but the problem with this is, only is just not only another medium, it is a totally different paradigm and totally different way of doing things so if you typically just try to take your 100 question survey and put the same 100 questions survey online, the mode of engagement, the way respondents answer your question will be significantly different compared to if they are using a traditional pen and paper where they sit down and answer or researcher is sitting down with them to ask them these questions while they answer, in that situation it is way less like for respondents to turn-away from the survey and disengage from answering. The methodology hasn't change but the problem with that is the methodology we have currently is not completely suited for online. *The researcher asked a follow up question, asking whether the experts thinks that there should be new types of methodology for online based market research.* Yes, I think there should be a change in the methodology, For example we do not sell the same way we do in physical stores like we do in e-commerce websites, you do not have array of stores but you have companies fighting for shelf space in the sites. In the same way market research also changes when you are doing it online, it needs to be much more bite-size and quicker

	rather than 100 questions, if you have a long questionnaire it is more reasonable to cut it down into smaller questionnaires as with online the velocity of the whole process needs to be much faster and the size of the content needs to be reduced.

The interviewee has highlighted that even though the advancement of technology has moved several processes online over the years, the methodologies and tools used by market researcher inherently has remained the same. Surveys and focus groups still are the most effective and popular tools for conducting market research but he believes that the move to online is a whole new paradigm and changes need to be brought to the methodologies used to get more engagements and better responses online as the old techniques are not fully suited for online surveys and focus-groups.

Question 3	Can you briefly describe the workflow of your team from data collection to analysis while conducting online focus groups and surveys and the challenges faced?
Answer	*The interviewee answered this question with context of open-ended text data.* So mainly for analytics of open-ended text data, we as a company has struggled a bit with this. It has been quite hard to find good patterns on how to analyse it. Generally open-ended question gives you the best data because you are not constraining them to categories and answer, basically you are not constraining them to your assumptions. However, they are also the hardest to analyse because of a couple of factors. One of the challenges is language, people tend to mix languages when they are responding to a survey or questionnaire. Secondly, while you may collect open-ended responses, companies tend to ask for defined answer more toward a YES/NO context or choice A or choice B. For example, if a company wants to know if their products are expensive and you ask the respondents if they think product A is expensive, and give them a open-ended text box to respond in the ended the company does not want the open-ended opinion they just want to know if the consumer thinks the product is expensive or not in a very binary context. Finally, there is also a problem of human bias,

	researcher may have their own biases, it less like compared to regular human because as researcher they are more aware of biases, but it still tends to affect in some amount. So, at Vase, how we currently approach that is we have looked into the past 500+ projects that we have ran and we try to find common patterns on how the analysis on certain open-ended text are carried away and have developed an in-house ruled-based based system which most of the time the operations time will pass the data through the software we have built and the rules will just decide how open-ended text data will be analysed.

In this part, the interviewee has highlighted the struggles and challenges that is faced by his team in analysing open-ended text responses. Even though he strongly believes that open end response does produce better insights they are really time consuming to analyse and most client expect to provide binary and more defined answers rather that the deeper opinion of each respondent so they need to manually tag each response as being YES/NO or MAYBE and POSITIVE/NEGATIVE or NEUTRAL. Also, the task becomes more complicated due to the mixture of language and local slangs that are used in the response. So, to tackle this issue the team at X has developed a rule-based system to filter and classify response to generate insights. This also clearly highlights and are in per with the assumptions and problem defined by the research that open-ended text data in analysed manually or through rule-based making the process highly resource intensive. The interviewee has also discussed the issues of human bias introduced by researcher to the insight because they are done manually, even though its comparatively less, but the issues still exist all is in par with the business problems defined by this project.

Question 4	Can you kindly specify what is the average time frame for completion of an average analytics process from data collection to report generation?
Answer	The average time for completion for each project honestly depends on the size of the project and the type of questions being asked to the respondents. If the questions are more categorical or linear the time required for analysis is significant lesser since we have automated analysis tools for such responses compared to surveys which contain a lot of open-ended responses and may take 1-2 weeks to process.

The response shows that the time of completion is highly variable depending on data and the requirements of clients, but on an average takes a few weeks.

Question 5	Can you specify the amount of time spent on analysing open-ended responses for surveys and focus group conversations?
Answer	The variation is quite high on this one unfortunately as it depends on the types of questions that are being asked to the respondents and the size and scale of the survey. Typically, my team`s members obviously depending on their roles will spend around 25-50% of their time analysing just open-ended text responses from various sources. To put it better in numbers, each open-ended question in a survey adds about 5 hours of work for the analyst to analyse, so if a project has 5 open-ended questions for example, that add 25 hours of work.

The response clearly indicates analysing open-ended text responses is a very resource intensive and time-consuming process. As highlighted the analytics team typical spends 25-50% of their time analysing open-ended text responses depending on their role and each open-ended question in a survey takes roughly 5 hours of work to analyse. This clearly indicates how time consuming the process of analysing open-ended text is and lines up with the objectives of this project.

Question 6	What kind of tools and software are you guys using at X for analysing open-ended text responses?
Answer	Most of the team members usually use excel or sheets and the in-house rule-based system we have developed to analyse open-ended text data. We do have our own proprietary reporting tool but most of the open-ended stuff is done using excel and some other rule-based system we have developed in-house.

The response highlights the team only uses excel and sheets for most of their open-ended text analysis, which is mostly manual. The team also uses an in-house developed rule-based system to classify and filter responses and do not use any kind of automated tools.

Question 7	Do you currently use any machine learning or sentiment analysis tool and techniques for open-ended text analytics?
Answer	As I mentioned previously, we mostly do it manually or using the rule-based system we developed and nope we currently do not use and machine learning model or sentiment analysis tools in our workflow. However, we did have a couple of POCs from clients before but we found it does not work as well, for example if a respondent uses a mixture of Malay and English in their response and Malaysian slangs the sentiment analysis model kind off goes out of the window.

The response highlights that the company currently does not use and machine learning or sentiment analysis-based models in their workflow because most of the open-ended text responses have a combination of different language such as Malay and English and local slangs making it very difficult for sentiment analysis based model to classify and tag them properly. The accuracy was poor and not worth the effort.

Question 8	Can you highlight some of the limitations and challenges the researcher might face in this project?
Answer	As you mentioned that the target market for this system will be Malaysia, localisation in terms of language will be huge issue. So, it is better to use a local dataset from past surveys and online-focus group conversations and augment it with another publicly available dataset such as twitter dataset popular for sentiment analysis research. But it would be better if it's a Malaysian twitter dataset, that will help with the localisation issues.

The expert highlighted that since the target market for this tool is Malaysia, localisation will be a huge issue since Malaysian speak a number of different languages like English, Malay, Chinese and Tamil and like to use a mixture of different languages and local slangs in their responses making it very difficult for machine learning models to classify the responses. The expert also suggested that using local datasets from past survey and focus group conversations and other Malaysian datasets such as tweets data may somewhat help in developing an accurate model to automatically classify the responses using sentiment analysis.

Question 9	In your opinion, do you think the proposed system will be helpful in improving the current workflow of open-ended text analysis at X?
Answer	So, the interesting point about that is currently, tagging and classifying already takes us 5 hours per open-ended question, we do not sentiment analysis because adding that on, most clients don't ask us for. I think it is not because they do not want it, I think it's mostly because they are not aware that something like this is entirely possible. So, we do not current do it because clients have not asked us to do it specifically and because of the available sentiment analysis tools we have explored just adds more time to our workflow. So, if we do get something like this (the proposed system) available to us we will make it a part of our deliverables which I feel will make the clients happier and more satisfied. Even do we necessary do it know I don't not know if "that it will save time" but the hypothetically if we do implement something like might help us save some time or add a few extra hours. *As a follow-up question the researcher asked the interview that if the clients are not aware about sentiment analysis, what kind of consumer behavioural outputs do they expect from the research.* As I mentioned the clients do not necessarily know or are fully aware of what sentiment analysis technically is but they want the results to be a binary output from their respondents such as "yes this product is expensive" or "no this product is not expensive" to "it is kind of expensive given the quality" with obviously other demographics factors but mostly they look for a defined and clear output from us. So, if the insight is that 80% of the respondent say its expensive, company will take that into account and will think that no I cannot raise my price further and that how it will affect critical business decisions.

According to the opinion of the expert, the proposed system will have a welcome addition to the workflow of the analytics team and might be able to save some time of the open-ended text data can be analysed, classified, filtered and tagged accordingly. However, as the process already take a significant amount of time to complete (roughly 5 hours for each question) so if

the model adds on the that time it will not be feasible for team to use it in their workflow. This highly depends on the accurate and output of the system and needs to be tested out once the system is ready for demonstration. But nevertheless, the sentiment analysis-based model which can successful tag localised open-ended text responses will be a very helpful tool for the market research industry and will help increase client satisfaction.

6.2 Analysis of data – Questionnaire

MODE OF QUESTIONNAIRE: ONLINE (GOOGLE FORMS)
RESPONDENTS: DATA ANALYTICS TEAM (Vase Technologies - Market Research Company)

Question 1: Can you describe your tasks on a daily basis?

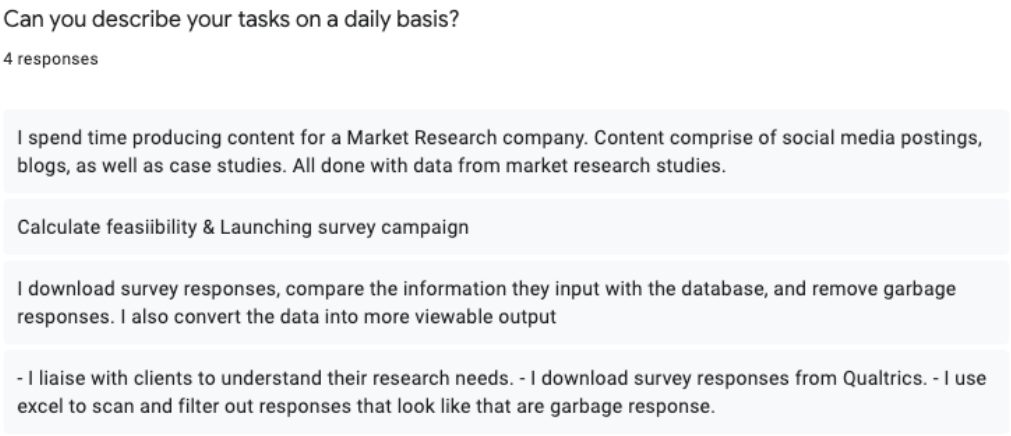


Figure 4 Screenshot for responses from question 1 (self-captured, 2020).

The responses show that most of the respondent work with open-ended text data as part of their data tasks in the company, making them a suitable candidate for this validation questionnaire to validate the responses of the proposed system.

Question 2: On average, what percentage of time do you spend on analysing open-ended text data?

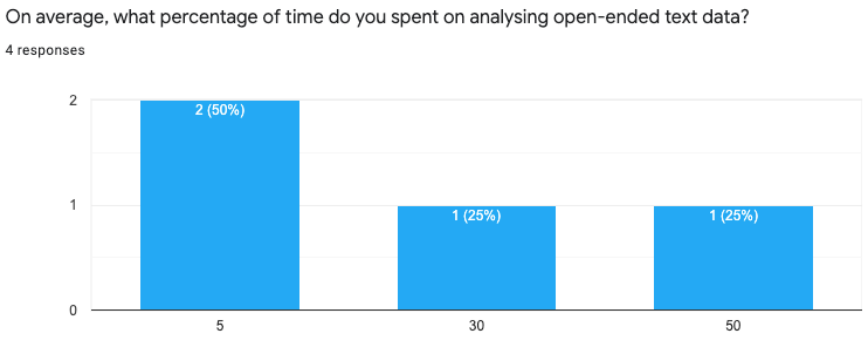


Figure 5 Screenshot of the responses from Q2 of the survey (self-captured, 2020).

The response highlights that all the respondents spend some time analysing open-ended text data but 50% of the respondents who work with analysing survey responses and conducts market research studies spend 30-50% of their time analysing open-ended text responses.

Question 3: Describe the current process in place to analyse open-ended text responses from surveys and online focus group conversations.

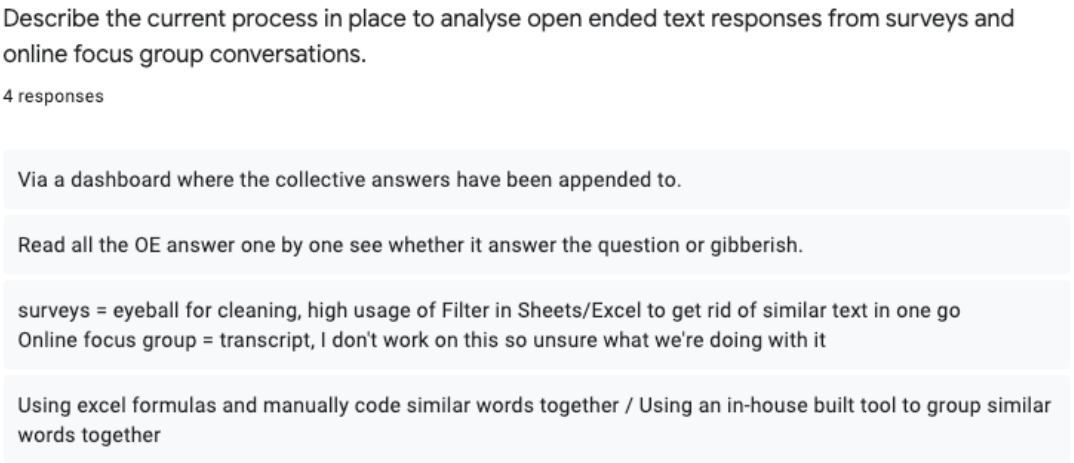


Figure 6 Screenshot of the responses from Q3 of the survey (self-captured, 2020).

The responses show that most of the task of analysing the open-ended text responses are carried out manually. The team either use eyeballing or excel/sheets formulas to filter the data. This

supports that researcher claim that most of the analysis conducted from open-ended text responses for surveys and online-focus groups are currently done manually by human researchers.

Question 4: Please list down the software you use to analyse open ended text data? If you don't please put in N/A.

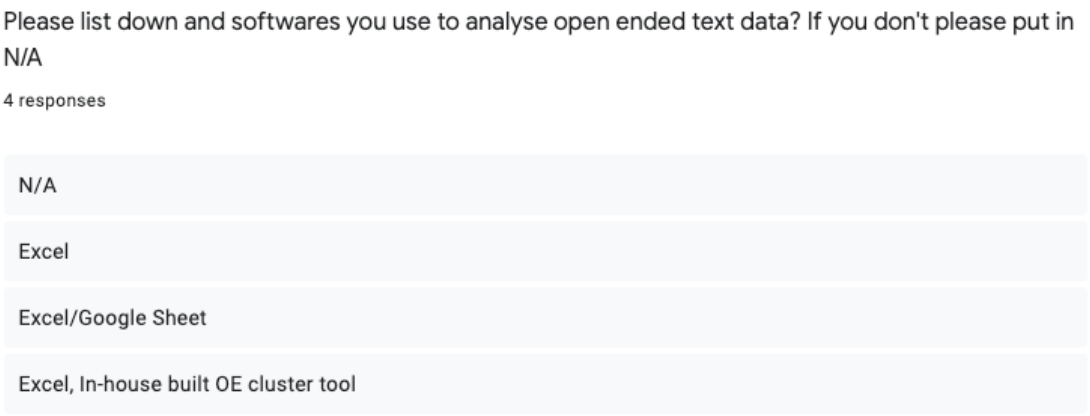


Figure 7 Screenshot of the responses from Q4 of the survey (self-captured, 2020).

The responses show that Microsoft excel, and Google Sheets are the most used software packages in analysing open-ended text responses, followed by an in-house built rule-based OE clustering tool. These further echoes the claim of the researcher that no machine learning based models and tool are currently popular for analysing open-ended text data in Malaysia.

Question 5: If you listed any software above, please describe how you use them.

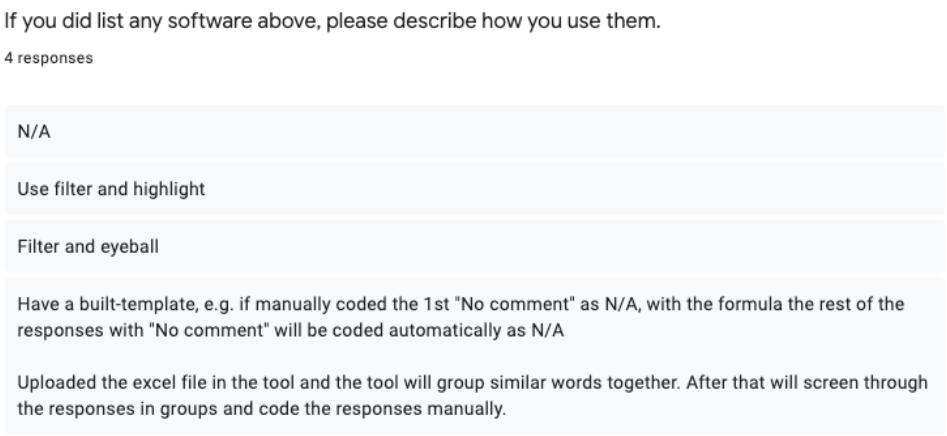


Figure 8 Screenshot of the responses from Q5 of the survey (self-captured, 2020).

The responses show as no automated tools are used in the process, Excel and sheet formulas and templates are the most popular option for filtering out responses. Also, eyeballing and manual processing are also used to filter out the responses. This further shows that an automated system will significantly help with speeding up the process and solve some of the business problems of manually combing through the data for researcher which is massively time intensive.

Question 6: Do you currently use sentiment analysis as part of your workflow?

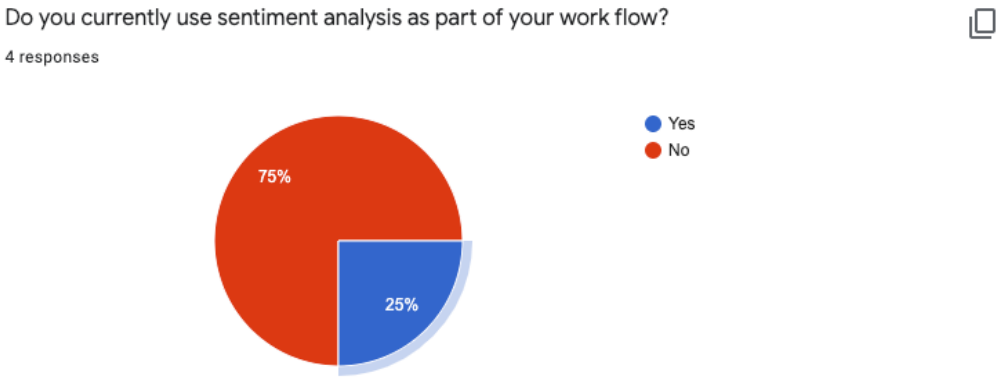


Figure 9 Screenshot of the responses from Q6 of the survey (self-captured, 2020).

The responses show that 75% of respondents do not use any sentiment analysis tools or techniques for their workflow. This highlights there is scope for an automated sentiment analysis tool as proposed in the project for the market research industry in Malaysia.

Question 7 : If you are using sentiment analysis, please describe how you currently use it?

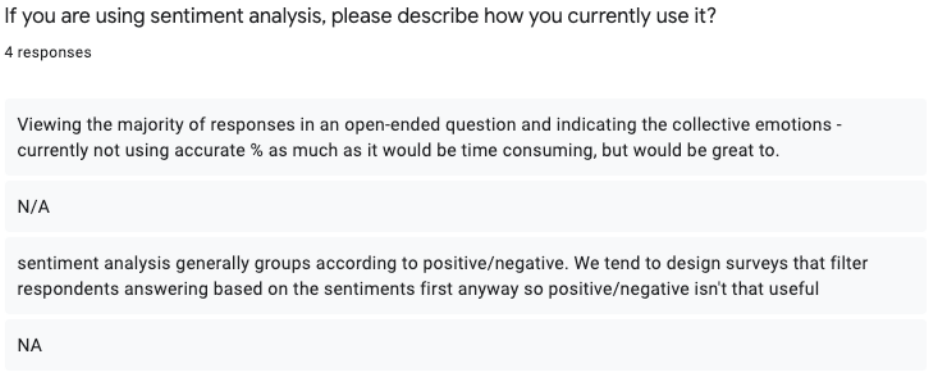


Figure 10 Screenshot of the responses from Q7 of the survey (self-captured, 2020).

The response show that most do not use any kind of sentiment analysis in their current workflow. However, some mentioned the team designs the questionnaires in a way that some automatically filter some responses as positive or negative. Some also mentioned they classification of sentiment done manually by reading all the open-ended responses one by one and tagging the sentiments. They also mentioned a sentiment analysis-based model will not be feasible since the accuracy isn't good enough.

6.3 Summary

After conducting a thorough analysis of the data collected from the expert interview and questionnaire is that currently in the market research industry the researcher faces a severe challenge in analysis and tagging open-ended text responses. Even though open-ended questions in surveys and focus groups generate the best insights they are significantly time consuming to analysis as the expert interview highlighted that each open-ended interview question in survey takes up to 5 hours of analysis which shows so resource intensive the task is. Also, the analysis has further highlighted on the facts that most of the analysis for opened-ended text data are carried out manually using eyeballing. However statistical software packages such as Microsoft Excel and Google Sheets are popular choices for researcher to conduct text-based analysis using formulas and templates.

The analysis also showed that no machine learning or sentiment analysis models are currently used in the workflow due to the lack of accuracy and extra work required. However, in-house built rule-based classifier are used to tag and classify some open-ended text responses which helps the process. The analytics teams mostly rely on statistical software and human researcher to manually generate insights and produce reports.

The analysis also shows that some of the challenges and limitation that will be faced by the project would be localisation as the target market for the proposed system is Malaysia where the respondents use multiple languages and local slangs making automatic classification difficult for machine learning models. However, using local datasets containing historical Malaysia text data might help in making the model more accurate in generating better results.

The analysis has been in per with most of the claims made by the research and also has yielded to similar business problems and challenges from the data collected and analysis that has been

carried out which validates the objectives set for this project. The analysis also shows that there is welcoming room for such an automated sentiment analysis-based system in the Malaysian market research industry provided in produces accurate results and help cut down and analysis time for open-ended text responses from online surveys and questionnaires. The analysis has helped the researcher validate all the assumptions made and also solidify the business problems and objective defined for this project.

7.0 System Architecture

This section of the paper will provide a detailed outline of the core features of the proposed system (Focusen) and all provide details about the abstract architecture of the system.

7.1 Introduction

Focusen: real-time sentiment analysis to understand consumer behaviour is a system designed mainly for market-research companies and marketing teams within companies to help them better understand their consumer through data collected from various source such as social-media, online focus group conversations and surveys. As mentioned in the sections above, it has become increasingly important for companies and brands to understand their consumer in order remain competitive and profitable in the market and one of the key ways to achieve that is by providing customer satisfaction through understanding their opinion and listening to their feedback. As the amount of data points available on consumer have grown rapidly over the past decade, it provides both an opportunity and challenges for market researchers and brands to comb through this vast amount of data and generate actionable insights to influence business decisions. Focusen was designed to meticulously meet the needs of its users and provide solutions to their challenges which have been pin-pointed and validated by the research and analysis that has been conducted by the researcher as thoroughly discussed in the section above. The goal of Focusen, is the provide market researchers and marketing teams with a fast, reliable, consistent and scalable way of analysing the vast of data they are collecting to understand consumer behaviour in real time by utilizing the powers of AI (artificial intelligence) and sentiment analysis to generate accurate insights in real-time which can help influence data-driven business decisions. The core-features of the system will be discussed below:

1. **Real-time Data Ingestion Pipeline**

One of the features of the system is that, Focusen provides a real-time data ingestion pipeline which is capable of collecting and processing text data from various sources using an API and storing it in a database. This is crucial components as users can use the system to easily collect data from various sources such as social-media, websites, forums, blogs and other publicly available data sources where consumers are constantly providing their opinions on the different

brands and services they are using. The pipeline can also be integrated with proprietary data collection platforms such as online-focus group and survey systems (example: Every by Vase) using an API to directly stream the data for real-time processing and insights. To demonstrate this feature of the system, the developer has used to Twitter API to collect and process tweets in real-time which are then stored in a relational database management system (RDMS) for further processing.

