

**Sentiment Analysis of Unstructured, Website-based Client Reviews and Feedback Using Natural Language Processing**

Project Documentation Submitted

to the Faculty of School of

Computing and Information Technologies

of

Asia Pacific College

In Fulfillment of the Requirements for the subject

Systems Analysis & Detailed Design for CS-SS

by

Baldelomar, James Daniel M.

Dimaranan, Hugo S.

Melad, Nathaniel D.

Morales, Joshua M.

Valguna, Rene Hervy G.

August 22, 2017

**Table of Contents**

I. Introdution**1**

Type chapter title (level 2)2

**Type chapter title (level 1)4**

Type chapter title (level 2)5

Type chapter title (level 3)6

**I. Introduction**

**1.1 Project Context**

The SM Hotels company has been struggling to create reports efficiently. Their traditional methods that would require personnel to sacrifice too much time reading individual reviews. This just doesn’t scream inefficiency it may also be counterproductive depending who’s assigned for the work. Their point-of-view could be biased as there is no standard in what is a positive and negative review. This in-turn, could lead to slow production of potentially faulty reports.

Our client ‘SM Hotels’ has given us the opportunity to create for them; a program that would fully automate their previous traditional methods. Their specifications and needs have been consulted with on several visits/seminars.

This would be an opportune moment to upgrade their system to a fully automated machine. Rather than hustling monthly to produce sentimental reports, a simple click of a button would suffice to not only analyze the reviews from external sites but also produce the report which is already tabulated and ready to be presented to the respective managers.

**1.2 Purpose and Description**

This project is aimed to improve SM Hotels’ efficiency in producing reliable monthly sentimental reports. Currently these reports are produced traditionally by manually reading each individual review on mainstream websites. Once read the staff that is assigned for the task would then have to classify them as either a positive or negative review and tabulate them neatly on excel file where it is later submitted to their respective manager.

What we intend to create is a program that would do all the work mentioned before with click of a button. The project is designed to refine their previously inefficient methods with the use of computers. Not only does this eliminate the need to assign staff for menial tasks it would only need an employee to supervise the program and turn over the reports to their respective managers.

**1.3 Objectives**

Sentiment analysis of unstructured, website-based client reviews and feedback focuses on one main resolve, to completely automate the whole process of assessing online reviews using Natural Language Processing. The method of manually harvesting reviews from different hotels and traveling sites is still being implemented by SM Hotels and Conventions Corporation just to generate accurate managerial plans and come up with the congruent strategic solution to the problems indicated among the reviews.

To be specific, one of narrowed down objectives of the study is to gather feedback from websites without doing it manually. Web scraping, the automated process of extracting data from websites, will be the appropriate method of getting the job done. Another objective is not only to simplify the way of gathering input but also to teach computers to analyze date by themselves so that it would be a lot easier. Humans can perform arithmetic and logical operations as effective as how computers do the same. However, when tasks come in vast numbers, there is no way a person could outmatch a computer when it comes to performing simultaneous and numerous tasks in an instant.

**1.4 Scope and Limitations**

The project will cover one of the monthly reports that is submitted by SM hotels the “Sentimental Analysis”. This is regarding the reviews that could be found on external websites that promote and showcase several hotel selections such as ‘Trip adviser’ etc. The project would only apply to the mainstream review sites as mentioned before. The project will not include individual blogs from different personalities as stated by the client.

The project aims to increase the efficiency by the traditional method done by the staff members of SM hotels. The traditional method is done by manually reading each review for the duration of the month by the end of the month on a monthly basis. Categorize each as either a positive or negative review. Then tally them on an excel table and present it. The project aims to improve this in such a way that it would only take a click of a button for a single employee to gather data from predetermined external websites, sort the data and fill it out in a table for presentation.

The project would also include the Natural Language Processing (NLP) algorithm as it should be able to detect preset words to determine whether it is a positive or negative review. However, the NLP is still unreliable if the review were to consist of sarcasm as the language tends to be obscure. Also, as specified by the client, the project would also record reoccurring words for each category. We will only tally the reoccurring words and list how often it was used. We will NOT include the reasons to why it is reoccurring as specified also by the client.

The project also wants to verify if whether the review given by the customers are genuine or not, but for now, our focus is just analyzing the reviews and extracting sentiments out of it. We will include it in our future research.

**II. Review of Related Literature**

**2.1 Finding the Right Algorithm for Sentiment Analysis**

Sentiment Analysis plays a big role when it comes to identifying and classifying opinions of certain documents, and many people are still debating on which among invented algorithms works best for it.

Pang, B., Lee, L., & Vaithyanathan, S. (2002) conducted a study about sentiment classification using each of the three machine learning languages: Naïve Bayes, Maximum Entropy Classification and Support Vector Machine. The authors wanted to find out which of those three has the best accuracy when it comes to identifying a sentiment. The authors had used movie reviews from IMDb, which is a famous movie review site, as their data for their experimentation and comparison of the three algorithms and. To do so, they had first collected lists of categorized human words that was used to identify emotions and they had randomly collected 700 positive-sentiment and 700 negative-sentiment reviews.

The authors had used different methods of parsing the review statements for all each algorithm and those unigrams, bigrams, part of speech, adjectives and position. They also used an appropriate pair of parsing methods for each algorithm to deeply know the comparison. The result of the experiment shows that the Support Vector Machine performs the best while Naive Bayes performs the worst on the test. The authors said that the differences of each accuracy score are not