2. Data Processing & Sentiment Analysis

The other core feature of Focusen is that the system can take the data collected from the specified source using an API and process the text-data in real-time to prepare it to be feed into model for sentiment classifications. In this step of the process the system uses different python libraries such as NLTK (Natural Language Toolkit), TextBlob and RE (Regular Expression) to pre-process the text data performing a set of standard text pre-processing procedures such as tokenization, normalization, removal of unwanted characters and etc. to prepare the data for classification. After the text has been pre-processed Focusen uses a pre-trained sentiment analysis classification model which under the hood uses a Naïve Bayes Classifier to classify the text-data according to its polarity and subjectivity. All of these processing happens in real-time without any noticeable lag in the system providing a fast, reliable and scalable solution for market researcher for a task that would usually takes weeks if not months to be done manually while still having drawbacks such as human errors and biases. The system provides an automated solution which uses sentiment analysis to generate reliable insights in seconds compared to weeks and eliminating issues such as biases and human errors.

3. Real-time Insights Dashboard

There is no point of having a real-time data pipeline or an automated and highly accurate sentiment analysis classification system if the results and insights are not readily available for marketer and business to make informed data-driven decisions. To tackle this challenge, Focusen provides a real-time insights dashboard for brands and market researcher to constantly monitor what consumers and media is saying about their brands or certain topics. The Focusen dashboard provides a time-series analysis graph which tracks every positive, negative or neutral statement made online about the topic, brand, product or service that's being tracked by the system. To demonstrate the functionalities of the Focusen dashboard the researcher has

used the Twitter API to collected tweets about a certain Keyboard that can be specified by the user. The system generates a time-series scatter plot to track all the positive, negative and neutral tweets in real-time according to their polarity (-1 being negative, 0 being neutral and 1 being a positive sentiment) and subjectivity. The system also provides details about the number of tweets posted in a certain timeframe, the potential impression (reach) of the tweets and also that percentage of positive, negative and neutral tweets for a specific amount of time which can be programmed into the system. Additional features of the dashboard also include Keyword tracking which can be used to track the top keywords that has been mentioned in the tweets which can be used by researcher and brands to pin-point the subjectivity behind consumer`s opinion and factors that are affecting consumer sentiment towards the topic in real-time. The dashboard also provides a geographic segmentation features demonstrated by using USA as an example which help marketers and researcher better understand the location and population distribution of their consumers as the system provides a state-wise representation showing amount of activity by consumer in each state towards the specified topic. More details of about each of the features will be discussed on the implementation section of this paper.

7.2 Abstract Architecture

This section will discuss about the abstract design and architecture of the proposed application to give a more detailed look into the design planning process of developing the Focusen application.

7.2.1 System Design – Focusen Use Case Diagram

The diagram below shows the use-case diagram for the Focusen application. The main users (actors) of Focusen are researchers, marketers and business leaders.

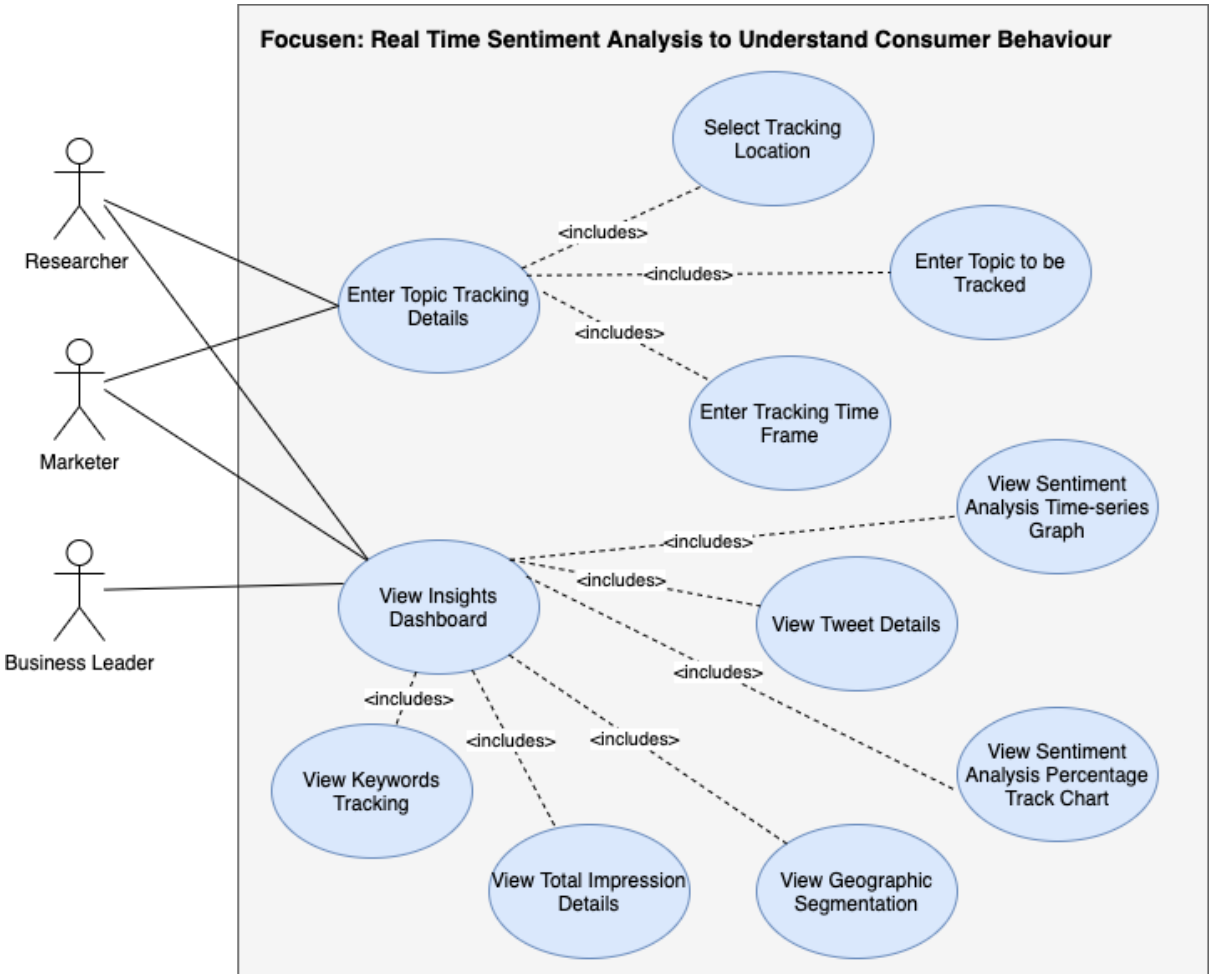


Figure 11 Use-case diagram for Focusen application (self-captured,2020).

The section below will provide an overview for the specifications for the use-cases of the Focusen: Real Time Sentiment Analysis to Understand Consumer Behaviour application:

Function Name	Enter topic tracking details
Description	The application will allow users to input the details about the topic or keyword they want to track using the system.
Actor	Researcher, Marketer
Assumptions	The user will only want to track one topic or keywords at a time using the system.
Default Workflow	Actor enters the details of the topic, time-frame and geographic location for the system to track and filter.

Function Name	View insights dashboard
Description	The web-application will provide an interactive dashboard for users to view the sentiment analysis results in real-time.
Actor	Researcher, Marketer, Business Leader
Assumptions	The user will view the results for only one topic or keywords tracking at a given time.
Default Workflow	Actor connects to the application using the URL to view the interactive dashboard.

Function Name	View tweet details
Description	This function can be used by users to get an overview of the tweets being tracked such as total no. of tweets, impressions and tweet frequency change over a given time-frame.
Actor	Researcher, Marketer, Business Leader
Assumptions	The user will want an overall summary of the tweet details and not at a micro level.
Default Workflow	Actor view tweet details using the dashboard.

Function Name	View keyword tracking

Description	User can use the dashboard to track the most-popular keywords in the tweets according to the topic specified using a frequency histogram.
Actor	Researcher, Marketer, Business Leader
Assumptions	Keyword tracking will help users understand the current trends and topics that are influencing user sentiment.
Default Workflow	Actors view keyword-tracking histogram using the dashboard.

Function Name	View geographic segmentation
Description	This feature can be used by the users to view the location of their consumers and which geographic areas have the most activity using a heat-map of the country or state being tracked.
Actor	Researcher, Marketer, Business Leader
Assumptions	Geographic segmentation will help researchers and brands better understand the demographics of their consumers and make more informed decisions based on location data.
Default Workflow	Actors view geographic segmentation heat-map using the Focusen dashboard.

7.2.2 System Design – Focusen Activity Diagram

The diagram below shows the system activity diagram from Focusen: Real-time Sentiment Analysis to Understand Consumer Behaviour Application.

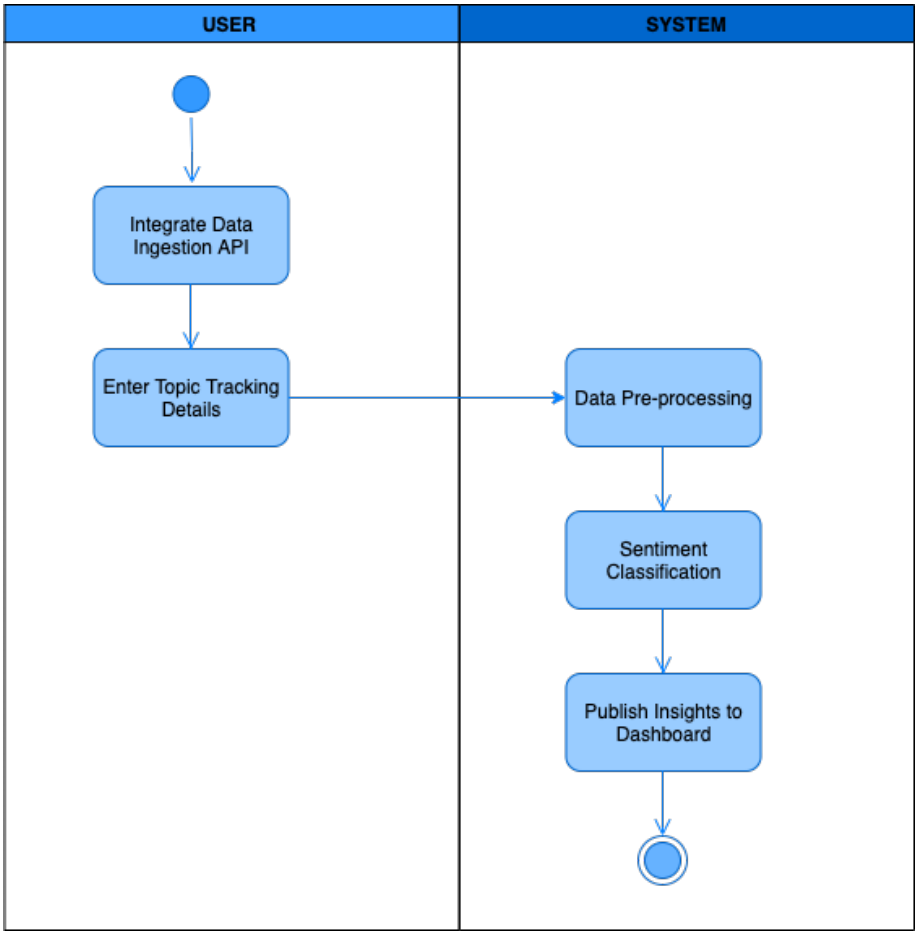


Figure 12 Focusen system activity diagram (Self-captured,2020).

7.3 Focusen Database Design

This section will discuss about the database architecture of the Focusen application. The database design is an essential component of any application development since it highly contributes to the efficiency, reliability and scalability of the application. As mentioned in the technical research section, the developer has decided to use a relational database management system (RDMS) for this application due to the structured nature of the data being ingested and stored in the tables.

7.3.1 Entity Relationship Diagram (ERD)

The image below shows the entity relationship diagram (ERD) for the Focusen system database, which provides an overview for the structure of the database:

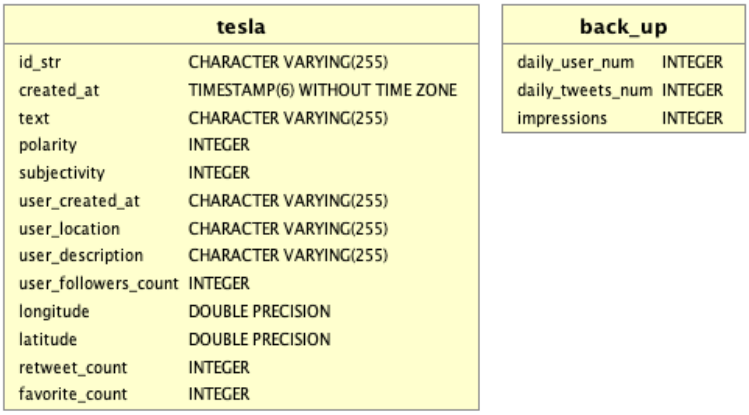


Figure 13 Focusen system database ERD diagram (self-captured, 2020).

Since the database requirements of the application is fairly simple as it is only being used to store and retrieve the tweets data from the Twitter API there are no dependencies our relationships between the tables in the database. SQL quires are being used in the code to input data in the tables and an aggregator function and SQL query is being used to populate the backup table which contain summary attributes for data in the main (topic/keyword) table. The system uses a relational PostgreSQL database in deployment.

7.3.2 Database Table Structure

This section will provide the details of the table structures in the Focusen application database.

Table Name: Tesla (Topic/Keyword)

Column Name	Description	Data Type	Remarks
id_str	Unique ID for the tweets from fetched using the Twitter API.	VARCHAR (255)	
created_at	Timestamp of when the tweet is created.	TIMESTAMPTZ (6)	Used or time-series analysis
text	Text body of the tweet.	VARCHAR (255)	

polarity	Polarity score of the tweet.	INTEGER	Calculated using ML model.
subjectivity	Subjectivity rating of the tweet.	INTEGER	Calculated using ML model.
user_created_at	Creation date of Twitter user account.	VARCHAR (255)	
user_location	Location of the Twitter user.	VARCHAR (255)	Used for geographic segmentation.
user_description	Description of Twitter user account.	VARCHAR (255)	
user_followers_count	Total follower count of the user.	INTEGER	
longitude	Longitude for user location.	DOUBLE PRECISION	Used for geographic segmentation.
latitude	Latitude for user location.	DOUBLE PRECISION	Used for geographic segmentation.
retweet_count	Number of total Retweets.	INTEGER	
favourite_count	Number of total favourites count.	INTEGER	

Table Name: Back_Up

Column Name	**Description**	**Data Type**	**Remarks**
daily_users_num	Daily number of active Twitter users while tracking is enabled.	INTEGER	
daily_tweets_num	Total number of tweets which include tracked tweets.	INTEGER	
impressions	Total impression from tweets	INTEGER	

7.4 Interface Design

Interface design and wireframing is a critical phase during the process of software development as it provides developer a way to think and create to best UI/UX experience that will solve user problem while also fulfilling the functional business requirements of the system. The image below shows a low fidelity wireframe designed for the Focusen application dashboard:

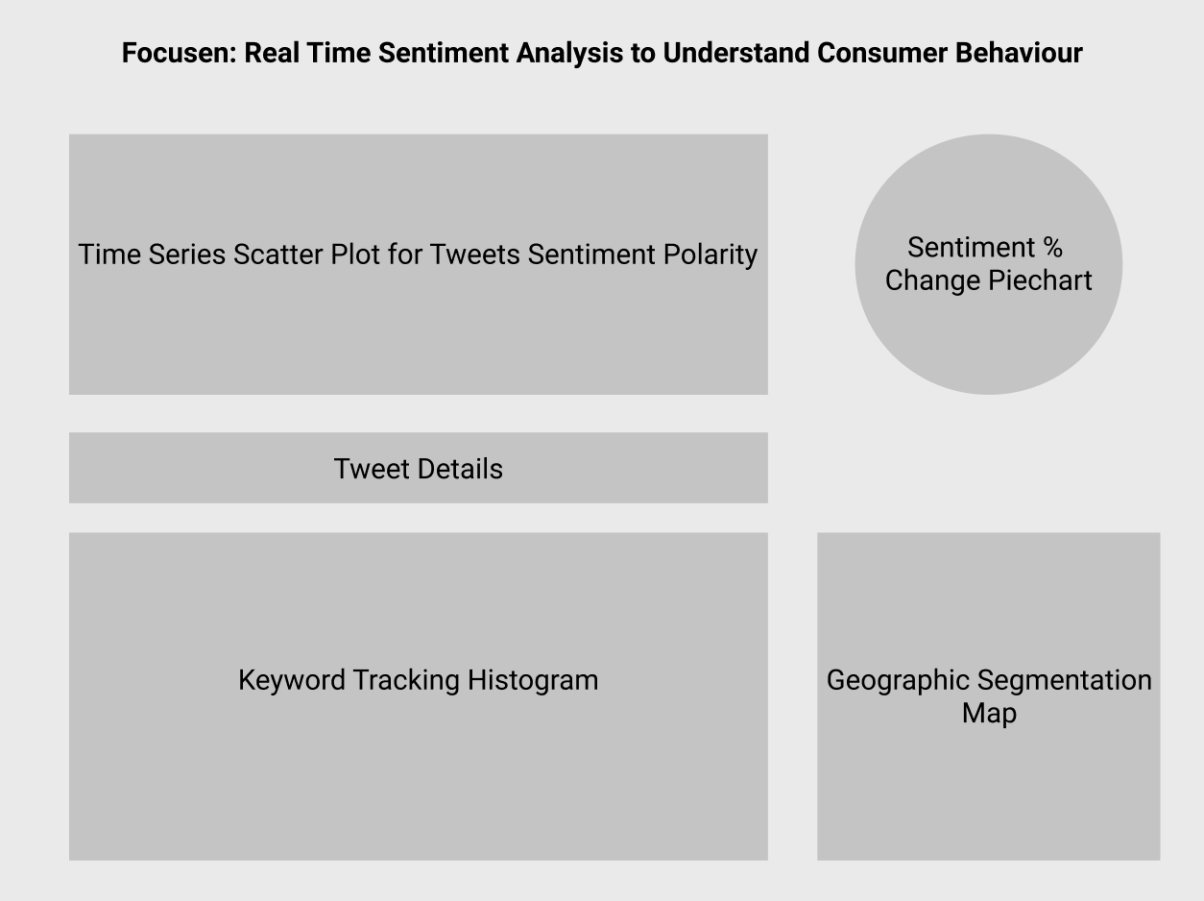


Figure 14 Low-fidelity wireframe for Focusen dashboard design (Self-captured,2020).

8.0 Project Plan

In this section of the paper the developer will discuss the detailed release plan for the proposed system. As mentioned in the project methodology section that project is divided into two components, the data science and machine learning component and the web development component. A detailed project timeline for this project is provided using a Gantt chart in the appendix section. The project will be released according to the several milestones that have been set for this project. The detailed release plan for this project will be discussed below.

8.1 Release Plan

This section will provide the details about the different version release plans for Focusen: Real-time sentiment analysis to understand consumer behaviour project.

8.1.1 Focusen Version 1.1

After the developer has conducted through technical research on the different libraries and packages available that can be utilised to successfully complete the proposed system. The developer has designed release plan for this project. The first version of the project which is due to be released in the first week of November 2020 will consist of a fully functional data ingestion pipeline component of this application developed using a local instance of a Jupyter notebook. Since the first phases of the project consist of the data science and machine learning components of the project, all the development and testing will be conducted using Jupyter notebook. As mentioned previously, for the demonstration of this project the developer has decided to use the Twitter API to fetch tweets in real-time to perform sentiment analysis the developer will use a RDMS and a python library called Tweepy to successfully develop and integrate the data ingestion pipeline. In the version 1 release, users will be able to enter their unique Twitter API authentication credentials and also define the topic or keyword they want to track in Twitter and the test will be able to fetch and display the tweets in real-time in version 1 of the Focusen system release.

8.1.2 Focusen Version 1.2

In the second release of the project which is due to be released on the second week on November 2020, the developer will release the fully-functional data pre-processing and ML

(machine learning) components of the project. In this phase the data the application will successfully be able to ingest the text data using the Twitter API in real-time, process the data using the NLTK python library to tokenize, normalize and format the data after which a pre-trained Textblob sentiment analysis ML model will be used to classify the sentiment of the Tweets using polarity and subjectivity after which the data will be stored in an SQL table for further processing and visualization.

8.1.3 Focusen Version 1.3

In the third release of the project, the developer will release the fully-functional data visualization component of the project which will be released on the third week of November 2020. The release will consist of a fully operational dashboard in Jupyter notebook which will contain several infographics and visualizations as discussed in the interface design section. The system will be able to show a time-series scatter plot tracking the sentiments of the Tweets in real-time. The version will also contain a fully-functional keyword tracking histogram which can be used to track the most popular keywords in the tweets posted about the topic or keyword being tracked using the system. A geographical segmentation map will also be available in this release for users to track the location their consumers to check which states in the USA has the most and least activity. The third release will have a full-functional version of Focusen: Real-time sentiment analysis to understand consumer behaviour and will be the final release for the data science and machine learning (ML) component of this project.

8.1.4 Focusen Version 2.0 (Demo)

The final release for this project timeline will be Focusen Version 2.0 which will be the final demo version of this project and will be released on the first week of December 2020. In this version of the project the developer will deploy and release a fully-functional web-based dashboard for this project available for demonstration and usage by the target user of this proposed system. The dashboard will contain all the functionalities of the Version 1.3 available on a interactive dashboard with a user-friendly UI to make it functional for end users. The version will contain same added features specific to the final version such as a detailed summary of the tweets, an interactive pie-chart to track the change in tweet sentiments over a specified time-frame and other. This release will be the final release for this research which will be used to demonstrate this project.

8.2 Project Test Plan

A test plan is a critical component for any project as it ensures that all the components of the system are functioning properly. As the developer has adopted a test-driven development strategy, it is critical that the all the components of the application are functional properly before the final version of the project is released. Details of the testing plan for the Focusen application has will be discussed below:

- To ensure this, the developer will conduct unit testing for each component of the system to ensure all the components are functioning properly and are providing the expected results. A detailed case by case unit testing plan will be used to conduct the unit testing for the Focusen application, the detailed plan will be discussed in the following section.
- After each unit of the system has been tested individually the developer will move on to integration testing where all the different units of the application are integrated together and checked to see whether there are any issues or bugs in the system after all the modules have be integrated together. A detailed integration testing plan will be discussed in the following section.
- Finally, after the system passes all the test cases for unit and integration testing and is ready for deployment, the developer will move into the user-testing phase which is another critical step for a test-driven development approach. Even though the system has passed both unit and integration test and is deemed fully functional, it is very important to conduct user testing with the target users of this system to ensure that all the users are comfortable while using the system and all their functional and business requirements have been met using the system. To ensure a valid user testing scenario and get genuine feedback on the system the developer will conduct the user testing with a data scientist, a market researcher, a business leader and a technical solutions architect who are experts in the field and are also the target user of the proposed Focusen application. A detailed user-testing plan for the system will be discussed in the section below.

8.2.1 Unit Testing Plan

The table below shows the detailed case by case unit testing plan for the Focusen system.

Case ID	Test Case	Test Function	Sample Data	Expected Result
1	Receiving data using Twitter API while the application is running.	Data Ingestion Pipeline	Twitter API	Successful data streaming incoming.
2	Automatic SQL table creation according to keyword being tracked.	Data Ingestion Pipeline	Twitter API	Table has been created in the SQL database according to the topic being tracked.
3	Data being entered properly in the current table.	Data Ingestion Pipeline	Twitter API	Data is being written in the correct table according to the keyword/topic that is being tracked.
4	Data stream is in real-time.	Data Ingestion Pipeline	Twitter API	Data in the table is correct and matches current UCT time.
5	SQL CRUD queries are functional.	Data Pipeline	Twitter API	All the SQL queries are executing correctly.

6	Tweet attributes extraction from Twitter API.	Data Ingestion Pipeline	Twitter API	All the attributes are correctly being extracted and stored in the database.
7	Database connection established.	Data Pipeline	Twitter API	Database connection has been established successfully with the local SQL database.
8	Text Tokenization is functional.	Data Pre-processing	Twitter API	Tweets are being tokenized correctly.
9	Tweet stemming is functional.	Data Pre-processing	Twitter API	Text stemming is accurate using python regular expression (RE).
10	Removal of emoji and unwanted characters from tweets.	Data Pre-processing	Twitter API	Emojis and unwanted characters such as punctuation are being removed properly before being stored in the database.
11	ML model is functional	Machine Learning	Twitter API	ML model is classify the tweets and proving a result.

12	Accuracy of the ML model.	Machine Learning	Twitter API	The accuracy of the output from the ML model is standard and acceptable.
13	Check sentiment classification output.	Machine Learning	Twitter API	Tweets are being classified properly according using their polarity and subjectivity.
14	UCT to PDT time conversion is correct.	Data Pre-processing	Twitter API	The creation time for tweets are successfully being converted from UCT time to PDT time.
15	Check scatter plot is functional and time series distribution is accurate.	Dashboard	Twitter API	The scatter plot is functional and display classification in real-time.
16	Sentiment Analysis Pie-chart is functional.	Dashboard	Twitter API	Sentiment Analysis summary pie-chart is current according to the time frame defined.
17	Twitter summary data is accurate (no. of tweets,	Dashboard	Twitter API	Twitter summary data is accurate according to the

	impressions, change in 10 mins)			input time-frame defined in the system.
18	Topic Tracking is accurate.	Dashboard	Twitter API	The correct topic is tracked by the system as defined by the user.
19	US geographical segmentation map is functional.	Dashboard	Twitter API	The USA geo-segmentation frequency heat-map is updating properly.
20	Top keyword tracking is functional.	Dashboard	Twitter API	Top keywords are being tracked currently according to the topic defined.

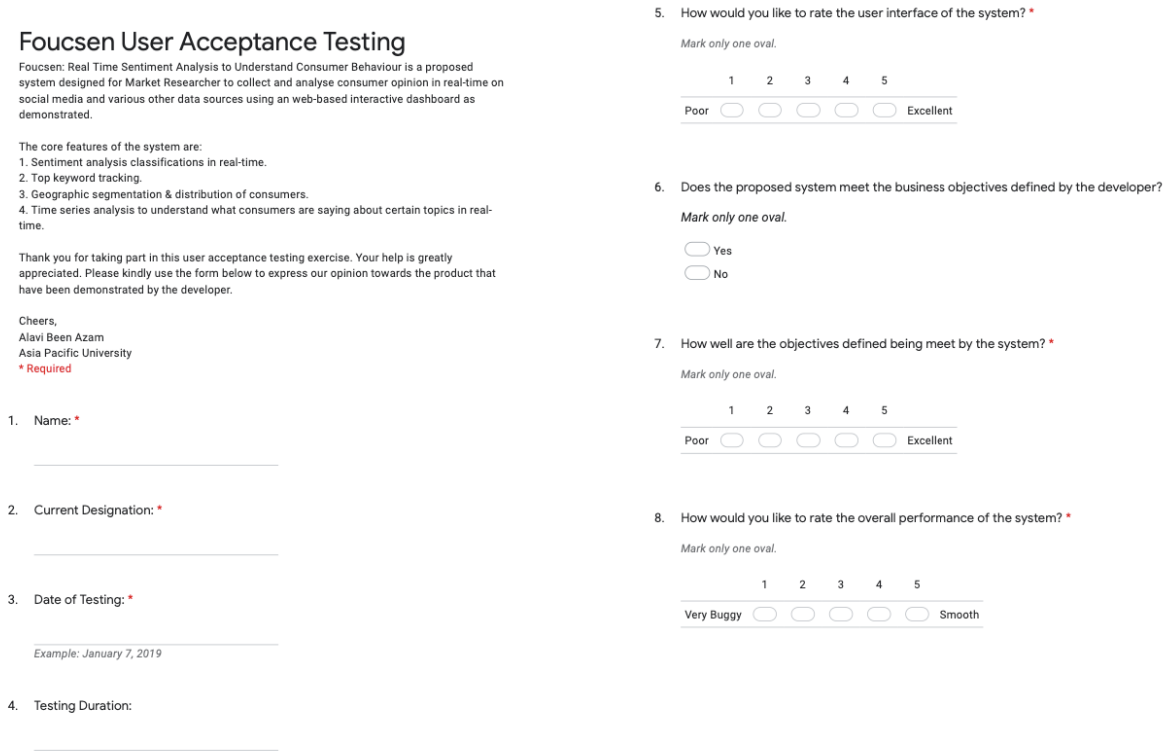
8.2.2 Integration Testing Plan

This section will provide the detailed integration testing plan for the Focusen application to ensure all the modules of the system are integrated currently and are fully functional.

Case ID	Test Case	Test Function	Sample Data	Expected Result
1	Twitter API integration is successful.	Data Ingestion Pipeline	Twitter API	The Twitter API has been successfully integrated into the application.
2	Database integration is successful.	Data Pipeline	Twitter API	The SQL database has been successfully integrated with the python application.
3	Focusen dashboard is functional.	Dashboard	Twitter API	The front-end dashboard has been successfully connected to the backend application using python Dash library.
4	Model integration	Machine Learning	Twitter API	The ML model has been successfully deployed to the frontend dashboard of the application.

8.2.3 User Acceptance Testing Plan

This section will provide the details about the user acceptance testing plan for the application to ensure the users are comfortable with using the system and the application meets will the business functional requirements that have been defined throughout the research. Due to the on-going Covid19 pandemic lockdowns and social-distancing rules, the user-acceptance testing was conducted online via video-conference and user feedback was recorded using a google form. The image below shows the template of the form used to collect user feedback after system demonstration:



Foucsen User Acceptance Testing

Foucsen: Real Time Sentiment Analysis to Understand Consumer Behaviour is a proposed system designed for Market Researcher to collect and analyse consumer opinion in real-time on social media and various other data sources using an web-based interactive dashboard as demonstrated.

The core features of the system are:
1. Sentiment analysis classifications in real-time.
2. Top keyword tracking.
3. Geographic segmentation & distribution of consumers.
4. Time series analysis to understand what consumers are saying about certain topics in real-time.

Thank you for taking part in this user acceptance testing exercise. Your help is greatly appreciated. Please kindly use the form below to express our opinion towards the product that have been demonstrated by the developer.

Cheers,
Alavi Been Azam
Asia Pacific University
* Required

1. Name: *
2. Current Designation: *
3. Date of Testing: *
   Example: January 7, 2019
4. Testing Duration:
5. How would you like to rate the user interface of the system? *
   Mark only one oval.
   1 2 3 4 5
   Poor ( ) ( ) ( ) ( ) ( ) Excellent
6. Does the proposed system meet the business objectives defined by the developer?
   Mark only one oval.
   ( ) Yes
   ( ) No
7. How well are the objectives defined being meet by the system? *
   Mark only one oval.
   1 2 3 4 5
   Poor ( ) ( ) ( ) ( ) ( ) Excellent
8. How would you like to rate the overall performance of the system? *
   Mark only one oval.
   1 2 3 4 5
   Very Buggy ( ) ( ) ( ) ( ) ( ) Smooth

9. How would you like to rate the user friendliness of the system?

Mark only one oval.

	1	2	3	4	5	
Poor	◯	◯	◯	◯	◯	Excellet

10. Feedback from Tester: *

11. Additional comments: *

This content is neither created nor endorsed by Google.

Google Forms

Note: A high-quality version of the user acceptance testing form has been attached in the appendix section.

9.0 Implementation

In this section the developer will provide a detailed walk-through of the core functionalities of the application using screenshots also discuss the code that has been used to develop the Focusen: Real-time Sentiment Analysis to Understand Consumer Behaviour.

9.1 UI Design and Outputs

9.1.1 Twitter Data Streaming

As mentioned earlier, the project has 2 distinct components, first the data-science and machine learning component and second, the web dashboard component. The screenshot below shows the data being streamed from Twitter using the Twitter API and displayed into the Focusen application.

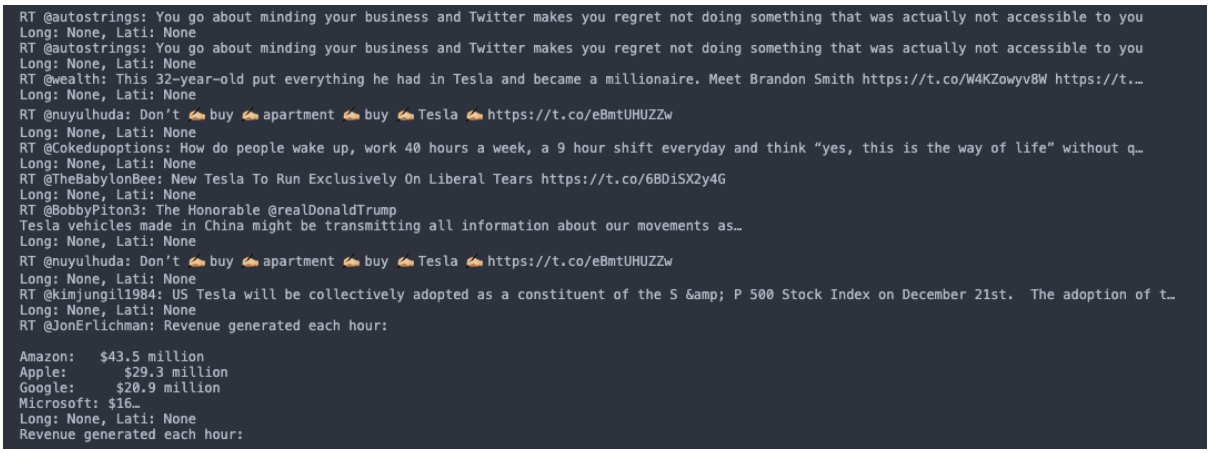


Figure 15 Focusen Tweets streaming (Self-captured, 2020).

The screenshot above shows the brand "Tesla" being tracked in real-time on Twitter. The result shows the tweets that have been posted during the actual time on when the screenshot was taken. The data stream also shows details about the tweets that have been retweeted by other twitter users giving market researcher and brands an easy way to track what exactly consumers are saying about their brands in real-time on social media.

9.1.2 Sentiment Analysis Scatter Plot

The screenshot below shows the time-series scatter plot on the Focusen application Dashboard tracking consumer sentiment in real-time using Twitter data feed.

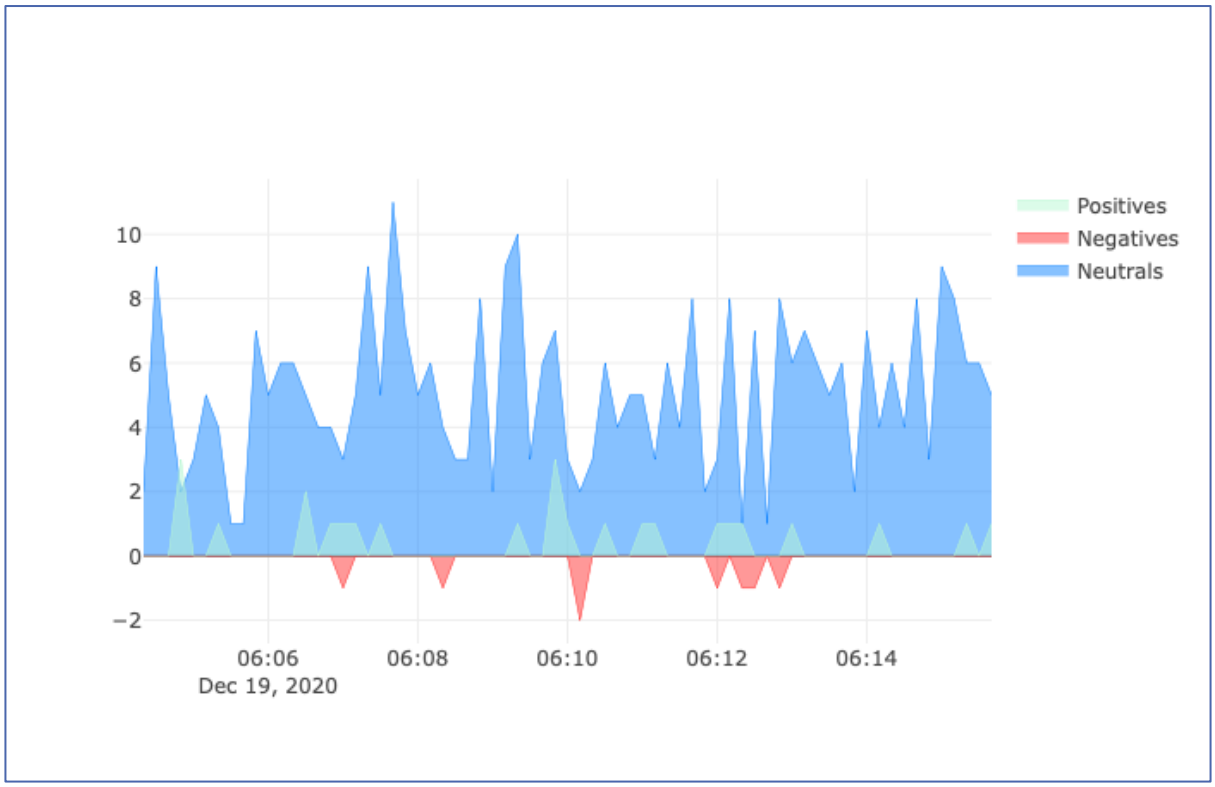


Figure 16 Scatter plotting showing real-time twitter sentiment analysis on Focusen dashboard (Self-captured, 2020).

As shown above the sentiment analysis model is being used to tag and categorized tweets in real-time which are then displayed using the scatter plot above in real-time. This gives users an easy way to track their brands on social-media and understand the sentiment of their consumers in real-time using the polarity and subjectivity of the tweets. As shown, the neutral tweets are plotted in blue, positive in green and negative in red according to their polarity. User can also hover over the graph to see the frequency of exactly how many positive, negative or neutral tweets were posted at any given timeframe using the plot. Users can also choose to easily save an image of the plot at any given time with just one click.

9.1.3 Top Keyword Tracking

The screenshot below shows the top keywords tracking in tweets feature of the Focusen application.

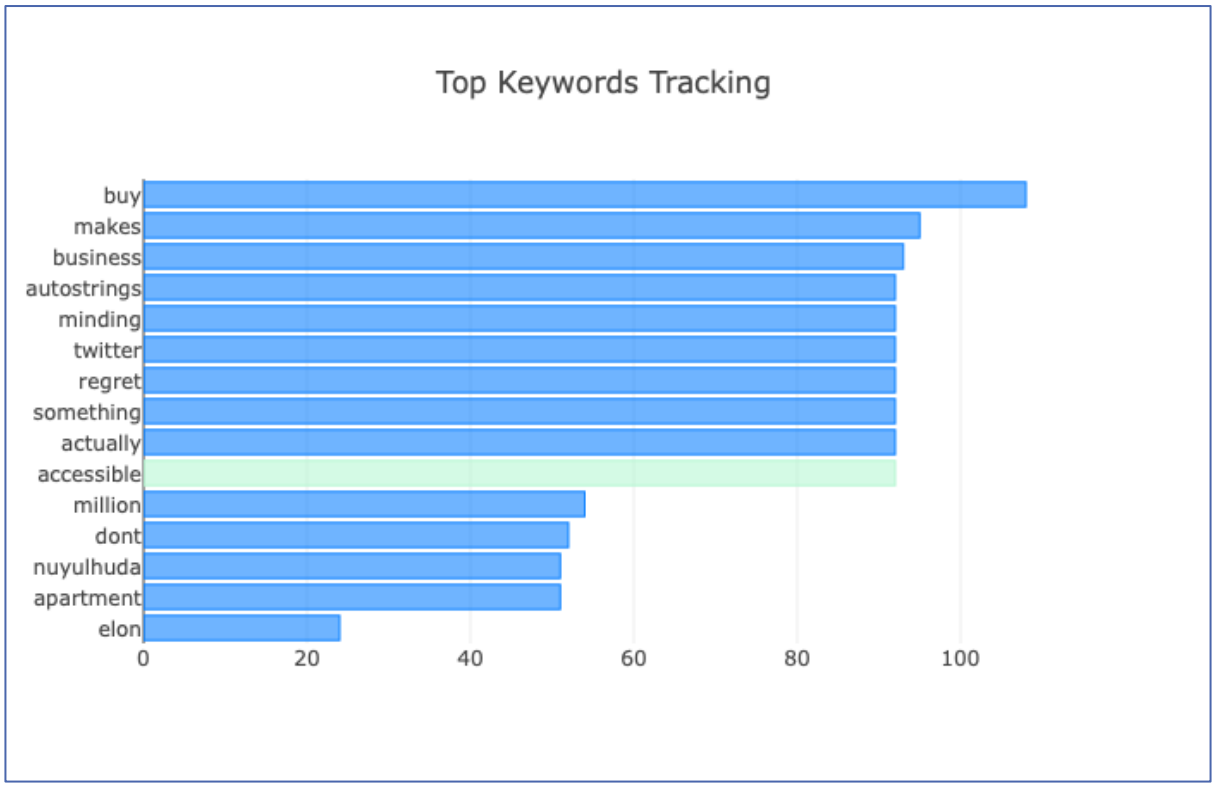


Figure 17 Top Keyword Tracking on Focusen (Self-captured, 2020).

The Focusen dashboard as mentioned before can also be used to track the top positive, negative and neutral keywords in tweets in real-time which is shown using the histogram above. User can easily use this histogram to track to most popular words consumers are using in their sentences while posting about their brands or the topic they are tracking using the application. The chart displays the top keywords against their frequency and positive words are highlighted in green, negatives in red while neutral words are highlighted in blue giving users (researchers, marketers and business leaders) to quickly understand to sentiment of consumers on social media platforms. Currently the example shows the company “Tesla” being tracked on twitter.

9.1.4 Consumer Geographic Segmentation

The screenshot below shows the customer location and geographic segmentation tracking feature of the Focusen application.

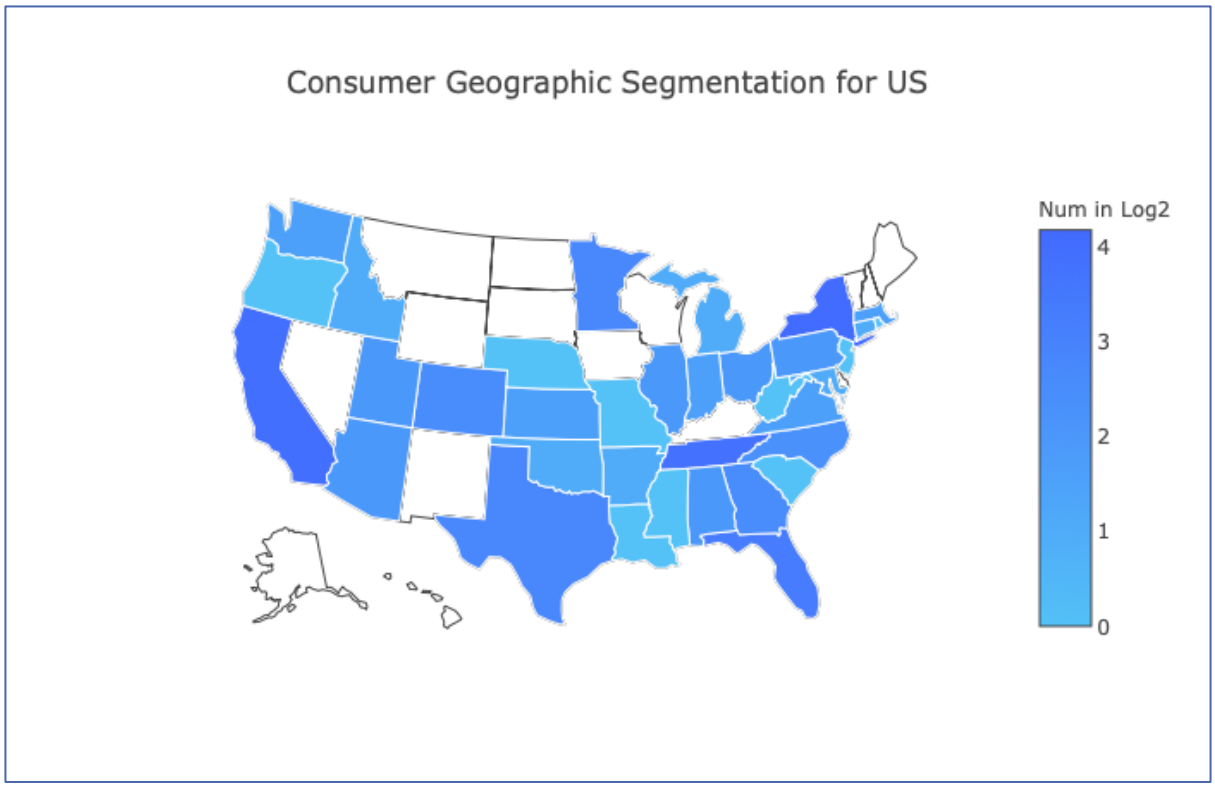


Figure 18 Consumer geographic segmentation on Focusen dashboard (Self-captured, 2020).

The image above shows the real-time consumer location tracking and geographic segmentation feature of the Focusen application. Users can use this heat-map to track the location of twitter users to get a better understand of where their consumers are located on states their products, services are brands are performing in each state by tracking their users on social media. The different shades of blue represent the frequency of tweets from each state in the United States where darker tone means more activity while white means no activity. The example shows the brand “Tesla” being tracked early morning in US.

9.1.5 Sentiment Analysis Pie-Chart

The screenshot below shows the Sentiment Analysis pie-chart summary feature on the Focusen web dashboard.

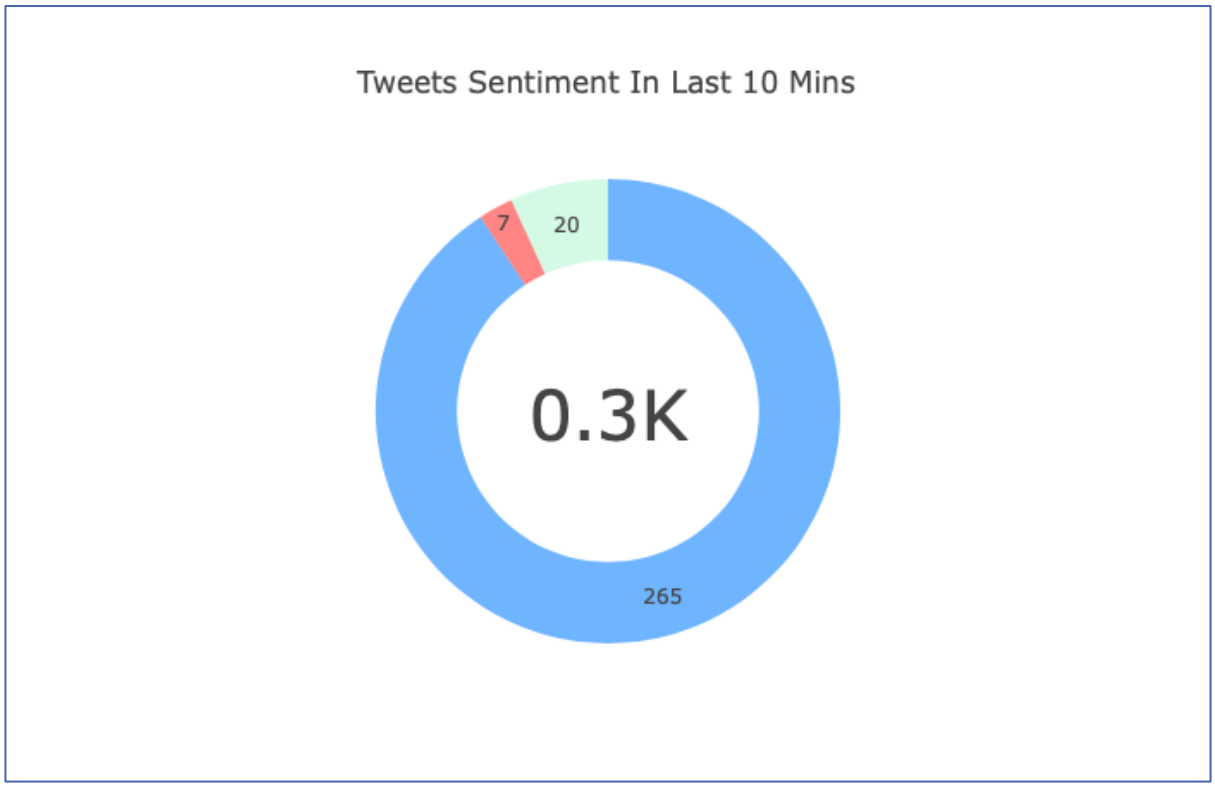


Figure 19 Focusen sentiment analysis summary pie-chart (Self-captured, 2020).

The image above shows the sentiment analysis summary feature of the Focusen application. User can use this feature to get a quick summary on the no. of tweets posted on a given time-frame, the example shows the no. of tweets posted in last 10 mins while tracking the brand "Tesla" on twitter when the screenshot was taken. Users can hover above the pie-chart to see the percentage of positive, negative and neutral tweets posted in the specified time frame. The numbers of tweets posted classified positive, negative and neutral can also been seem using the dashboard. This lets users quickly get a summary of consumer sentiment towards their brand in a specified time-frame in real-time to help them make more informed data-driven decisions quickly. The chart can also be downloaded easily for report purposes similar to all the other charts and graphs on the dashboard.

9.1.6 Social Media Tracking Details

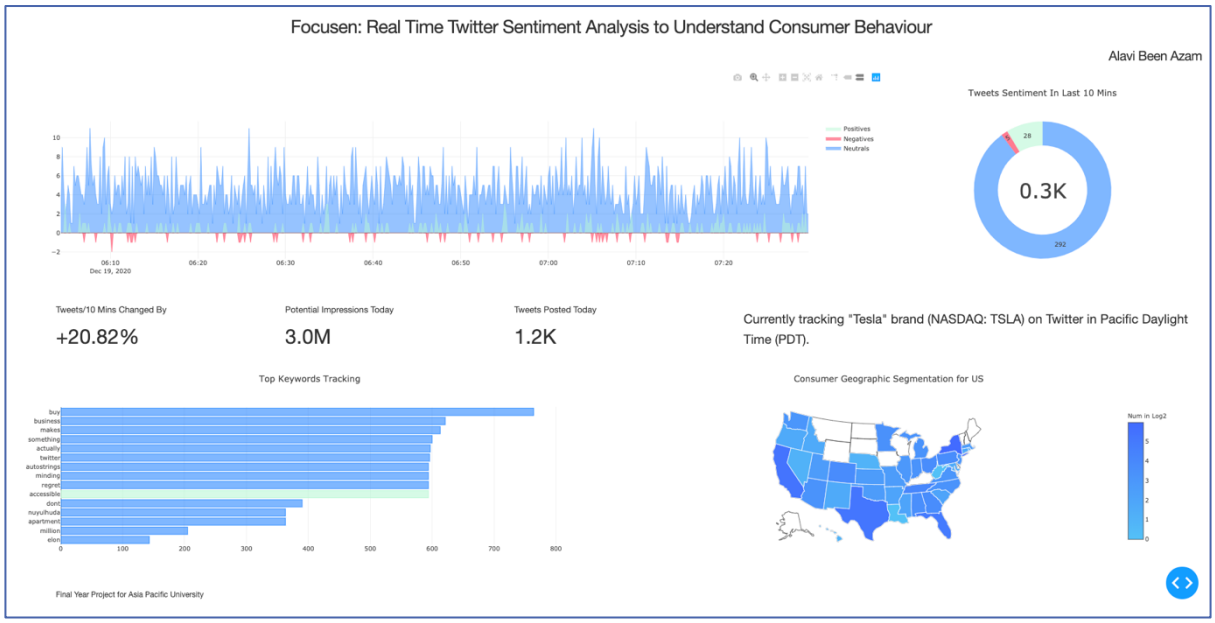
The screenshot below shows the component of the dashboard which provides a summary of the tracking details for social media platforms on the Focusen dashboard.

Figure 20 Social media tracking summary on Focusen Dashboard. (Self-captured, 2020).

The image above displays the tracking summary feature of the Focusen application dashboard. Users can use this feature to get real-time insights on the details such as the total amount of tweets posted about their brands or topic, they have been tracking in 1 day, the total number of impressions for the tweets posted about the topic and also the percentage change display the number of tweets posted in a given timeframe. In this example, the brand “Tesla” is being tracked early morning on PDT time and shows the 900 tweets have been posted about the brand within an hour which has a total impression reach of 2.9M (million) and the frequency of the tweets have increased at a percentage of 5.46% within the last minutes. This gives users quick and real-time insights about the consumers they are tracking on social media and also given an idea on how activity their brands or topics are on social media on the day specified. This can be potential useful on days of new product launches or PR pitfalls where companies can quickly and easily track the activity of consumer regarding their brands, products and services being tracked using social media data.

9.1.7 Focusen Full Dashboard

The screenshot below shows the full dashboard for the Focusen: Real Time Sentiment Analysis to Understand Consumer Behaviour system.



The image above shows the full web-based dashboard of the Focusen application with all the features discussed in the sections above. The application provides target users (researchers, marketers and business leaders a scalable, fast and accurate way to track consumer behaviour in real-time on social media. For this demonstration, the developer is using Twitter data using the Twitter API to portray the functionality of the application by tracking the company “Tesla” the leading electric car manufacturing company in the word. The dashboard lets users track a topic or keyword on twitter in real-time and get a detailed sentiment analysis report using the dashboard which can help them better understand their consumers sentiment and get actionable insights in real-time to make better data driven decisions to increase consumer satisfaction and profitability.

9.2 Sample Implementation Code

In this section the developer will discuss the code used in the core components of the system and justify and explain his solution and approach towards developing the Focusen application. The implementation had two phases. First was the data science and machine learning phase, where the developer used a Jupyter notebook for development and testing of the overall skeleton of the project with all the core features consisting of the data ingestion pipeline, text data pre-processing, model development and sentiment classification and visualization following the CRISP-DM methodology and for the web-development phases the developer used the RAD methodology as it has been explained in the details in the methodology section of this paper.

9.2.1 Libraries used for Data Ingestion, Processing & Model

```
import credentials # Import api/access_token keys from credentials.py
import settings # Import related setting constants from settings.py
import os
import psycopg2
import re
import tweepy
import pandas as pd
from textblob import TextBlob
```

Figure 21 Libraries used for scrapping, pre-processing and sentiment classification (Self-captured, 2020).

The above code snippet shows the libraries that has been used to develop the data ingestion pipeline, data pre-processing and sentiment analysis classification model. Psycopg2 library is being used to connect the application to a PostgreSQL database as the Focusen system uses PostgreSQL RDMS in deployment. The Python Regex (regular expression) library is being used to normalize the tweets and remove unwanted characters. Tweepy is being used to accessing and interacting with the python API. Pandas is popular data-processing libarary for python used data analysis and manipulation while the TextBlob library is being used for text data processing and sentiment analysis.

9.2.2 Extract Attributes from Tweets & Sentiment Analysis

The code snippet below shows the code used to extract relevant attributes from the Twitter API and how the data is being processed before being stored in the SQL table.

```
# Extract data from Twitter API

    def on_status(self, status):

        if status.retweeted:
            # Avoid retweeted info, and only original tweets will be received
            return True
        # Extract attributes from each tweet
        id_str = status.id_str
        created_at = status.created_at
        text = deEmojify(status.text)    # Pre-processing the text
        sentiment = TextBlob(text).sentiment
        polarity = sentiment.polarity
        subjectivity = sentiment.subjectivity

        user_created_at = status.user.created_at
        user_location = deEmojify(status.user.location)
        user_description = deEmojify(status.user.description)
        user_followers_count =status.user.followers_count
        longitude = None
        latitude = None
        if status.coordinates:
            longitude = status.coordinates['coordinates'][0]
            latitude = status.coordinates['coordinates'][1]

        retweet_count = status.retweet_count
        favorite_count = status.favorite_count
```

Figure 22 The code used to extract Attributes from Twitter API, data pre-processing & sentiment classification (Self-captured, 2020).

The snippet shows the different attributes being extracted from the Twitter API by defining local variable for each of the attributes that are being streamed using Tweepy and the Twitter API before they are stored in the local database. The deEmojify function from the Python RE library is being used to remove any emojis from tweet`s text body. The TextBlolb pre-trained sentiment analysis model is being used to classify the sentiment of the tweet by calculating the polarity and subjectivity of the tweet text. Positive tweets have a polarity of +1, negative -1 and neutral tweets are scored 0. Under the hood the Textblob sentiment analysis model API uses a Naïve Bayes classifier to tag and classify the tweets according to it polarity, by default the model calculates the polarity and subjectivity in each sentence the determine the tweet of the text provided. All the other required attributes are tagged with local variable before being stored into the database.

9.2.3 PostgreSQL Database Connection

```
conn = psycopg2.connect(dbname="twitterdb", user="postgres", password="postgres", host="localhost", port="5432")
cur = conn.cursor()
```

Figure 23 Database connection code (Self-captured, 2020).

The code snippet above shows the code being used to connect the application to a local hosted PostgreSQL database instance where the tweet attributes are stored after being processed.

```
# Store all data in PostgreSQL
cur = conn.cursor()
sql = "INSERT INTO {} (id_str, created_at, text, polarity, subjectivity, user_created_at, user_location, user_description, user_followers_count, longitude, latitude, retweet_count, favorite_count)
val = (id_str, created_at, text, polarity, subjectivity, user_created_at, user_location, \
    user_description, user_followers_count, longitude, latitude, retweet_count, favorite_count)
cur.execute(sql, val)
conn.commit()
```

Figure 24 SQL query and function to store data in database after processing (Self-captured, 2020).

The above snippet shows the function and SQL query being used to insert to data into the proper table after the data has been pre-processed and classified using the sentiment analysis machine learning (ML) model.

9.2.4 Twitter API Authentication & Data Streaming

```
auth  = tweepy.OAuthHandler(credentials.API_KEY, credentials.API_SECRET_KEY)
auth.set_access_token(credentials.ACCESS_TOEKN, credentials.ACCESS_TOKEN_SECRET)
api = tweepy.API(auth)

myStreamListener = MyStreamListener()
myStream = tweepy.Stream(auth = api.auth, listener = myStreamListener)
myStream.filter(languages=["en"], track = settings.TRACK_WORDS)

# Press STOP button to finish the process.
conn.close()


print('Done') #check if scrapping is running.
```

Figure 25 Twitter API authentication and data streaming code (Self-captured, 2020).

The code in the above snippet shows functions being used to authenticate access to the. Twitter API using Tweepy. The authentication details are stored in a different file called credentials.py which was imported at the start of the code. This has been done for better security and code modularity and keep the credentials safe and separate during deployment. The Tweepy data stream only fetches tweets containing the keyword which is defined and stored in another separate python file called settings.py.

9.2.5 Define Tracking Details

```
TRACK_WORDS = ['Tesla']
TABLE_NAME = "Tesla"
TABLE_ATTRIBUTES = "id_str VARCHAR(255), created_at TIMESTAMPTZ, text VARCHAR(255), \
            polarity INT, subjectivity INT, user_created_at VARCHAR(255), user_location VARCHAR(255), \
            user_description VARCHAR(255), user_followers_count INT, longitude float8, latitude float8, \
            retweet_count INT, favorite_count INT"
```

Figure 26 Defining topic Tracking Details & Database table settings (Self-captured, 2020).

The above code snippet shows the code in the settings.py file which is being used to define what topic or keyword will be tracked on Twitter in real time using the API. Users can

dynamically define the topic (brands, products, services etc.) they want to track on Twitter and the table will be created for the specific topic in the database to ensure that all the topic and details are stored separately for further analysis if required.

9.2.6 Libraries used for Dashboard & Deployment

The code snippet below shows all the python libraries that have been used for the devolvement on the web-based dashboard and application deployment.

```
import dash
import dash_core_components as dcc
import dash_html_components as html
from dash.dependencies import Input, Output
import pandas as pd
import plotly.graph_objs as go
import settings
import itertools
import math
import base64
from flask import Flask
import os
import psycopg2
import datetime

import re
import nltk
#nltk.download('punkt')
#nltk.download('stopwords')
from nltk.probability import FreqDist
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from textblob import TextBlob
import pytz
```

Figure 27 Libraries used for Focusen Dashboard development and deployment (Self-captured, 2020).

The frontend web dashboard has been developed using python Dash which uses the Plotly graphical library for data visualizations and Flask for server-side deployments at its core. The NLTK library is being used for text tokenization and the top keyword tracking function. TextBlob is used for text analysis and sentiment analysis as mentioned in the previous section. Detailed information and justification about all the libraries in the development of the proposed system being used has been provided in the technical research section of this paper.

9.2.7 Time Series Scatter Plot

The code snippet below shows the code being used to develop the time-series scatter plot graph on the dashboard in real time.

```
# Create the graph
children = [
            html.Div([
                html.Div([
                    dcc.Graph(
                        id='crossfilter-indicator-scatter',
                        figure={
                            'data': [
                                go.Scatter(
                                    x=time_series,
                                    y=result["Num of '{}' mentions".format(settings.TRACK_WORDS[0])][result['polarity']==0].reset_index(drop=True),
                                    name="Neutrals",
                                    opacity=0.8,
                                    mode='lines',
                                    line=dict(width=0.5, color='rgb(15, 131, 255)'),
                                    stackgroup='one'
                                ),
                                go.Scatter(
                                    x=time_series,
                                    y=result["Num of '{}' mentions".format(settings.TRACK_WORDS[0])][result['polarity']==-1].reset_index(drop=True).apply(lambda x: -x),
                                    name="Negatives",
                                    opacity=0.8,
                                    mode='lines',
                                    line=dict(width=0.5, color='rgb(255, 50, 50)'),
                                    stackgroup='two'
                                ),
                                go.Scatter(
                                    x=time_series,
                                    y=result["Num of '{}' mentions".format(settings.TRACK_WORDS[0])][result['polarity']==1].reset_index(drop=True),
                                    name="Positives",
                                    opacity=0.8,
                                    mode='lines',
                                    line=dict(width=0.5, color='rgb(184, 247, 212)'),
                                    stackgroup='three'
                                )
                            ]
                        }
                    )
                ], style={'width': '73%', 'display': 'inline-block', 'padding': '0 0 0 20'}),
```

Figure 28 Time-series sentiment plot code (Self-captured, 2020).

The track word function from the setting.py file is being used to track the defined keywords and collect data accordingly from the database table using their polarity score. Different colours have been used to mark the different sentiment classification as shown in the sample screenshot in the above section. Unstack-stack function in pandas is being used to stack the tweets in 10 sec intervals in the graph and show the plot for the past 30 minutes of Twitter scrapping data on the topic specified along with its sentiment classification.

```
# Clean and transform data to enable time series
result = df.groupby([pd.Grouper(key='created_at', freq='10s'), 'polarity']).count().unstack(fill_value=0).stack().reset_index()
result = result.rename(columns={"id_str": "Num of '{}' mentions".format(settings.TRACK_WORDS[0]), "created_at":"Time"})
time_series = result["Time"][result['polarity']==0].reset_index(drop=True)

min10 = (PDT_now - datetime.timedelta(minutes=10))
min20 = (PDT_now - datetime.timedelta(minutes=20))
#print(min10)

neu_num = result[result['Time']>min10]["Num of '{}' mentions".format(settings.TRACK_WORDS[0])][result['polarity']==0].sum()
neg_num = result[result['Time']>min10]["Num of '{}' mentions".format(settings.TRACK_WORDS[0])][result['polarity']==-1].sum()
pos_num = result[result['Time']>min10]["Num of '{}' mentions".format(settings.TRACK_WORDS[0])][result['polarity']==1].sum()
#print (result[result['Time']>min10])
```

Figure 29 Code used to enable time-series analysis (Self-captured, 2020).

The code above shows the code being used to enable time-series plotting on the Focusen Dashboard.

9.2.8 Word Tokenization & Keyword Frequency Tracking

The snippet below shows the code being used to text tokenization using the python NLTK library to enable top keyword in the dashboard using a histogram which shows the frequency of the top keywords being tracked along with its polarity for sentiment classification of the word as show on the graph screenshot in the section above.

```
tokenized_word = word_tokenize(content)
stop_words=set(stopwords.words("english"))
filtered_sent=[]
for w in tokenized_word:
    if (w not in stop_words) and (len(w) >= 3):
        filtered_sent.append(w)
fdist = FreqDist(filtered_sent)
fd = pd.DataFrame(fdist.most_common(16), columns = ["Word","Frequency"]).drop([0]).reindex()
fd['Polarity'] = fd['Word'].apply(lambda x: TextBlob(x).sentiment.polarity)
fd['Marker_Color'] = fd['Polarity'].apply(lambda x: 'rgba(255, 50, 50, 0.6)' if x < -0.1 else \
    ('rgba(184, 247, 212, 0.6)' if x > 0.1 else 'rgba(15, 131, 255, 0.6)'))
fd['Line_Color'] = fd['Polarity'].apply(lambda x: 'rgba(255, 50, 50, 1)' if x < -0.1 else \
    ('rgba(184, 247, 212, 1)' if x > 0.1 else 'rgba(15, 131, 255, 1)'))
```

Figure 30 Word tokenization and top keyword tracking function (Self-captured, 2020).

The snipper below shows the code used for developing the top keyword tracking histogram on the Focusen dashboard.

9.2.9 Location Tracking & Geographic Segmentation Map

The code snippet above shows the code used to develop the geographic distribution map on the Focusen dashboard to track the location and summarize the location of the Twitter users. The map will display the number of tweets created from each state in the United States for users who have location enabled on Twitter or are using geo-tagging. The states in US have been defined using constants in the state's variable. The data then is being cleaned and transformed to enable geo-distribution plotting using the Plotly library.

```
# Clean and transform data to enable geo-distribution
is_in_US=[]
geo = df[['user_location']]
df = df.fillna(" ")
for x in df['user_location']:
    check = False
    for s in STATES:
        if s in x:
            is_in_US.append(STATE_DICT[s] if s in STATE_DICT else s)
            check = True
            break
    if not check:
        is_in_US.append(None)

geo_dist = pd.DataFrame(is_in_US, columns=['State']).dropna().reset_index()
geo_dist = geo_dist.groupby('State').count().rename(columns={"index": "Number"}) \
    .sort_values(by=['Number'], ascending=False).reset_index()
geo_dist["Log Num"] = geo_dist["Number"].apply(lambda x: math.log(x, 2))


geo_dist['Full State Name'] = geo_dist['State'].apply(lambda x: INV_STATE_DICT[x])
geo_dist['text'] = geo_dist['Full State Name'] + '<br>' + 'Num: ' + geo_dist['Number'].astype(str)
```

Figure 31 Location tracking and geographic segmentation functions (Self-captured, 2020).

The number of tweets posted in each state are counted and logarithmic values are used to normalize the distribution on the map and for better visualization output. The code snippet below shows the code used for the map visualization.

```
    html.Div([
        dcc.Graph(
            id='y-time-series',
            figure = {
                'data':[
                    go.Choropleth(
                        locations=geo_dist['State'], # Spatial coordinates
                        z = geo_dist['Log Num'].astype(float), # Data to be color-coded
                        locationmode = 'USA-states', # set of locations match entries in `locations`
                        #colorscale = "Blues",
                        text=geo_dist['text'], # hover text
                        geo = 'geo',
                        colorbar_title = "Num in Log2",
                        marker_line_color='white',
                        colorscale = ["#54c1f7","#426bff"],
                        #autocolorscale=False,
                        #reversescale=True,
                    )
                ],
                'layout': {
                    'title': "Consumer Geographic Segmentation for US",
                    'geo':{'scope':'usa'}
                }
            }
        )
    ], style={'display': 'inline-block', 'width': '49%'})
]
```

Figure 32 Code used for geographic distribution plotting (Self-captured, 2020).

9.2.10 Sentiment Percentage & Pie-chart Plotting

The snippet below shows the code being used to display the sentiment analysis percentage pie-chart on the Focusen dashboard.

```
# Percentage Number of Tweets changed in Last 10 mins

count_now = df[df['created_at'] > min10]['id_str'].count()
count_before = df[ (min20 < df['created_at']) & (df['created_at'] < min10)]['id_str'].count()
percent = (count_now-count_before)/count_before*100
```

Figure 33 Sentiment percentage calculation function (Self-captured, 2020).

The code below shows the code being used to display the pie-chart on the dashboard according to the sentiment classifications and the number of tweets posted in the last 10 mins.

```
    html.Div([
        dcc.Graph(
            id='pie-chart',
            figure={
                'data': [
                    go.Pie(
                        labels=['Positives', 'Negatives', 'Neutrals'],
                        values=[pos_num, neg_num, neu_num],
                        name="View Metrics",
                        marker_colors=['rgba(184, 247, 212, 0.6)','rgba(255, 50, 50, 0.6)','rgba(15, 131, 255, 0.6)'],
                        textinfo='value',
                        hole=.65)
                ],
                'layout':{
                    'showlegend':False,
                    'title':'Tweets Sentiment In Last 10 Mins',
                    'annotations':[
                        dict(
                            text='{0:.1f}K'.format((pos_num+neg_num+neu_num)/1000),
                            font=dict(
                                size=40
                            ),
                            showarrow=False
                        )
                    ]
                }

            }
        )
    ], style={'width': '27%', 'display': 'inline-block'})
]),
```

Figure 34 Pie-chart plotting code (Self-captured, 2020).

10.0 System Validation

10.1 Unit Testing Results – Focusen

This section will discuss the final unit-testing results for the Focusen application before the developer moves on the integration test to unsure all the individual modules and components of the system are functioning as intended.

Case ID	Test Case	Test Function	Sample Data	Expected Result	Actual Result	Remarks
1	Receiving data using Twitter API while the application is running.	Data Ingestion Pipeline	Twitter API	Successful data streaming incoming.	As expected.	Passed Testing.
2	Automatic SQL table creation according to keyword being tracked.	Data Ingestion Pipeline	Twitter API	Table has been created in the SQL database according to the topic being tracked.	As expected.	Passed Testing.
3	Data being entered properly in the current table.	Data Ingestion Pipeline	Twitter API	Data is being written in the correct table according to the keyword/topic that is being tracked.	As expected.	Passed Testing.

4	Data stream is in real-time.	Data Ingestion Pipeline	Twitter API	Data in the table is correct and matches current UCT time.	As expected.	Passed Testing.
5	SQL CRUD queries are functional.	Data Pipeline	Twitter API	All the SQL queries are executing correctly.	As expected.	Passed Testing.
6	Tweet attributes extraction from Twitter API.	Data Ingestion Pipeline	Twitter API	All the attributes are correctly being extracted and stored in the database.	As expected.	Passed Testing.
7	Database connection established.	Data Pipeline	Twitter API	Database connection has been established successfully with the local SQL database.	As expected.	Passed Testing.
8	Text Tokenization is functional.	Data Pre-processing	Twitter API	Tweets are being tokenized correctly.	As expected.	Passed Testing.
9	Tweet stemming is functional.	Data Pre-processing	Twitter API	Text stemming is accurate using python regular expression (RE).	As expected.	Passed Testing.

10	Removal of emoji and unwanted characters from tweets.	Data Pre-processing	Twitter API	Emojis and unwanted characters such as punctuation are being removed properly before being stored in the database.	As expected.	Passed Testing.
11	ML model is functional	Machine Learning	Twitter API	ML model is classify the tweets and proving a result.	As expected.	Passed Testing.
12	Accuracy of the ML model.	Machine Learning	Twitter API	The accuracy of the output from the ML model is standard and acceptable.	As expected.	Passed Testing.
13	Check sentiment classification output.	Machine Learning	Twitter API	Tweets are being classified properly according using their polarity and subjectivity.	As expected.	Passed Testing.

14	UCT to PDT time conversion is correct.	Data Pre-processing	Twitter API	The creation time for tweets are successfully being converted from UCT time to PDT time.	As expected.	Passed Testing.
15	Check scatter plot is functional and time series distribution is accurate.	Dashboard	Twitter API	The scatter plot is functional and display classification in real-time.	As expected.	Passed Testing.
16	Sentiment Analysis Pie-chart is functional.	Dashboard	Twitter API	Sentiment Analysis summary pie-chart is current according to the time frame defined.	As expected.	Passed Testing.
17	Twitter summary data is accurate (no. of tweets, impressions, change in 10 mins)	Dashboard	Twitter API	Twitter summary data is accurate according to the input time-frame defined in the system.	As expected.	Passed Testing.

18	Topic Tracking is accurate.	Dashboard	Twitter API	The correct topic is tracked by the system as defined by the user.	As expected.	Passed Testing.
19	US geographical segmentation map is functional.	Dashboard	Twitter API	The USA geo-segmentation frequency heat-map is updating properly.	As expected.	Passed Testing.
20	Top keyword tracking is functional.	Dashboard	Twitter API	Top keywords are being tracked currently according to the topic defined.	As expected.	Passed Testing.

10.2 Integration Testing Results – Focusen

This section will discuss the final integration test results for the Focusen system. Once the system has passed all the test cases for the unit testing phase, the developer moves on the integration testing to see if all the modules are working properly after they have been integrated together in the final system, this is the last phase of technical testing. Once the system passes all the test cases of integration testing and satisfies the set standards and benchmarks, the developer will move on to user-testing with targeted users of the system.

Case ID	Test Case	Test Function	Sample Data	Expected Result	Actual Result	Remarks
1	Twitter API integration is successful.	Data Ingestion Pipeline	Twitter API	The Twitter API has been successfully integrated into the application.	As expected.	Passed Testing.
2	Database integration is successful.	Data Pipeline	Twitter API	The SQL database has been successfully integrated with the python application.	As expected.	Passed Testing.
3	Focusen dashboard is functional.	Dashboard	Twitter API	The front-end dashboard has been successfully connected to the backend application	As expected.	Passed Testing.

				using python Dash library.		
4	Model integration	Machine Learning	Twitter API	The ML model has been successfully deployed to the frontend dashboard of the application.	As expected.	Passed Testing.

10.3 User Acceptance Testing Results – Focusen

The section will discuss the result from the user-testing round which is the final testing phase for the Focusen application. The results were collected using a google form and the demonstration of the system was conducted using video conferencing.

Name:

The screenshot below shows the name of all the participants that took part in the user acceptance testing of Focusen application.

Figure 35 Focusen UAT participants (Self-captured, 2020).

Job Title (Current Designation)

The screenshot below shows the current designations of the participants of the that took part in the Focusen UAT exercise.

Figure 36 Job titles of Focusen UTA participants (Self-captured, 2020).

Date of Testing

The screenshot below shows the date of when the user acceptance testing for the Focusen application was conducted.

Figure 37 Date of UAT conducted for Focusen (Self-captured, 2020).

Test Duration

The screenshot below shows the duration it took to conducted the user acceptance testing for Focusen with each individual tester.

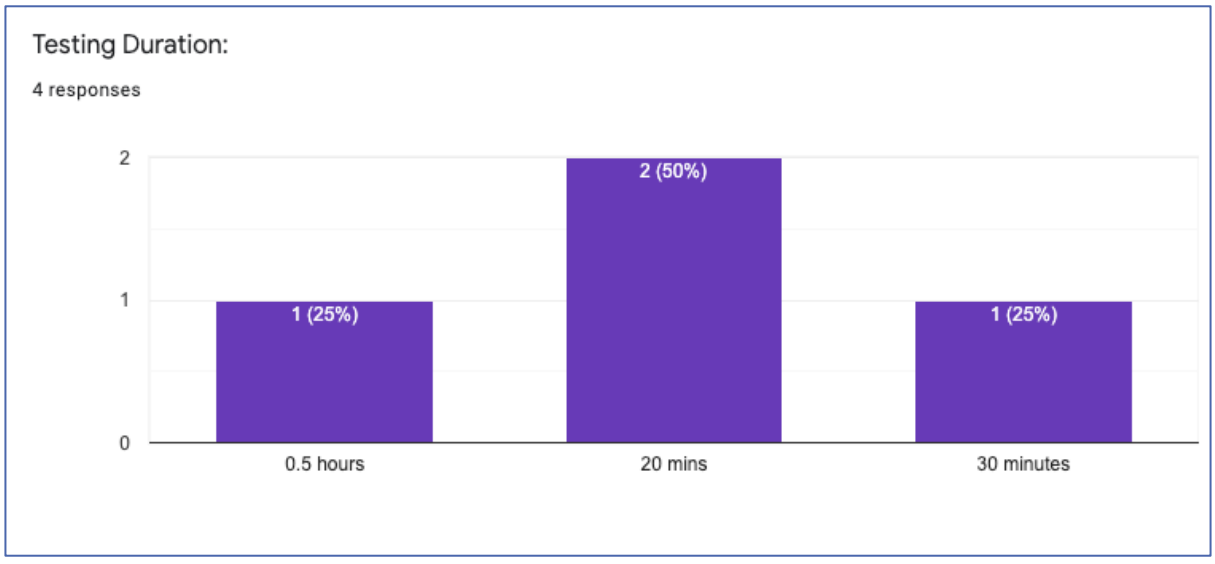


Figure 38 Test duration of Focusen UAT (Self-captured, 2020).

How would to like to rate the user interface of the system?

The screenshot below shows the summary results from tester on how would they rate the user interface (UI) of the Focusen application.

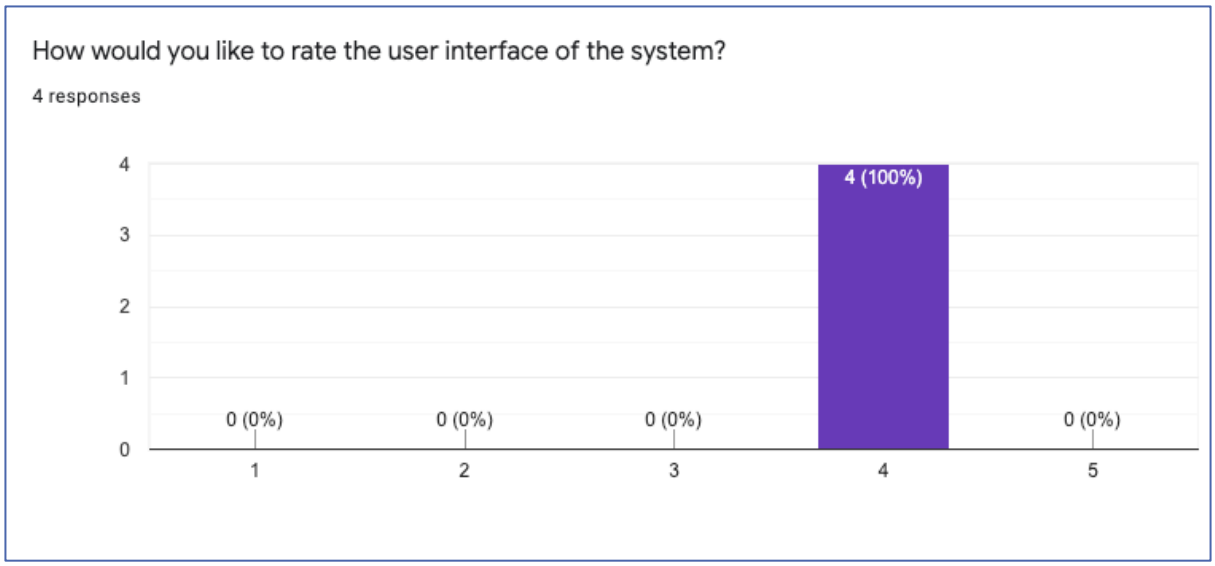


Figure 39 UAT UI design results summary (Self-captured, 2020).

Does the proposed system meet the business objectives defined by the developer?

The screenshot shows the opinion of testers on whether the proposed system meets the business objectives defined for the project.

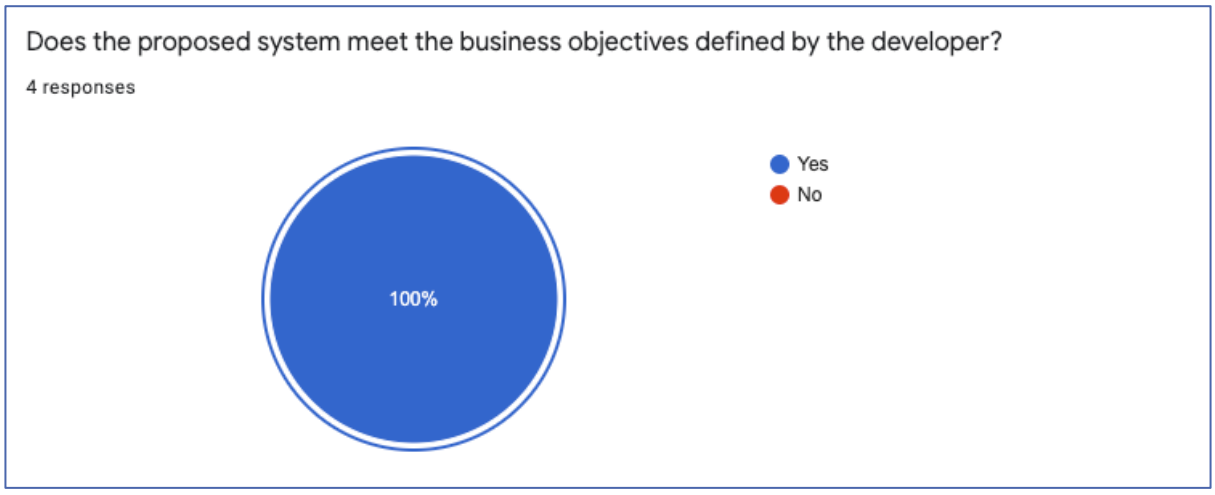


Figure 40 Business objective results of UAT. (Self-captured, 2020)

How well are the objectives defined being meet by the system?

The screenshot below shows the summary of opinion on how well the business objectives were meet by the proposed system Focusen.

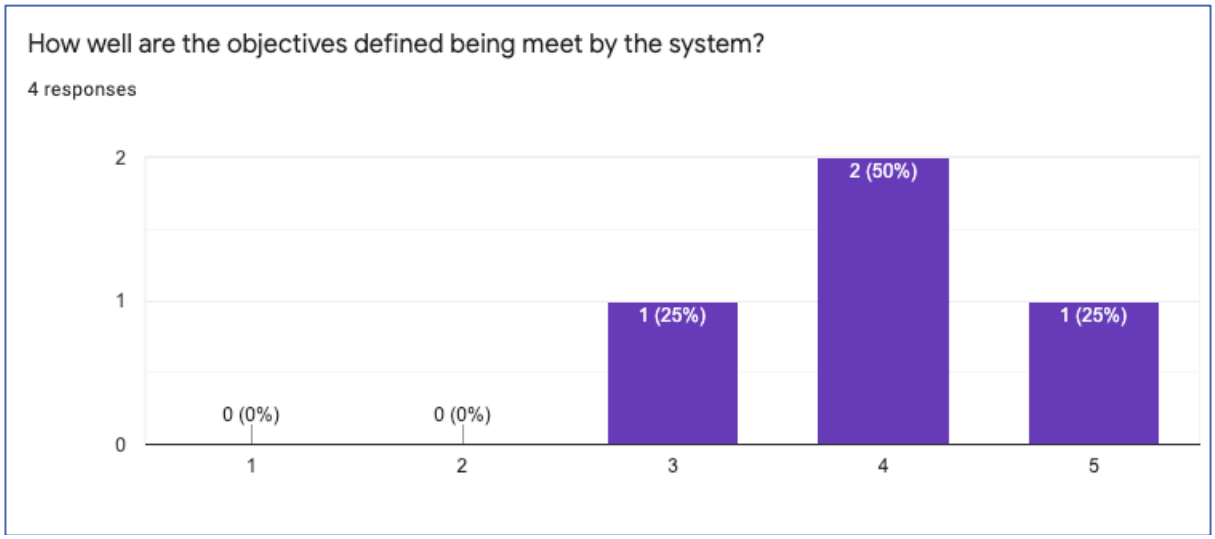


Figure 41 Summary of result from UAT on how well business objectives were met. (Self-captured, 2020).

How would you like to rate the overall performance of the system?

The screenshot below shows the testers opinion on the overall performance of the Focusen system.

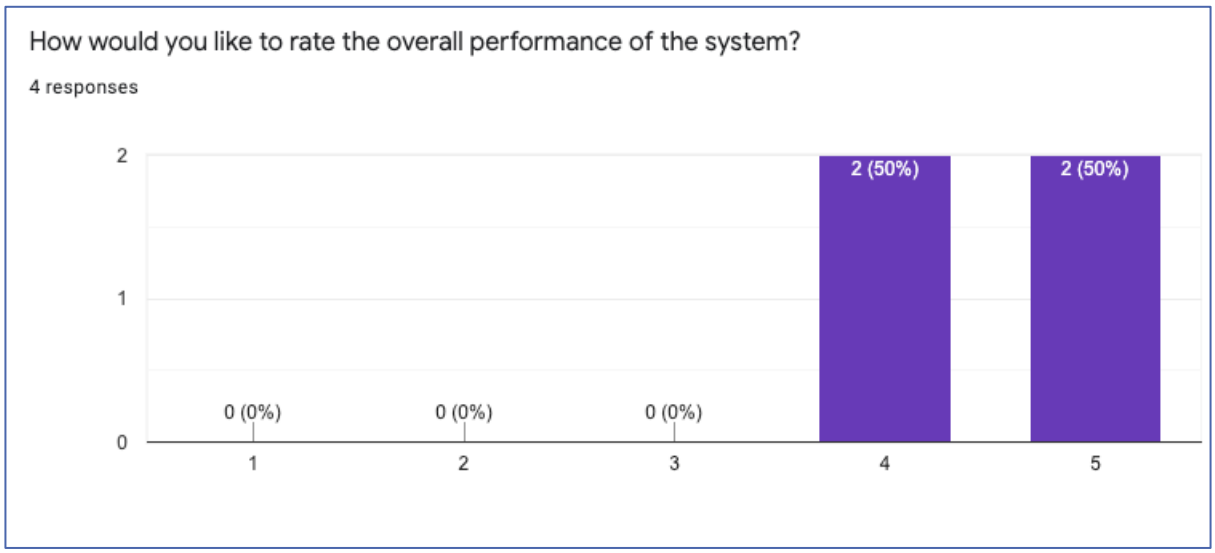


Figure 42 Testers opinion on overall system performance (Self-captured, 2020).

How would you like to rate the user friendliness of the system?

The screenshot below shows the summary of testers opinion on how user friendly the Focusen is as a whole.

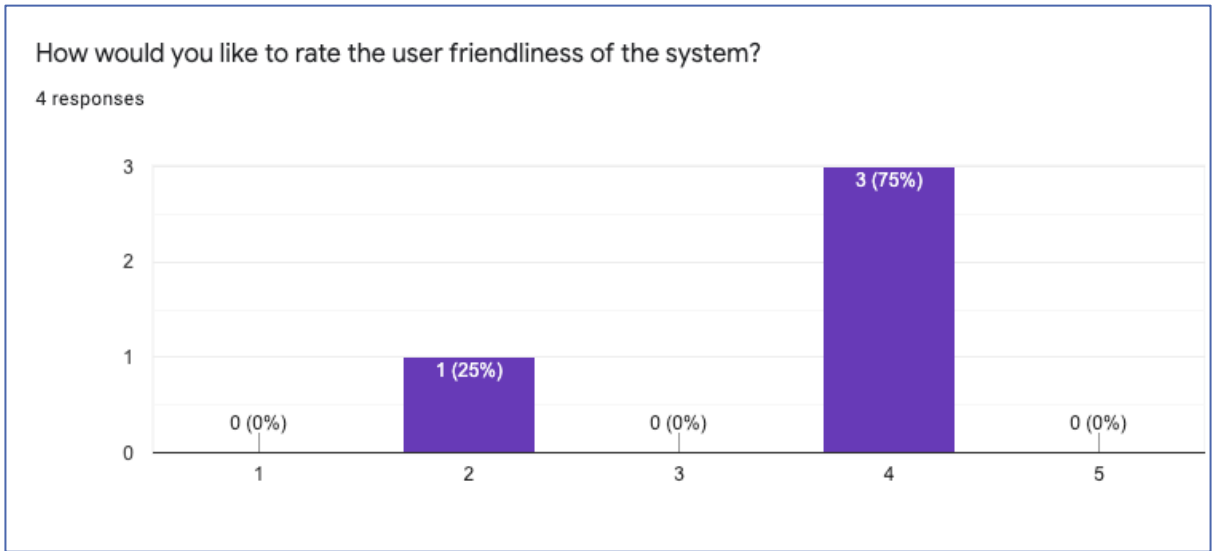


Figure 43 Testers response on user friendliness (Self-captured, 2020).

Tester Feedback:

The screenshot below shows the overall feedback from tester provided towards the Focusen application.

Figure 44 Overall feedback from testers (Self-captured, 2020).

Additional Comments:

This screenshot shows any additional comments that were provided by the testers towards to Focusen system demonstration and testing.

Figure 45 Additional comments provided by testers (Self-captured, 2020).

Note: Individual responses from the user acceptance testing (UTA) will be attached at the end of this document.

10.4 Summary

As the developer followed a test-driven approach, unit-testing, integration test and user acceptance testing was conducted from the Focusen system to ensure that the proposed solution meet all the requirement standards and was fully-functional. The system has successfully passed all the cases of unit and integration testing. A through demonstration and user acceptance testing (UTA) was conducted with the right users to ensure that the system meet all the functional and non-functional requirements and objectives stated for this research. The UTA results show that most of the testers were very satisfied with the overall performance, user-friendliness and user interface of the Focusen application. All the testers agreed that the solution proposed and system developed meet all the objectives defined by the research. Some tester did provide additional feedback that a input field should be added to the frontend of the application to let the users choose the tracking topic and the that the dashboard should be made mobile responsive. Some tester also suggested some UI/UX improvements which the developer has taken into consideration for the next version of the application. Overall, it can be concluded that the Focusen system has successfully passed the testing phase of the research.

11.0 Conclusion and Reflections

This section will wrap up this paper and provide a detailed critical evaluation for this project. The section will also provide a general conclusion and reflections from the researcher and his future plans for this project.

11.1 Critical Evaluation

In the final section of this paper, the developer would summary his thoughts and provide a critical evaluation to reflect on the overall outcome of the Focusen: Real Time Sentiment Analysis to Understand Consumer Behaviour project. The researcher will compile his thoughts in terms of the objectives and benefits of this project and how they have been met, evaluation on the methodology chosen, reflect on the implementation and its challenges and finally evaluate to overall effective of this solution in solving the business problems identified during the undertaking of this project.

11.1.1 Objectives & Benefits of the Project

The aim of this project was to develop a system that can be used by market researching companies and organization to better understand the sentiment of their consumers and understand consumer behaviour to make better data-driven decisions as discussed in details in the background and problem context section of this paper. To solve this problem, the core objectives were to develop an AI (Artificial Intelligence) based system that can automatically tag social media data, survey responses and online-focus group conversations to classify their sentiment with high accuracy in real-time. The system had to be had to be robust, consist and highly scalable to that it can handle the vast of data that is available on the internet where consumers are constantly provide their opinion about the product and services they are using. The developer worked with a leading market research company in Kuala Lumpur, Malaysia Vase Technologies who operates across ASEAN to validate the problem context and set the business goals, functional and non-functional requirements for this project as thoroughly discussed in the research and analysis section,

After fine-tunning the requirements according to the feedback collected and validating the idea the researcher planned to develop an AI application called Focusen which will use sentiment

analysis to help understand consumer behaviour. The benefits of this project are that the application provides a real-time dashboard which companies and individual market researchers can use to track consumer sentiment towards a chosen topic or keyword. The system can be used to track brands, products, services or another other topic on social media to identify trends and to get a better understanding of what consumers are saying about topic in real-time. As data is said to be the new goals and the importance of understanding consumer behaviour for companies have been highlighted throughout the project, Focusen provides organizations with an immense opportunity to monitor and collect feedback from their customers (consumers) in real-time to make more informed and data-driven decision. As highlighted in the problem context such research usually takes weeks to months be done manually without any AI solutions are often are not accurate enough due to human errors and biases. Also, it is almost impossible to manually process the vast amount of that that is available on the internet and generate actionable insights to make better decisions. Focusen use artificial intelligence and machine learning to do just that will great accuracy making it an ideal solution for market researchers and organisation looking to better understand their consumers.

11.1.2 Methodology Evaluation

The point of using a methodology for any research or development is to have a structured approach in solving the problem and is a critical factor which highly contributes to the success of any project. Having a structured approach and set methodology immensely helped the developer in completing the project within the set time-frame as the methodology provides a step-by-step approach in developing the proposed solution. For the project the developer decided to use a combination of two different methodologies, CRISP-DM which is a popular data science methodology and RAD which is another popular system development methodology. The step-by-step approach and structure of CRISP-DM from business objective discovery to model deployment helped the developer break the project down in smaller components and complete it in steps which highly contributed to the accuracy and successfully completions of project through meeting the business requirements. Similarly, having a structured methodology for the web-application development helped the developer readily prototype, receive feedback and iterate the design to unsure that the application was satisfy all the needs and business requirements of the target users. The methodologies helped the developer complete the tasks according to the checklist and prioritization and follow up regular with the project supervisor to show project and receive feedback on over steps contributing to

the successful completion of the project. Overall, the author is highly satisfied with the methodologies that were chosen for this project.

11.1.3 Implementation Evaluation and Challenges

The developer has opted to use the python program language to fully develop this project because of prior knowledge on the language and the flexibility and ease of usage compared to other languages as discussed in the technical research section. As python is a popular language for data science and machine learning applications, a lot of resources were available online for the developer to refer to and solve any challenges or bugs that occurred during the implementation stage of the project. As the project mostly uses open-source libraries and packages which are free and provide excellent documentation and examples which thoroughly helped the developer in completing the project from a technical perspective. The developer faced some challenges while developing the backend of the web-dashboard as he does not have much experience with backend development but the vast of online guides, articles and example code was available which provided great assistance in successfully implementing the proposed solutions according to the objectives and defined business requirements.

11.1.4 Target User Evaluation

During the demonstration and user-testing phase of the application, the target users and testers were satisfied with the solution developed taking into account the academic nature of the project and the current situation (Covid19 pandemic). The users were impressed with the simplistic UI and user-friendly UX of the application as the system does not require much effort to apart and the only input that is required by the user is the topic that they want to track. However, the users did have feedback and suggestions about the accuracy and limitations of the system to prepare it for real-world deployment and usage. Overall users were satisfied that the project has achieved the overall objective and goals defined for the project and is a step in the right direction for developing a system which can actually deploy in a real-world scenario with further research and technical expertise. For details about the limitations the system faces and further research plans will be discussed in the next sections

11.1.5 Evaluation Summary

This sections above discussed the critical evaluation of the project from the authors perspective and feedback collected from target users and project supervisor during the research, development and testing of the Focusen application. Further reflection and limitations of the project will be discussed in the conclusion section below.

11.2 Conclusion

This research paper is part of an academic project written and developed by the author himself with guidance from the project supervisor and industry mentors who provided immense support throughout the project. The first chapter of the research, the developer discussed the details the background of the project, the problem context, benefit, objectives and deliverables of the project followed by a detailed timeline of the entire project. Next the author moves on to conduct a thorough literature reviews and research similar systems can solutions that have been proposed by other researchers or are currently available in the market. After the literature review, the paper contains a through technical research conducted to identify the best technical resources that are available to successfully complete this project. Then the paper, discuss the methodologies chosen for the project and break-down the different stages of the proposed system development. Next the paper provides an overview about the research conducted to validate the proposed solution and its accompanying analysis. After the analysis, the author provides the details about the system architecture and project release plan of the project followed by a detailed guide on how the project was implemented with sample code and overview of all the core functionalities of the system. Finally, the paper discussed about the testing and validation conducted for the proposed system followed by an evaluation to wrap up the research conducted.

11.2.1 System Limitations

Even to the system has managed to fulfil all the aims and objectives of defined by research, there are several limitations and drawbacks to the proposed system. Firstly, the system demonstration is conducted using USA as the target market and tracking global companies which are very popular by mining data from social-media. Due to the strategic selection of the region and companies, abundance of data was available through tweets to track the sentiments of the consumers towards the brand, however for smaller companies it will become difficult to

track consumer through only relying on social media and they need to rely on surveys and focus groups to collect data. Even though the system is technically capable of using proprietary APIs from market research companies and other third-party organisations, such test cases have not been tested and will require significant changes to the data ingestion and processing pipeline making the real-world deployments and on-boarding harder as a SaaS (software as a service solution). Next, the demonstration uses only tracked tweets in English to understand consumer sentiments. As only a small fraction of the global population is native English speaker it will be hard to deploy the system in regions where the first language is not English and consumer speak a different language or a mixture of different languages, Localisation will be a huge challenges and limitation for the proposed system. Even though TextBlob is technically capable of handling a few popular languages such as French it is not reliable and the use-cases have not been tested. To deploy the proposed solution in countries where majority of the population a custom model is needed to be trained to make to system reliable and deployable. Also, as the system is developed to only focus on real-time data streaming scenarios and use-cases, the system does not have a way for users to upload their own data using a more common format such as CSV or JSON except fully custom API integrations making the system less versatile and ideal for only single API integration and monitoring at same time. Lastly, the system can only track one topic at a time, this is a limiting factor for brands who have multiple product or services and want to track them all using a single dashboard. The developer plans to overcome the limitations and develop the system further in the next versions of the application.

11.2.2 Research Limitation

Due to the strict timeframe of the research and other academic commitments of the research there were certain limitations in the research. More data could have been collected from analysis can validation from industry experts to provide a more solid validation and getting a better understanding of the problem. Also due to the Covid19 pandemic, several restrictions were imposed which made the research process and data gathering harder for the researcher.

11.2.3 Further Research Plans

The developer plans to further continue his work beyond the current scope of the project to improve the system further and fix the limitations that have been identified. The researcher

also plans to continue is work on the field of market research and machine learning to continue to collaborate with other Vase and other companies to better understand the problem and find better solutions that can be implemented in real-world use cases.

11.2.4 Personal Reflection

This academic project has helped the author further improve and develop his research and technical skills which will be immensely valuable in the industry. The structured nature of the project and strict deadlines have helped the author to become better at time management and strategic planning and research which are highly valuable and sort after skills. The author plans to continue working in this field and carry out further research in the field of Artificial Intelligence (AI) to help find solutions to the toughest problems in the world to make the world a better place for everyone.

List of References

2ndQuadrant (n.d.). PostgreSQL vs MySQL. *2ndQuadrant | PostgreSQL*. [Online]. Available from: https://www.2ndquadrant.com/en/postgresql/postgresql-vs-mysql/. [Accessed: 17 December 2020].

Ali, K., Dong, H., Bouguettaya, A., Erradi, A. & Hadjidj, R. (2017). Sentiment Analysis as a Service: A Social Media Based Sentiment Analysis Framework. In: *2017 IEEE International Conference on Web Services (ICWS)*. [Online]. June 2017, Honolulu, HI, USA: IEEE, pp. 660–667. Available from: http://ieeexplore.ieee.org/document/8029820/. [Accessed: 9 August 2020].

Almuqren, L. & Cristea, A.I. (2016). Framework for Sentiment Analysis of Arabic Text. In: *Proceedings of the 27th ACM Conference on Hypertext and Social Media*. HT '16. [Online]. 10 July 2016, New York, NY, USA: Association for Computing Machinery, pp. 315–317. Available from: http://doi.org/10.1145/2914586.2914610. [Accessed: 12 August 2020].

Amazon S3 (2020a). *Cloud Object Storage | Store & Retrieve Data Anywhere | Amazon Simple Storage Service (S3)*. [Online]. 2020. Amazon Web Services, Inc. Available from: https://aws.amazon.com/s3/. [Accessed: 9 August 2020].

Anderson, E.W., Fornell, C. & Lehmann, D.R. (1994). Customer Satisfaction, Market Share, and Profitability: Findings from Sweden. *Journal of Marketing*. 58 (3). p.pp. 53–66.

Anon (n.d.) a. *10 Essential Market Research Methods*. [Online]. Brandwatch. Available from: https://www.brandwatch.com/blog/market-research-methods/. [Accessed: 8 August 2020a].

Anon (n.d.) b. *2017-Scrum-Guide-US.pdf*. [Online]. Available from: https://www.scrumguides.org/docs/scrumguide/v2017/2017-Scrum-Guide-US.pdf#zoom=100. [Accessed: 10 August 2020b].

Anon (n.d.) c. *Colaboratory – Google*. [Online]. Available from: https://research.google.com/colaboratory/faq.html. [Accessed: 9 August 2020c].

Anon (2000). CRISP-DM. *Data Science Project Management*. [Online]. Available from: https://www.datascience-pm.com/crisp-dm-2/. [Accessed: 10 August 2020].

Anon (n.d.) d. *Customer Satisfaction: The Foundation of Business Success*. [Online]. Salesforce.com. Available from: https://www.salesforce.com/hub/service/importance-of-customer-satisfaction/. [Accessed: 8 August 2020d].

Anon (n.d.) e. *Enterprise AI Platform*. [Online]. RapidMiner. Available from: https://rapidminer.com/products/. [Accessed: 9 August 2020e].

Anon (2019) f. Key Benefits of Sentiment Analysis for Businesses. *CommSights*. [Online]. Available from: https://www.commsights.com/benefits-of-sentiment-analysis-for-businesses/. [Accessed: 7 August 2020].

Anon (n.d.) g. *Natural Language Toolkit — NLTK 3.5 documentation*. [Online]. Available from: https://www.nltk.org/. [Accessed: 9 August 2020f].

Anon (n.d.) h. *Pandas Basics - Learn Python - Free Interactive Python Tutorial*. [Online]. Available from: https://www.learnpython.org/. [Accessed: 9 August 2020g].

Anon (n.d.) I. *SciPy.org — SciPy.org*. [Online]. Available from: https://www.scipy.org/. [Accessed: 9 August 2020h].

Anon (n.d.) j. *Visual Studio Code - Code Editing. Redefined*. [Online]. Available from: https://code.visualstudio.com/. [Accessed: 9 August 2020i].

Anon (n.d.) k. *What Is an IDE?* [Online]. Codecademy. Available from: https://www.codecademy.com/articles/what-is-an-ide. [Accessed: 9 August 2020j].

Anon (n.d.) l. *What main methodology are you using for your analytics, data mining, or data science projects? Poll*. [Online]. Available from: https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html. [Accessed: 10 August 2020k].

Azevedo, A. & Santos, M.F. (2008). *KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW*. p.p. 6.

Barysevich, A. (2020). *10 Sentiment Analysis Tools to Measure Brand Health*. [Online]. April 2020. Social Media Today. Available from: https://www.socialmediatoday.com/news/10-sentiment-analysis-tools-to-measure-brand-health/575334/. [Accessed: 9 August 2020].

Beekeeper Studio (2020). *Open Source SQL Editor and Database Manager*. [Online]. 2020. Beekeeper Studio. Available from: https://www.beekeeperstudio.io/. [Accessed: 17 December 2020].

Bishop, S. (n.d.). *pytz: World timezone definitions, modern and historical*. [Online]. Available from: http://pythonhosted.org/pytz. [Accessed: 17 December 2020].

Cambria, E., Schuller, B., Liu, B., Wang, H. & Havasi, C. (2013). Knowledge-Based Approaches to Concept-Level Sentiment Analysis. *IEEE Intelligent Systems*. 28 (2). p.pp. 12–14.

Capozzi, C. (2017). *What Is the Difference Between Tangible & Intangible Benefits?* [Online]. September 2017. Bizfluent. Available from: https://bizfluent.com/info-8094073-difference-between-tangible-intangible-benefits.html. [Accessed: 7 August 2020].

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). *Step-by-step data mining guide*. p.p. 76.

Day, M.-Y. & Lin, Y.-D. (2017). Deep Learning for Sentiment Analysis on Google Play Consumer Review. In: *2017 IEEE International Conference on Information Reuse and Integration (IRI)*. [Online]. August 2017, San Diego, CA: IEEE, pp. 382–388. Available from: http://ieeexplore.ieee.org/document/8102961/. [Accessed: 7 August 2020].

Donegan, C. (2019). *State of the Connected Customer Report Outlines Changing Standards for Customer Engagement*. [Online]. June 2019. Salesforce.com. Available from: https://www.salesforce.com/company/news-press/stories/2019/06/061219-g/. [Accessed: 7 August 2020].

Gell, T. (2019). *Using Sentiment Analysis in Surveys | Market Research Tips*. [Online]. January 2019. Available from: https://www.driveresearch.com/market-research-company-blog/using-sentiment-analysis-in-surveys-market-research-tips/. [Accessed: 9 August 2020].

Gregorio, F.D. (n.d.). *psycopg2: psycopg2 - Python-PostgreSQL Database Adapter*. [Online]. Available from: https://psycopg.org/. [Accessed: 17 December 2020].

Gunter, B., Koteyko, N. & Atanasova, D. (2014). Sentiment Analysis: A Market-Relevant and Reliable Measure of Public Feeling? *International Journal of Market Research*. 56 (2). p.pp. 231–247.

Javalgi, R. (Raj) G., Martin, C.L. & Young, R.B. (2006). Marketing research, market orientation and customer relationship management: a framework and implications for service providers. *Journal of Services Marketing*. 20 (1). p.pp. 12–23.

Keras (2020b). *Keras: the Python deep learning API*. [Online]. 2020. Available from: https://keras.io/. [Accessed: 9 August 2020].

Khan, M.T., Durrani, M., Ali, A., Inayat, I., Khalid, S. & Khan, K.H. (2016). Sentiment analysis and the complex natural language. *Complex Adaptive Systems Modeling*. 4 (1). p.p. 2.

Kim, J.-H. (2019). Imperative challenge for luxury brands: Generation Y consumers' perceptions of luxury fashion brands' e-commerce sites. *International Journal of Retail & Distribution Management*. 47 (2). p.pp. 220–244.

Laksono, R.A., Sungkono, K.R., Sarno, R. & Wahyuni, C.S. (2019). Sentiment Analysis of Restaurant Customer Reviews on TripAdvisor using Naïve Bayes. In: *2019 12th International Conference on Information Communication Technology and System (ICTS)*. July 2019, pp. 49–54.

Marshall, G. (2005). The purpose, design and administration of a questionnaire for data collection. *Radiography*. 11 (2). p.pp. 131–136.

Mitchell, V. (1992). Understanding Consumers' Behaviour: Can Perceived Risk Theory Help? *Management Decision*. 30 (3). p.p. 00251749210013050.

Mohajan, H. (2018). *Qualitative Research Methodology in Social Sciences and Related Subjects*. [Online]. 10 December 2018. Available from: https://mpra.ub.uni-muenchen.de/85654/. [Accessed: 11 August 2020].

Mohammad, S. (2011). From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales. In: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. [Online]. June 2011, Portland, OR, USA: Association for Computational Linguistics, pp. 105–114. Available from: https://www.aclweb.org/anthology/W11-1514. [Accessed: 9 August 2020].

Mohammad, S. & Yang, T. (2011). Tracking Sentiment in Mail: How Genders Differ on Emotional Axes. In: *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*. [Online]. June 2011, Portland, Oregon: Association for Computational Linguistics, pp. 70–79. Available from: https://www.aclweb.org/anthology/W11-1709. [Accessed: 9 August 2020].

NumPy (2020c). *NumPy*. [Online]. 2020. Available from: https://numpy.org/. [Accessed: 9 August 2020].

Ozgur, C. & University, V. (2017). *MatLab vs. Python vs. R*. p.p. 18.

Pallets Projects (2020d). *Welcome to Flask — Flask Documentation (1.1.x)*. [Online]. 2020. Available from: https://flask.palletsprojects.com/en/1.1.x/. [Accessed: 9 August 2020].

Pandas (2020e). *pandas - Python Data Analysis Library*. [Online]. 2020. Available from: https://pandas.pydata.org/. [Accessed: 9 August 2020].

Pandey, D., Khan, W. & Pandey, V. (2012). *ROLE OF REQUIREMENT VALIDATION IN REQUIREMENT DEVELOPMENT*. In: 21 November 2012.

Pang, B. & Lee, L. (2008). *Opinion mining and sentiment analysis*. p.p. 94.

Plotly (2020a). *Dash Overview*. [Online]. 2020. Available from: /dash. [Accessed: 17 December 2020].

Plotly (2020b). *Plotly Python Graphing Library*. [Online]. 2020. Available from: https://plotly.com/python/. [Accessed: 16 December 2020].

PostgreSQL Global Development (2020). *PostgreSQL*. [Online]. 16 December 2020. PostgreSQL. Available from: https://www.postgresql.org/. [Accessed: 17 December 2020].

Project Jupyter (2020f). *Project Jupyter*. [Online]. 2020. Available from: https://www.jupyter.org. [Accessed: 9 August 2020].

Rowley, J. & Slack, F. (2001). Leveraging customer knowledge – profiling and personalisation in e-business. *International Journal of Retail & Distribution Management*. 29 (9). p.pp. 409–416.

Saltz, J.S. (2015). The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. In: *2015 IEEE International Conference on Big Data (Big Data)*. [Online]. October 2015, Santa Clara, CA, USA: IEEE, pp. 2066–2071. Available from: http://ieeexplore.ieee.org/document/7363988/. [Accessed: 10 August 2020].

Sarkar, S. (2018). Benefits of Sentiment Analysis for Businesses. *Analytics Insight*. [Online]. Available from: https://www.analyticsinsight.net/benefits-of-sentiment-analysis-for-businesses/. [Accessed: 7 August 2020].

Schouten, K. & Frasincar, F. (2016). Survey on Aspect-Level Sentiment Analysis. *IEEE Transactions on Knowledge and Data Engineering*. 28 (3). p.pp. 813–830.

Schwaber, K. (1997). SCRUM Development Process. In: J. Sutherland, C. Casanave, J. Miller, P. Patel, & G. Hollowell (eds.). *Business Object Design and Implementation*. [Online]. London: Springer London, pp. 117–134. Available from: http://link.springer.com/10.1007/978-1-4471-0947-1_11. [Accessed: 10 August 2020].

Shafique, U. & Qaiser, H. (2014). *A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)*. 12 (1). p.p. 6.

Sims, S. (2015). Sentiment Analysis 101. *KDnuggets*. [Online]. Available from: https://www.kdnuggets.com/sentiment-analysis-101.html/. [Accessed: 7 August 2020].

Singh, A. (2019). *What Is Rapid Application Development (RAD)?* [Online]. December 2019. Available from: https://blog.capterra.com/what-is-rapid-application-development/. [Accessed: 10 August 2020].

Statt, D.A. (2013). *Consumer Behaviour*. p.p. 26.

Team, D.J. (2020). *The Importance of Consumer Behavior in Marketing*. [Online]. January 2020. Available from: https://www.demandjump.com/blog/the-importance-of-consumer-behavior-in-marketing. [Accessed: 7 August 2020].

Team, M. (2011). *The Importance of Market Research Explained, or Why You should Research Markets*. [Online]. Available from: https://mymanagementguide.com/the-importance-of-market-research-explained-or-why-you-should-research-markets/. [Accessed: 8 August 2020].

TextBlob (2020). *TextBlob: Simplified Text Processing — TextBlob 0.16.0 documentation*. [Online]. 2020. Available from: https://textblob.readthedocs.io/en/dev/. [Accessed: 16 December 2020].

Troisi, O., Grimaldi, M., Loia, F. & Maione, G. (2018). Big data and sentiment analysis to highlight decision behaviours: a case study for student population. *Behaviour & Information Technology*. 37 (10–11). p.pp. 1111–1128.

Tweepy (2020). *Tweepy*. [Online]. 2020. Available from: https://www.tweepy.org/. [Accessed: 16 December 2020].

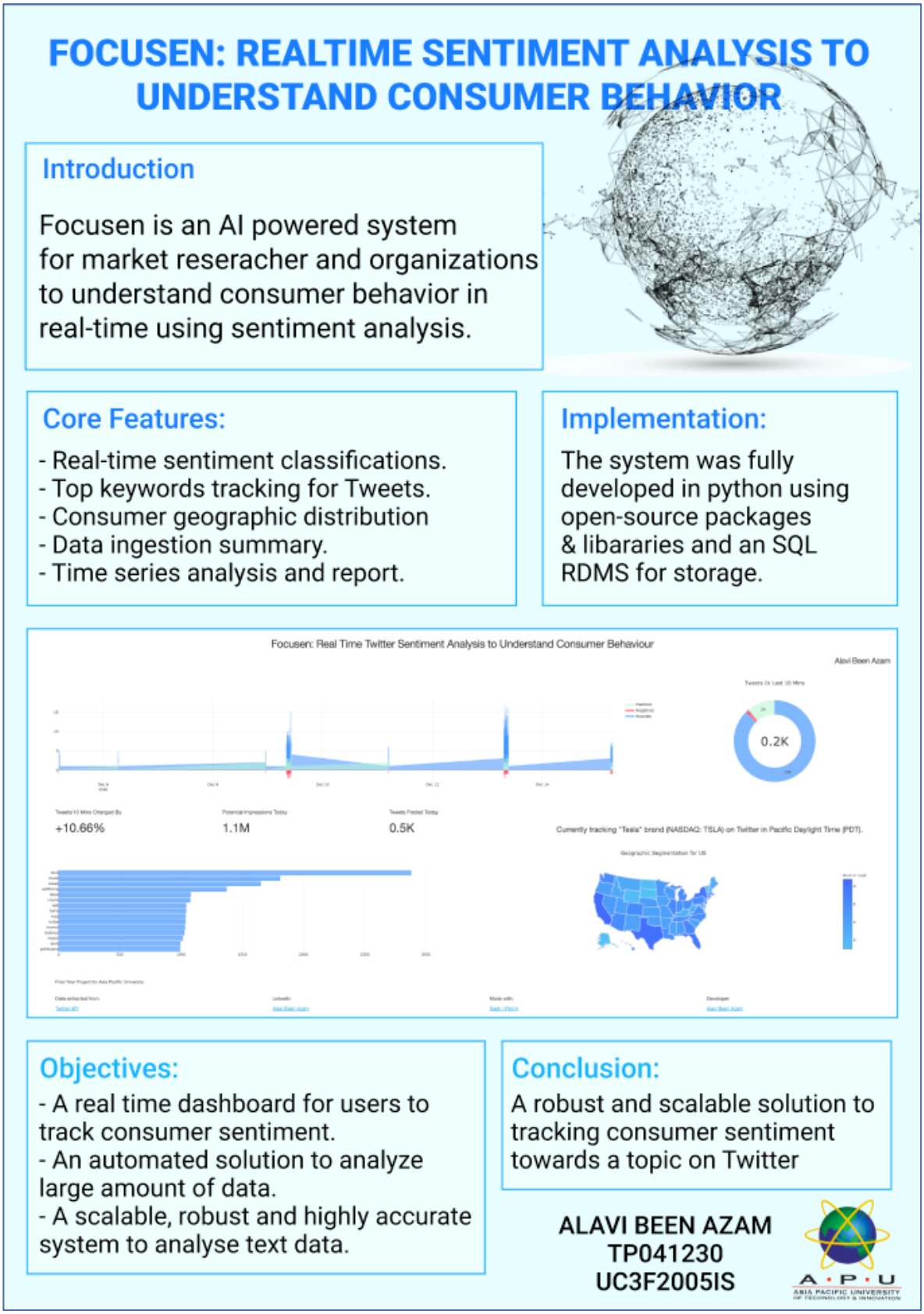
Vase (2020). *Learning Resources | Vase Actionable Intelligence*. [Online]. 2020. Learning Resources | Vase Actionable Intelligence. Available from: https://vase.ai/resources/. [Accessed: 15 December 2020].

Wirth, R. & Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*. p.p. 11.

Witell, L., Kristensson, P., Gustafsson, A. & Löfgren, M. (2011). Idea generation: customer co-creation versus traditional market research techniques. *Journal of Service Management*. 22 (2). p.pp. 140–159.

Appendix

1.0 Project Poster



2.0 Project Log Sheets

Log Sheet 1

(APU: Serial Number)

PLS V1.0

Project Log Sheet – Supervisory Session

Notes on use of the project log sheet:

1. This log sheet is designed for meetings of more than 15 minutes duration, of which there must be at minimum SIX (6) during the course of the project (SIX mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant sections of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. The student should leave a copy (after the session) of the Project Log Sheet with the supervisor and to the administrator at the academic counter. A copy is retained by the student to be filed in the project file.
6. It is recommended that students bring along log sheets of previous meetings together with the project file during each supervisory session.
7. The log sheet is an important deliverable for the project and an important record of a student's organisation and learning experience. The student **must** hand in the log sheets as an appendix of the final year documentation, with sheets dated and numbered consecutively.

Student's name: ALAVI BEEN AZAM **Date:** 26/06/2020 **Meeting No:** 01

Project title: Focusen: A sentiment analysis tool for online focus groups and surveys to understand consumer behaviour. **Intake:** UC3F2005IS

Supervisor's name: MR RAHEEM MAFAS **Supervisor's e-signature:** RMF

Items for discussion (noted by student before mandatory supervisory meeting):

1. Discuss about scope and objectives of the project.
2. Discuss about collaboration with market research company for the project and dataset.
3. Discussion about PSF.
4.

Record of discussion (noted by student during mandatory supervisory meeting):

1. Discussion and Feedback on the scope and objective of the project.
2. Discussion about company collab and dataset.
3. Discussion about PSF.
4.

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Complete PSF.
2. Discuss about official MPU with APU with company.
3.

Note: A student should make an appointment to meet his or her supervisor (via the consultation system) at least ONE (1) week prior to a mandatory supervisor session – please see document on project timelines. In the event a supervisor could not be booked for consultation, the project manager should be informed ONE (1) week prior to the session so that a meeting can be subsequently arranged.

Project Log Sheet

Log Sheet 2

(APU: Serial Number)

PLS V1.0

Project Log Sheet – Supervisory Session

Notes on use of the project log sheet:

1. This log sheet is designed for meetings of more than 15 minutes duration, of which there must be at minimum SIX (6) during the course of the project (SIX mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant sections of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. The student should leave a copy (after the session) of the Project Log Sheet with the supervisor and to the administrator at the academic counter. A copy is retained by the student to be filed in the project file.
6. It is recommended that students bring along log sheets of previous meetings together with the project file during each supervisory session.
7. The log sheet is an important deliverable for the project and an important record of a student's organisation and learning experience. The student **must** hand in the log sheets as an appendix of the final year documentation, with sheets dated and numbered consecutively.

Student's name: ALAVI BEEN AZAM **Date:** 30/07/2020 **Meeting No:** 02

Project title: Focusen: a sentiment analysis tool for online focus groups and surveys to understand consumer behaviour. **Intake:** UC3F2005IS

Supervisor's name: MR. RAHEEM MAFAS **Supervisor's e-signature:** RMF

Items for discussion (noted by student before mandatory supervisory meeting):

1. Discuss about requirment validation and how to the conduct research.
2. Discuss about the details of organizations and what to mention in the IR.
3. Discussion about ehithcs form approval and submission.
4. Ask for feedback on PSF.

Record of discussion (noted by student during mandatory supervisory meeting):

1. Do a questionnaire to collect data requirment validation along with interview to collect more data.
2. Collect data from 3 levels within the organizations (analyst, product manager, CTO).
3. Task the name of the organization to maintain confidentiality.
4. Follow up with questionnaire design next week.

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Complete Questionnaire design.
2. Collect data from organization.
3.

Note: A student should make an appointment to meet his or her supervisor (via the consultation system) at least ONE (1) week prior to a mandatory supervisor session – please see document on project timelines. In the event a supervisor could not be booked for consultation, the project manager should be informed ONE (1) week prior to the session so that a meeting can be subsequently arranged.

Project Log Sheet

Log Sheet 3

(APU: Serial Number)

PLS V1.0

Project Log Sheet – Supervisory Session

Notes on use of the project log sheet:

1. This log sheet is designed for meetings of more than 15 minutes duration, of which there must be at minimum SIX (6) during the course of the project (SIX mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant sections of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. The student should leave a copy (after the session) of the Project Log Sheet with the supervisor and to the administrator at the academic counter. A copy is retained by the student to be filed in the project file.
6. It is recommended that students bring along log sheets of previous meetings together with the project file during each supervisory session.
7. The log sheet is an important deliverable for the project and an important record of a student's organisation and learning experience. The student **must** hand in the log sheets as an appendix of the final year documentation, with sheets dated and numbered consecutively.

Student's name: ALAVI BEEN AZAM **Date:** 11/08/2020 **Meeting No:** 03

Project title: Focusen: A sentiment analysis tool for online focus groups and surveys to understand consumer behaviour. **Intake:** UC3F2005IS

Supervisor's name: MR RAHEEM MAFAS **Supervisor's e-signature:** RMF

Items for discussion (noted by student before mandatory supervisory meeting):

1. Get Feedback on IR
2.
3.
4.

Record of discussion (noted by student during mandatory supervisory meeting):

1. Feedback on IR draft.
2. Discussion about expert interview and questionnaire responses.
3.
4.

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. None.
2.
3.

Note: A student should make an appointment to meet his or her supervisor (via the consultation system) at least ONE (1) week prior to a mandatory supervisor session – please see document on project timelines. In the event a supervisor could not be booked for consultation, the project manager should be informed ONE (1) week prior to the session so that a meeting can be subsequently arranged.

Project Log Sheet

Log Sheet 4

(APU: Serial Number)

PLS V1.0

Project Log Sheet – Supervisory Session

Notes on use of the project log sheet:

1. This log sheet is designed for meetings of more than 15 minutes duration, of which there must be at minimum SIX (6) during the course of the project (SIX mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session and noting these in the relevant sections of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. The student should leave a copy (after the session) of the Project Log Sheet with the supervisor and to the administrator at the academic counter. A copy is retained by the student to be filed in the project file.
6. It is recommended that students bring along log sheets of previous meetings together with the project file during each supervisory session.
7. The log sheet is an important deliverable for the project and an important record of a student's organisation and learning experience. The student **must** hand in the log sheets as an appendix of the final year documentation, with sheets dated and numbered consecutively.

Student's name: ALAVI BEEN AZAM **Date:** 16/10/2020 **Meeting No:** 04

Project title: Focusen: Real Time Sentiment Analysis to Understand Consumer Behaviour. Intake: UC3F2005IS

Supervisor's name: MR. RAHEEM MAFAS **Supervisor's signature:** RMF

Items for discussion (noted by student before mandatory supervisory meeting):

1. Development plan for the project.
2. Finalise project scope and objectives.
3. Feedback on IR.
4.

Record of discussion (noted by student during mandatory supervisory meeting):

1. Makes necessary changes provide in IR feedback form.
2. Research the best python libraries available for sentiment analysis.
3. Data pipeline, model and frontend dashboard required.
4.

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Complete final research and finalise the libraries and packages to be used.
2. Provide details in the next meeting.
3.

Note: A student should make an appointment to meet his or her supervisor (via the consultation system) at least ONE (1) week prior to a mandatory supervisor session – please see document on project timelines. In the event a supervisor could not be booked for consultation, the project manager should be informed ONE (1) week prior to the session so that a meeting can be subsequently arranged.

Project Log Sheet

Log Sheet 5

(APU: Serial Number)

PLS V1.0

Project Log Sheet – Supervisory Session

Notes on use of the project log sheet:

1. This log sheet is designed for meetings of more than 15 minutes duration, of which there must be at minimum SIX (6) during the course of the project (SIX mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session and noting these in the relevant sections of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. The student should leave a copy (after the session) of the Project Log Sheet with the supervisor and to the administrator at the academic counter. A copy is retained by the student to be filed in the project file.
6. It is recommended that students bring along log sheets of previous meetings together with the project file during each supervisory session.
7. The log sheet is an important deliverable for the project and an important record of a student's organisation and learning experience. The student **must** hand in the log sheets as an appendix of the final year documentation, with sheets dated and numbered consecutively.

Student's name: ALAVI BEEN AZAM Date: 18/11/2020 Meeting No: 05 ~~05~~

Project title: Focusen: Real Time Sentiment Analysis to Understand Consumer Behaviour **Intake:** UC3F2005IS

Supervisor's name: MR RAHEEM MAFAS **Supervisor's signature:** RMF

Items for discussion (noted by student before mandatory supervisory meeting):

1. Show progress on project implementation.
2. Finalise deliverables scope and objectives.
3.
4.

Record of discussion (noted by student during mandatory supervisory meeting):

1. Open-source libraries are okay to use.
2. Pre-trained sentiment analysis model will be fine.
3. Dashboard should be deployed even if its local deployment.
4.

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Show progress on data pipeline and model.
2.
3.

Note: A student should make an appointment to meet his or her supervisor (via the consultation system) at least ONE (1) week prior to a mandatory supervisor session – please see document on project timelines. In the event a supervisor could not be booked for consultation, the project manager should be informed ONE (1) week prior to the session so that a meeting can be subsequently arranged.

Project Log Sheet

Log Sheet 6

(APU: Serial Number)

PLS V1.0

Project Log Sheet – Supervisory Session

Notes on use of the project log sheet:

1. This log sheet is designed for meetings of more than 15 minutes duration, of which there must be at minimum SIX (6) during the course of the project (SIX mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session and noting these in the relevant sections of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. The student should leave a copy (after the session) of the Project Log Sheet with the supervisor and to the administrator at the academic counter. A copy is retained by the student to be filed in the project file.
6. It is recommended that students bring along log sheets of previous meetings together with the project file during each supervisory session.
7. The log sheet is an important deliverable for the project and an important record of a student's organisation and learning experience. The student **must** hand in the log sheets as an appendix of the final year documentation, with sheets dated and numbered consecutively.

Student's name: ALAVI BEEN AZAM Date: 4/12/2020 Meeting No: 06 06

Project title: Focusen: Real Time Sentiment Analysis to Understand Consumer Behaviour **Intake:** UC3F2005IS

Supervisor's name: MR. RAHEEM MAFAS **Supervisor's signature:** RMF

Items for discussion (noted by student before mandatory supervisory meeting):

1. Show completed data pipeline, text pre-processing and ML model.
2. Discuss components to be added to the dashboard.
3. Discuss deployment plan.
4.

Record of discussion (noted by student during mandatory supervisory meeting):

1. Pipeline, pre-processing and model is fine.
2. Add more components to the dashboard.
3. Model should be deployed and available for ease use by target user.
4.

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Show completed dashboard.
2. Starting compiling the final documentation.
3.

Note: A student should make an appointment to meet his or her supervisor (via the consultation system) at least ONE (1) week prior to a mandatory supervisor session – please see document on project timelines. In the event a supervisor could not be booked for consultation, the project manager should be informed ONE (1) week prior to the session so that a meeting can be subsequently arranged.

Project Log Sheet

Log Sheet 7

(APU: Serial Number)

PLS V1.0

Project Log Sheet – Supervisory Session

Notes on use of the project log sheet:

1. This log sheet is designed for meetings of more than 15 minutes duration, of which there must be at minimum SIX (6) during the course of the project (SIX mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session and noting these in the relevant sections of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. The student should leave a copy (after the session) of the Project Log Sheet with the supervisor and to the administrator at the academic counter. A copy is retained by the student to be filed in the project file.
6. It is recommended that students bring along log sheets of previous meetings together with the project file during each supervisory session.
7. The log sheet is an important deliverable for the project and an important record of a student's organisation and learning experience. The student **must** hand in the log sheets as an appendix of the final year documentation, with sheets dated and numbered consecutively.

Student's name: ALAVI BEEN AZAM **Date:** 17/12/2020 **Meeting No:** 07

Project title: Focusen: Real Time Sentiment Analysis to Understand Consumer Behaviour. **Intake:** UC3F2005IS

Supervisor's name: MR. RAHEEM MAFAS **Supervisor's signature:** RMF

Items for discussion (noted by student before mandatory supervisory meeting):

1. Show completed dashboard.
2. Discuss about documentation and source-code submission.
3.
4.

Record of discussion (noted by student during mandatory supervisory meeting):

1. Dashboard is okay.
2. Add titles to the different visualisations on the dashboard.
3. Change neutral sentiment colour to grey.
4.

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Send documentation draft for final feedback before submission.
2.
3.

Note: A student should make an appointment to meet his or her supervisor (via the consultation system) at least ONE (1) week prior to a mandatory supervisor session – please see document on project timelines. In the event a supervisor could not be booked for consultation, the project manager should be informed ONE (1) week prior to the session so that a meeting can be subsequently arranged.

Project Log Sheet

3.0 Project Proposal Form (PPF)

Focusen: A sentiment analysis tool for online focus groups and surveys to understand consumer behavior.

Introduction

Consumer behaviour is described as a psychologically based study of how individuals make purchase decisions; it is the understanding of what motivates individuals to purchase a specific product or service. Several key characteristics are classified under the study of consumer behaviour such as how a consumer feels about certain brands, products and services, what motivates a consumer to choose one product over the others and why, what factors in a consumer's everyday environment affect their buying decisions or brand perception and how consumers make decisions in groups or when they are alone. Multiple factors are needed to be taken into consideration to understand the characteristics and determine consumer behaviour such as social factors, psychological factors, economic factors and even simple personal traits.

It is increasingly becoming more important for businesses to understand their consumers in order to stay relevant in the market and generate revenue by understanding the behaviour of its consumers and providing the right products and services according to their wants and needs. According to a salesforce report, it shows that 76% of consumers expect companies to understand their needs and expectations, which translates to if a consumer isn't happy with a product or service, they will most probably move to a competitor alternative. Most successful organizations take this into account and make their product development and marketing decision based on insights generated from their customer behaviour data. There are several ways of how companies can generate consumer behaviour insights. These insights may either be generated from existing company marketing or sales platforms or through other channels such surveys and focus groups.

Traditional approaches for studying and understanding consumer behaviour such as focus groups and marketing surveys have been popular for a long time but require a lot of time and resources to generate actionable insights. Over the past few decades the constant growth of the internet, increasing popularity and advancement in web and mobile technologies have moved several processes online and reachable by a greater number of people, surveys and focus groups are no different. Several organizations including market research and analytics companies have been conducting surveys and focus groups online in order to generate consumer insights and understand their behavioural traits. Even though moving things online and technology have made the process of gathering data relatively faster with significantly less resources but the actual process of combing through the data to clean, organize and analyse it to generate actionable insights has still remain a tedious process which requires a high level of statistical and technical expertise along with domain knowledge. Several advancements in Artificial Intelligence such as machine learning has provided individuals and organizations with several tools and techniques in order to make faster and more accurate computations to generate insights automatically with minimal to no human supervision. One of these techniques is known as sentiment analysis.

Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. To identify the underlying tone of the expression, sentiment analysis uses natural language processing (NLP).

Different types of sentiment analysis use different methods to identify the undertone of a phrase, it is mainly classified into two major types, subjective/objective identification and feature/aspect based sentiment analysis. Sentiment analysis can be used by organizations to identify customer sentiment towards products, services and brands in online conversations and feedback. Use of sentiment analysis techniques has been popular among organizations to generate insights by mining text data from sales and marketing platforms, social media and other sources to understand their consumer better and make product and marketing decisions accordingly. The advancement of AI technologies and big-data has enabled sentiment analysis to easily capture, quantify, retrieve and analyse consumers more effectively which businesses can use to better understand their strengths and weaknesses. This research aims to research further into the importance of consumer behaviour, data collection techniques and sentiment analysis to develop a tool which can generate and classify consumer emotions atomically from online focus group conversations and surveys using machine learning and sentiment analysis techniques to help provide better and faster insights for organizations to make product and marketing decisions based on consumer behaviour.

Problem Statement

1. **It is resource intensive and time consuming to manually analyze the data to extract insights from text data in online focus groups and open-ended survey responses.**

The process of analyzing data manually to generate insights after qualitative text data has been generated from online focus group and survey responses is heavily labour intensive and time consuming. The process often takes days to weeks for researchers in the organization to manually comb through the data to discover and report relevant insights. For research companies who are constantly conducting focus groups and survey this process can be highly tedious and in-efficient.

2. **Ruled-based systems lack optimal accuracy and are limited.**

In some organizations, often some rule-based classifiers are used to classify the qualitative text data to generate insights faster, but these rule-based systems have limited capabilities and can produce inconsistent results with low accuracy due to its hard coded nature and inability to learn from historical data, making the insights generated from the data unreliable.

3. **Data analyzed manually can be in-consistent and be prone to human bias.**

Humans often get tired, disagree with one another and can have different ways of interpreting things based on personality and other psychological factors. So, it highly difficult for individuals and teams to tag text responses generated from surveys and focus groups consistent without fluctuations or biases which can result in inconsistent and biased insights may not be totally accurate towards understanding consumer behaviour.

4. **Real-time analysis and insights are not possible**

It is not possible for businesses to receive instant realtime feedback from their consumers using manual analysis and classification which can lead to missed opportunities and quick testing of ideas and hypothesis which may generate actionable insights for swift business decisions.

5. **Lack of scalability**

Market research firms and other research entities often produce thousands of text-based responses from surveys and focus groups every week, it is highly unrealistic for researchers to analyze such high volumes of data manually to generate reports on consumer behavior. Organizations need to dedicate extensive human resources to tackle this issue which can be dedicated to other human-dependent task if the process is automated which translates to lower operational costs making market research significantly cheaper and more accessible for companies.

Project Aim & Objectives

The aim of this project is to develop a NLP based sentiment analysis tool which will allow the research firm to automatically tag and analyze online focus group and survey responses in minutes to generate quick real-time insights on consumer sentiment towards brands, products and services and unlock other valuable consumer behaviour insights that can help business identify their strengths and weaknesses and also help make better product decisions influenced by consumers and backed by data.

The objectives of this project are:

- Developing a tool that can automatically tag and analyze text based qualitative responses using a sentiment analysis-based machine learning model to generate quick real-time insights into consumer behaviour trained using previous survey and focus group responses.
- To evaluate the accuracy of the model in tagging responses and classifying them as positive, negative or neutral.
- To make real-time prediction from new data provided through the web-app. The tool should be able to accurately predict the polarity score of as given text-based phrase or sentence extracted from the surveys and online focus group responses.

Literature Review

To evaluate the chosen project, data relevant to the topic, facts and other previous literature based on the concepts of consumer behaviour and sentiment analysis need to be reviewed and researched to support the notion of this project. There are a lot of previous literature that discusses the importance and different aspects of consumer behaviour and also similar literature can be found on sentiment analysis. But no specific literature can be found on using sentiment analysis for online focus groups to generate consumer behaviour insights. All the previous literature that is relevant to the project has been analysed so that the researcher can evaluate, learn and improve from the work previously done in this field and propose new ideas and solutions in solving the problem.

Consumer insight is the study about how the consumer is and what they think and feel (Stone, n.d.). Insight is not conscious behaviours or thoughts of consumers but most are affected by various external factors from the state of economy and society to the way a brand is marketed (Chamlertwat et al., n.d.). Consumer-based strategy is organizational strategy that is developed

based on insights about consumers. Such strategy can be developed based on understanding consumers' wants and needs (Lam et al.; Olson), the costs consumers incur to purchase and own goods and services (Choi et al. Zielke and Komor), the convenience of obtaining goods and services (Baker and Wakefield; Lim et al.), or what makes communication between the organization and the consumer more effective (Brasel and Gips; Mikolon et al.). All of these insights share a focus on consumers as the unit of analysis: data about needs, costs of purchase and ownership, convenience, and communication effectiveness can be collected for each consumer. A consumer-based strategy lets businesses drive their product and marketing decisions based on insight generated from consumers either through internal business sales and service platforms or through market research.

An important part of our information-gathering behaviour has always been to find out what other people think. With the growing availability and popularity of opinion-rich resources such as online review sites and personal blogs, new opportunities and challenges arise as people now can, and do, actively use information technologies to seek out and understand the opinions of others. The sudden eruption of activity in the area of opinion mining and sentiment analysis, which deals with the computational treatment of opinion, sentiment, and subjectivity in text, has thus occurred at least in part as a direct response to the surge of interest in new systems that deal directly with opinions as a first-class object (Pang & Lee, n.d.). Sentiment analysis can be used to extract insight to understand consumer emotion towards a product, service or brand.

Deliverables

- Develop and train a ML model for sentiment analysis using historical text data from survey and focus group responses to predict consumer sentiment and classify them as positive, negative or neutral.
- A web-app hosted in AWS that will let users upload csv/json documents for the model to analyze, predict and display results.
- An ETL pipeline to extract, load and transform user input data to pre-process it for the machine learning model to make predictions.

4.0 Project Specification Form (PSF)

1. Brief description on project background. (.i.e. problem context, rationale, description of problem area, nature of challenge.

1.1 Problem Context

It is increasingly becoming more important for businesses to understand their consumers in order to stay relevant in the market and generate revenue by understanding the behaviour of its consumers and providing the right products and services according to their wants and needs. According to a salesforce report, it shows that 76% of consumers expect companies to understand their needs and expectations, which translates to if a consumer isn't happy with a product or service, they will most probably move to a competitor alternative. Most successful organizations take this into account and make their product development and marketing decision based on insights generated from their customer behaviour data. Consumer behaviour is described as a psychologically based study of how individuals make purchase decisions; it is the understanding of what motivates individuals to purchase a specific product or service.

Traditional approaches for studying and understanding consumer behaviour such as focus groups and marketing surveys have been popular for a long time but require a lot of time and resources to generate actionable insights. Several organizations including market research and analytics companies are now conducting surveys and focus groups online in order to generate consumer insights and understand their behavioural traits. Even though moving things online and technology have made the process of gathering data relatively faster with significantly less resources but the actual process of combing through the data to clean, organize and analyse it to generate actionable insights has still remain a tedious process which requires a high level of statistical and technical expertise along with domain knowledge.

Several advancements in Artificial Intelligence such as machine learning has provided individuals and organizations with several tools and techniques in order to make faster and more accurate computations to generate insights automatically with minimal to no human supervision. One of these techniques is known as sentiment analysis. Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. To identify the underlying tone of the expression, sentiment analysis uses natural language processing (NLP). This project aims to use sentiment analysis in order to generate reliable and actionable insight from online focus group conversations and marketing surveys with quick turnaround time and minimal human input.

1.2 Rationale

All of the issues highlighted above in the problem context show the importance of consumer sentiment for brands and why it required to generate reliable insights fast which can influence critical products, services and business decisions for organizations. Therefore, marketing research teams and companies need to find a way to generate reliable and actionable insights fast instead of relying on human experts who has to manually comb through the data to classify responses to produce reports or rule based systems which have limited capabilities and can portray a significant amount of bias in the insight generated. Also, it is resource incentive to find experts who have strong knowledge in several domains and it also impossible to build rule-based system from every use-case since these are not intelligent and adaptative system.

By implementing an AI based sentiment analysis model trained with a diverse dataset to account for expert knowledge in several domain areas, organisations can easily generate insights in a shorter timeframe with higher accuracy and less bias while also saving a huge amount of resources which can be deployed in other business areas or help them provide their services at a more affordable cost to get more clients . It will also help market research companies to generate more insights and take on more clients and help them make better data driven decision based on reliable insights generated using sentiment analysis and ML techniques by automating the entire ETL to insight pipeline.

- Tangible benefits:
  - Analytics can be performed faster but automating the entire process from data collection to generating insights.
  - Generate more accurate and reliable insights for better data driven decision.
  - Minimise inconsistency in data and reduce bias formed by human experts.
  - The capability to generate real-time insights which can help in time-critical data driven decisions.
  - Provide a more scalable and adaptive approach to generating consumer insights.

- Intangible benefits:
  - Ensure a higher level of client satisfaction by generating reliable insights faster which can help them make better business decisions.
  - Help reduce operation cost for conducting online focus groups and survey since significantly less manpower is required to generate reports since most of the process are online and automated using machine learning.

1.3 Description of the problem

The problem area includes building a ETL (extract, transform & load) data pipeline which will take text data from online focus group conversations and surveys as input. Next the data will be passed through a sentiment analysis based machine learning model which will evaluate the data to determine the sentiment of the text, its polarity and other text based analytic to generate a report which will easily be available on a web-based dashboard to provide insight into the consumer sentiment and help drive more accurate data-driven decisions.

1.4 Nature of the challenge

The nature of the challenge is to develop a sentiment analysis based machine learning model which can adapt to data from several different industries and domain areas since most data analytics and market research companies work with a diverse set of clients who require insights to be generated in a variety of different fields. Another challenge would be localization since the main target market for the tool is Malaysia which has people speaking several different languages which can be a challenge since the model will pre-dominantly be trained on the English language so if the data has input in Malay, Chinese or Tamil mixed with English it can become difficult for the model to generate reliable insights.

2. Brief description of project objectives.(i.e. scope of proposal and deliverables).

2.1 Scope of the proposal

The scope of the proposal is to develop an end to end sentiment analysis tool which will take clean text data from online focus group conversations and survey responses using an ETL data pipeline and generate a text analytics report using a sentiment analysis based machine learning model on a web-based dashboard showing the sentiment of the text, its polarity and other text based insights.

2.2 Deliverables

The deliverables of this project will be a full end to end sentiment analysis tool which will include but are not limited to a data pipeline to process the input data, a ml model to evaluate the data to generate insights, a web-based report providing a detailed report on the output of the model and data analytics showing the a detailed breakdown on the sentiment report, data visualizations to make the report more interactive and easy to use. The developer will work with a popular market research company to understand the requirements better and develop a working solution for the problem stated.

- Design, architect and build a data pipeline to ingest and process cleaned text data for ml model.
- Perform extract, transform and load operations.
- Develop a ML model to generate sentiment analysis based analytics.
- Provide a web-based dashboard to make it easier for user to upload the data and display the report generated by the model.
- Data visualizations to make the report more interactive and engaging.
- To tool to generate real-time insights from online focus group conversations and survey data.

3. Brief description of the resources needed by the proposal (i.e. hardware, software, access to information / expertise, user involvement etc.)

3.1 Hardware

The projects requires some specific hardware to completed according to its SDLC requirements. The developer will use an Apple Macbook Pro 15inch (2019) as the primary device for research and development of this projects.

- Processor: 2.3 GHz 8-Core Intel Core i9
- Memory: 16GB 2400 MHz DDR4
- Graphics: Radeon Pro 560X 4GB | Intel UHD Graphics 630 1536 MB
- Solid State Storage SSD: Apple 500gb
- RJ45/ Wireless Fidelity (Wi-Fi)
- Other peripherals – monitors, wireless keyboard, wireless mouse and router (no specific model according to requirements and able to serve the required purpose.)

3.2 Software

Along with the hardware the project also has some specific software requirements which are listed below.

- Operating System: Apple OSX MacOS Catalina (Version 10.15.5) and newer
- Web Browser: Google Chrome (Version 83.0.4103), Apple Safari (Version 13.1.1)
- Integrated Development Environment (IDE): Visual Studio Code (Version 1.47.0), PyCharm (Version 2019.3.5), Jupyterlab and Google Colaboratory (Online).
- Programming Language: Python 3.7
- Developments frameworks: Flask 1.1.X, Vue.Js (Version 2.6.11 Stable)
- Command Line tools: Terminal, iterm2, Vue CLI, AWS SDK
- Cloud Service: Amazon Web Services (For deployment and model training jobs) – Amazon EC2, Amazon S3, Amazon API Gateway
- API Management – Postman
- Documentation and Planning – Microsoft Office word, Microsoft Excel and Microsoft Project (latest version available).

3.3 Access to information / expertise

In order to complete this project, consultation with a market research company is required to understand the requirements and performance criteria of the sentiment analysis tool along with online research being conducted by the developer. The data will be collected through a series of online interviews and all the data used from the organizations will be strictly under the supervision of project supervisor, project manager and other relevant authorities.

3.4 User Involvement

The primary users of this tool will be the researcher and analyst with the market research company. Some feedback and requirements will also be collected from the users through online interviews to better understand the functional requirements for the tool and also to evaluate the success matrix of the tool once deployed compared to human analysts and rule-based systems.

4. **Academic research being carried out and other information, techniques being learned.(i.e. what are the names of books you are going to read / data sets you are going to use).**

In order to fully understand the functional and non-functional requirements of the project and develop a tool which can handle the requirements and follow software development best practise the developer has conducted extensive research on the subject matter. Research has also been conducted to understand and evaluate pervious literature available in within the domain of the project.

- Knowledge areas in focus:

  - Consumer behaviour and insights.
  - Sentiment analysis and machine learning.
  - Data pipeline and ETL architecture.
  - Text Analysis
  - Data analytics and visualization.
  - AWS cloud architecture.

Also for further learning and understanding some books have also been referenced for research.

1. Data Science for Scratch (2nd Edition) by Joel Grus (2019) O`Reilly
2. Hacker`s Guide to Machine Learning by Venelin Valkov (2019) Leanpub
3. Hands-on Machine Learning with Scikit-Learn, Keras & Tensorflow (2nd Edition) by Aurelein Geron (2020) O`Reilly
4. Automate the Boring Stuff with Python by Albert Sweigart (2015) No Strach Press

Also several online resources have been referenced while conducting research for this project, which are referenced below.

Anon (2018). *Sentiment Analysis*. [Online]. 20 June 2018. MonkeyLearn. Available from: https://monkeylearn.com/sentiment-analysis. [Accessed: 30 May 2020].

Anon (n.d.). *State of the Connected Customer Report Outlines Changing Standards for Customer Engagement*. [Online]. Salesforce.com. Available from: https://www.salesforce.com/company/news-press/stories/2019/06/061219-g/. [Accessed: 30 May 2020a].

Anon (n.d.).

Chamlertwat, W., Bhattarakosol, P., Rungkasiri, T. & Haruechaiyasak, C. (n.d.). *Discovering Consumer Insight from Twitter via Sentiment Analysis*. p.p. 20.

economist, S.S. is an aspiring, data, analytics enthusiast S. has completed her M. of S. in E. from the U. of C.S. articulates insights covering various technologies including big, analytics, blockchain & More, M. (2018). Benefits of Sentiment Analysis for Businesses. *Analytics Insight*. [Online]. Available from: https://www.analyticsinsight.net/benefits-of-sentiment-analysis-for-businesses/. [Accessed: 31 May 2020].

Pang, B. & Lee, L. (n.d.). *Opinion Mining and Sentiment Analysis | Foundations and Trends in Information Retrieval*. [Online]. Available from: https://dl.acm.org/doi/10.1561/1500000011. [Accessed: 1 June 2020].

Stone, M. (n.d.). *Consumer Insight: How to Use Data and Market Research to Get Closer to Your Customer (Market Research in Practice)*.

Team, D.J. (n.d.). *The Importance of Consumer Behavior in Marketing*. [Online]. Available from: https://www.demandjump.com/blog/the-importance-of-consumer-behavior-in-marketing. [Accessed: 30 May 2020].

Dataset:

Also to develop the model the developer has requested access for a private dataset from a market research company which contains sizeable raw text data from past online focus group conversations and survey response from primarily Malaysian consumers on a variety of subject across several consumer industries.

5. Brief description of the development plan for the proposed project.(i.e. which software methodology and why, the major areas of functions to be developed and the order in which developed)

The researcher has chosen to adopt the CRISP-DM software development methodology for this project as it provide a good framework tailored towards data science and machine learning projects. Although the methodology is geared more towards data-mining the framework argues to provide a solid base for all kinds of data science projects. The methodology has six major iterative phases, each with its own defined tasks and set of deliverables. The phases are:

- Business Understanding: determine the business objectives, asses situations and determine the output goals.
- Data Understanding: collect initial data, describe data and verify data quality
- Data Preparation: Select data, clean data, construct data, integrate data and format data. (This step is the most time consuming)
- Modelling: Select modelling technique, generate test design, build model and asses the model.
- Evaluation: Evaluate results, review process and determine the next steps.
- Deployment: Plan deployment, plan monitoring and maintenance, produce final report and review project.

The order in which the project will be developed is:

1. Data pipeline
2. Model development and evaluation
3. Building web-based dashboard for reporting.
4. Deployment to aws cloud services.

6. Brief description of the evaluation and test plan for the proposed project.(i.e. what is the success criteria and how will be evaluated & implementation will be tested, indicate the estimated size of the demonstration/test database).

Success Criteria:

The success criteria of the sentiment analysis tool can be determined based on the accuracy of its predication on test data in terms of sentiment polarity and other text analytics factors. The

prediction accuracy needs to meet a certain pre-determined threshold to be considered a success.

Evaluation & Implementation test:

The dataset will be split into training, testing and evaluation subsets. After the model is trained using the training dataset, the test and evaluation datasets will be used to determine its accuracy and evaluate its performance criteria based on the model`s accuracy, precision, recall, F1-score, ROC, AUC, Regression Metrics and several other standard ml model evaluation metric measures to determine whether the model is ready for deployment. Once the model has based the evaluation is will be connected to a data pipeline which will text csv dataset or plain text input using a web dashboard and the model will return a text analytics report based on the data provided as input.

Estimated Size of Data:

The dataset size is estimated to be around 20GB of text-based online focus group conversations and survey responses.

5.0 Ethics Form

Office Record	Receipt – Fast-Track Ethical Approval
Date Received:	Student name: Student number:
Received by whom:	Received by: Date:

APU / APIIT FAST-TRACK ETHICAL APPROVAL FORM (STUDENTS)

Tick one box (level of study):
- ☐ POSTGRADUATE (PhD / MPhil / Masters)
- ☑ UNDERGRADUATE (Bachelors degree)
- ☐ FOUNDATION / DIPLOMA / Other categories

Tick one box (purpose of approval):
- ☑ Thesis / Dissertation / FYP project
- ☐ Module assignment
- ☐ Other: _____

Title of Programme on which enrolled … BSc (Hons) in Intelligent Systems …

Tick one box: ☑ Full-Time Study or ☐ Part-Time Study

Title of project / assignment … Focusen: A sentiment analysis tool for online focus groups and surveys to understand consumer behaviour. …

Name of student researcher … ALAVI BEEN AZAM …

Name of supervisor / lecturer… MR. RAHEEM MAFAS …

Student Researchers- please note that certain professional organisations have ethical guidelines that you may need to consult when completing this form.

Supervisors/Module Lecturers - please seek guidance from the Chair of the APU Research Ethics Committee if you are uncertain about any ethical issue arising from this application.

		YES	NO	N/A
1	Will you describe the main procedures to participants in advance, so that they are informed about what to expect?	✓		
2	Will you tell participants that their participation is voluntary?	✓		
3	Will you obtain written consent for participation?	✓		
4	If the research is observational, will you ask participants for their consent to being observed?	✓		
5	Will you tell participants that they may withdraw from the research at any time and for any reason?	✓		
6	With questionnaires and interviews will you give participants the option of omitting questions they do not want to answer?	✓		
7	Will you tell participants that their data will be treated with full confidentiality and that, if published, it will not be identifiable as theirs?	✓		
8	Will you give participants the opportunity to be debriefed i.e. to find out more about the study and its results?	✓		

If you have ticked **No** to any of Q1-8 you should complete the full Ethics Approval Form.

		YES	NO	N/A
9	Will your project/assignment deliberately mislead participants in any way?		✓	
10	Is there any realistic risk of any participants experiencing either physical or psychological distress or discomfort?		✓	
11	Is the nature of the research such that contentious or sensitive issues might be involved?		✓	

If you have ticked **Yes** to 9, 10 or 11 you should complete the full Ethics Approval Form. In relation to question 10 this should include details of what you will tell participants to do if they should experience any problems (e.g. who they can contact for help). You may also need to consider risk assessment issues.

Fast-Track Ethical Approval Form ver. 3.0 (Dec 2015) Page 1 of 4

			YES	NO	N/A
12	Does your project/assignment involve work with animals?			✓	
13	Do participants fall into any of the following special groups? **Note that you may also need to obtain satisfactory clearance from the relevant authorities**	Children (under 18 years of age)		✓	
		People with communication or learning difficulties			
		Patients			
		People in custody			
		People who could be regarded as vulnerable			
		People engaged in illegal activities (eg drug taking)			
14	Does the project/assignment involve external funding or external collaboration where the funding body or external collaborative partner requires the University to provide evidence that the project/assignment had been subject to ethical scrutiny?			✓	

If you have ticked **Yes** to 12, 13 or 14 you should complete the full Ethics Approval Form. There is an obligation on student and supervisor to bring to the attention of the APU Research Ethics Committee any issues with ethical implications not clearly covered by the above checklist.

STUDENT RESEARCHER

Provide in the boxes below (plus any other appended details) information required in support of your application, THEN SIGN THE FORM.

	Please Tick Boxes
I consider that this project/assignment has no significant ethical implications requiring a full ethics submission to the APU Research Ethics Committee.	✓
Give a brief description of participants and procedure (methods, tests used etc) in up to 150 words. Questionnaires will be used to collect data from subject matter experts to validate the requirments of this project.	
I also confirm that: i) All key documents e.g. consent form, information sheet, questionnaire/interview are appended to this application. Or	
ii) Any key documents e.g. consent form, information sheet, questionnaire/interview schedules which need to be finalised following initial investigations will be submitted for approval by the project/assignment supervisor/module lecturer before they are used in primary data collection.	✓

E-signature... ... ... Print Name... ALAVI BEEN AZAM ... ... Date 5/08/2020
(Student Researcher)

Please note that any variation to that contained within this document that in any way affects ethical issues of the stated research requires the appending of new ethical details. New ethical consent may need to be sought.

The completed form (and any attachments) should be submitted for consideration by your Supervisor/Module Lecturer

SUPERVISOR/MODULE LECTURER
PLEASE CONFIRM THE FOLLOWING:

	Please Tick Box
I consider that this project/assignment has no significant ethical implications requiring a full ethics submission to the APU Research Ethics Committee	✔
i) I have checked and approved the key documents required for this proposal (e.g. consent form, information sheet, questionnaire, interview schedule) Or	✔
ii) I have checked and approved draft documents required for this proposal which provide a basis for the preliminary investigations which will inform the main research study. I have informed the student researcher that finalised and additional documents (e.g. consent form, information sheet, questionnaire, interview schedule) must be submitted for approval by me before they are used for primary data collection.	

SUPERVISOR AND SECOND ACADEMIC SIGNATORY

STATEMENT OF ETHICAL APPROVAL (please delete as appropriate)

1) **THIS PROJECT/ASSIGNMENT HAS BEEN CONSIDERED USING AGREED APU/SU PROCEDURES AND IS NOW APPROVED**

2) **THIS PROJECT/ASSIGNMENT HAS BEEN APPROVED IN PRINCIPLE AS INVOLVING NO SIGNIFICANT ETHICAL IMPLICATIONS, BUT FINAL APPROVAL FOR DATA COLLECTION IS SUBJECT TO THE SUBMISSION OF KEY DOCUMENTS FOR APPROVAL BY SUPERVISOR (see Appendix A)**

E-signature RMF ... Print Name Raheem Mafas ... Date 06.08.2020
(Supervisor/Lecturer)

E-signature ... Print Name ... Date ...
(Second Academic Signatory)

Office Record	Receipt – Appendix A (Fast-Track Ethics Form)
Date Received:	Student name: Student number:
Received by whom:	Received by: Date:

APPENDIX A
AUTHORISATION FOR USE OF KEY DOCUMENTS

Completion of Appendix A is required when for good reasons key documents are not available when a fast track application is approved by the supervisor/module lecturer and second academic signatory.

I have now checked and approved all the key documents associated with this proposal e.g. consent form, information sheet, questionnaire, interview schedule

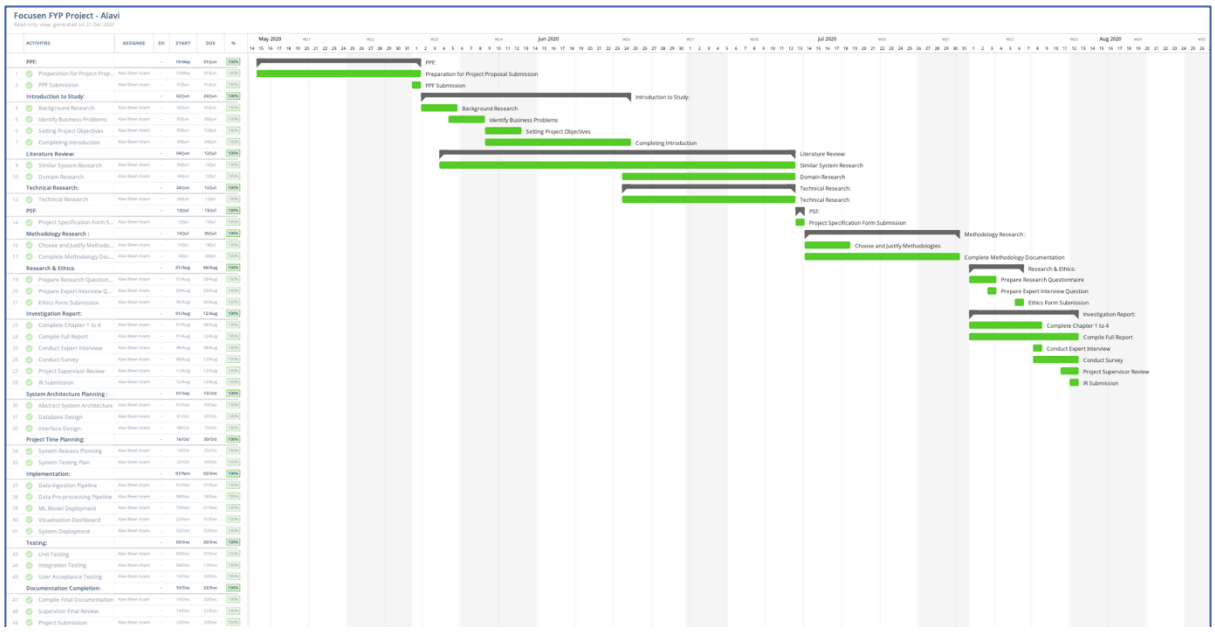
Title of project/assignment... Focusen: A sentiment analysis tool for online focus groups and surveys to understand consumer behaviour. .....

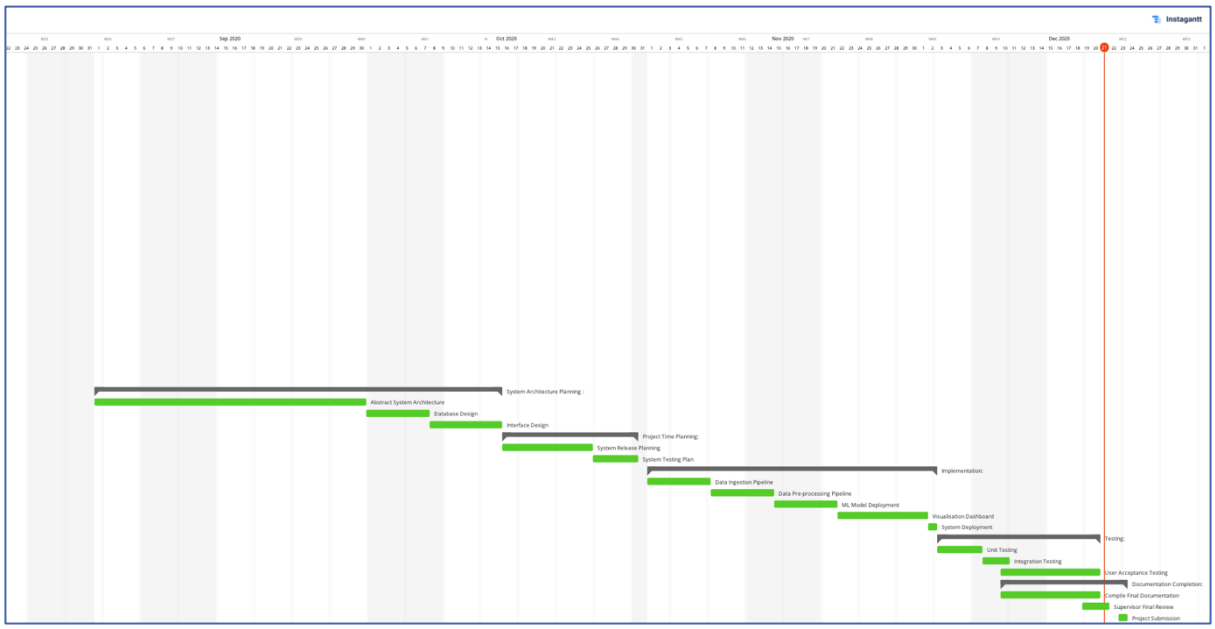
Name of student researcher ... ALAVI BEEN AZAM ....

Student ID: ..... TP041230 ..... Intake: ... UC3F2005IS ...

E-signature... ... Print Name... MR. RAHEEM MAFAS Date... ...
(Supervisor/Lecturer)

6.0 FYP Gantt Chart





Foucsen User Acceptance Testing

Foucsen: Real Time Sentiment Analysis to Understand Consumer Behaviour is a proposed system designed for Market Researcher to collect and analyse consumer opinion in real-time on social media and various other data sources using an web-based interactive dashboard as demonstrated.

The core features of the system are:
1. Sentiment analysis classifications in real-time.
2. Top keyword tracking.
3. Geographic segmentation & distribution of consumers.
4. Time series analysis to understand what consumers are saying about certain topics in real-time.

Thank you for taking part in this user acceptance testing exercise. Your help is greatly appreciated. Please kindly use the form below to express our opinion towards the product that have been demonstrated by the developer.

Cheers,
Alavi Been Azam
Asia Pacific University

Name: *

Omar Shabab

Current Designation: *

ML Engineer

Date of Testing: *

MM DD YYYY

12 / 20 / 2020

Testing Duration:

20 mins

How would you like to rate the user interface of the system? *

	1	2	3	4	5	
Poor	○	○	○	◉	○	Excellent

Does the proposed system meet the business objectives defined by the developer?

- ◉ Yes
- ○ No

How well are the objectives defined being meet by the system? *

	1	2	3	4	5	
Poor	○	○	◉	○	○	Excellent

How would you like to rate the overall performance of the system? *

	1	2	3	4	5	
Very Buggy	○	○	○	◉	○	Smooth

How would you like to rate the user friendliness of the system?

	1	2	3	4	5	
Poor	○	○	○	◉	○	Excellet

Feedback from Tester: *

Overall good execution, more features need to be added for deployment

Additional comments: *

None

Foucsen User Acceptance Testing

Foucsen: Real Time Sentiment Analysis to Understand Consumer Behaviour is a proposed system designed for Market Researcher to collect and analyse consumer opinion in real-time on social media and various other data sources using an web-based interactive dashboard as demonstrated.

The core features of the system are:
1. Sentiment analysis classifications in real-time.
2. Top keyword tracking.
3. Geographic segmentation & distribution of consumers.
4. Time series analysis to understand what consumers are saying about certain topics in real-time.

Thank you for taking part in this user acceptance testing exercise. Your help is greatly appreciated. Please kindly use the form below to express our opinion towards the product that have been demonstrated by the developer.

Cheers,
Alavi Been Azam
Asia Pacific University

Name: *

Hang Zhi Theng

Current Designation: *

Data scientist

Date of Testing: *

MM DD YYYY

12 / 14 / 2020

Testing Duration:

30 minutes

How would you like to rate the user interface of the system? *

	1	2	3	4	5	
Poor	○	○	○	◉	○	Excellent

Does the proposed system meet the business objectives defined by the developer?

- ◉ Yes
- ○ No

How well are the objectives defined being meet by the system? *

	1	2	3	4	5	
Poor	○	○	○	◉	○	Excellent

How would you like to rate the overall performance of the system? *

	1	2	3	4	5	
Very Buggy	○	○	○	◉	○	Smooth

How would you like to rate the user friendliness of the system?

	1	2	3	4	5	
Poor	○	○	○	◉	○	Excellet

Feedback from Tester: *

Good in overall but could add more interactive feature

Additional comments: *

-

Foucsen User Acceptance Testing

Foucsen: Real Time Sentiment Analysis to Understand Consumer Behaviour is a proposed system designed for Market Researcher to collect and analyse consumer opinion in real-time on social media and various other data sources using an web-based interactive dashboard as demonstrated.

The core features of the system are:
1. Sentiment analysis classifications in real-time.
2. Top keyword tracking.
3. Geographic segmentation & distribution of consumers.
4. Time series analysis to understand what consumers are saying about certain topics in real-time.

Thank you for taking part in this user acceptance testing exercise. Your help is greatly appreciated. Please kindly use the form below to express our opinion towards the product that have been demonstrated by the developer.

Cheers,
Alavi Been Azam
Asia Pacific University

Name: *

Lee Kah Kin

Current Designation: *

Web Developer

Date of Testing: *

MM DD YYYY

12 / 20 / 2020

Testing Duration:

0.5 hours

How would you like to rate the user interface of the system? *

	1	2	3	4	5	
Poor	○	○	○	◉	○	Excellent

Does the proposed system meet the business objectives defined by the developer?

- ◉ Yes
- ○ No

How well are the objectives defined being meet by the system? *

	1	2	3	4	5	
Poor	○	○	○	◉	○	Excellent

How would you like to rate the overall performance of the system? *

	1	2	3	4	5	
Very Buggy	○	○	○	○	◉	Smooth

How would you like to rate the user friendliness of the system?

	1	2	3	4	5	
Poor	○	◉	○	○	○	Excellet

Feedback from Tester: *

Colour scheme can be a little difficult to see. Wish the fonts are slightly bigger so that I can see it without zooming in the web page.

Additional comments: *

-

Foucsen User Acceptance Testing

Foucsen: Real Time Sentiment Analysis to Understand Consumer Behaviour is a proposed system designed for Market Researcher to collect and analyse consumer opinion in real-time on social media and various other data sources using an web-based interactive dashboard as demonstrated.

The core features of the system are:
1. Sentiment analysis classifications in real-time.
2. Top keyword tracking.
3. Geographic segmentation & distribution of consumers.
4. Time series analysis to understand what consumers are saying about certain topics in real-time.

Thank you for taking part in this user acceptance testing exercise. Your help is greatly appreciated. Please kindly use the form below to express our opinion towards the product that have been demonstrated by the developer.

Cheers,
Alavi Been Azam
Asia Pacific University

Name: *

Mohammed Fazalullah Qudrath

Current Designation: *

Senior Solutions Architect

Date of Testing: *

MM DD YYYY

12 / 20 / 2020

Testing Duration:

20 mins

How would you like to rate the user interface of the system? *

	1	2	3	4	5	
Poor	○	○	○	◉	○	Excellent

Does the proposed system meet the business objectives defined by the developer?

- ◉ Yes
- ○ No

How well are the objectives defined being meet by the system? *

	1	2	3	4	5	
Poor	○	○	○	○	◉	Excellent

How would you like to rate the overall performance of the system? *

	1	2	3	4	5	
Very Buggy	○	○	○	○	◉	Smooth

How would you like to rate the user friendliness of the system?

	1	2	3	4	5	
Poor	○	○	○	◉	○	Excellet

Feedback from Tester: *

Keyword input field in the dashboard will be a nice to have.

Additional comments: *

Mobile friendly dashboard to stack the charts vertically as a future roadmap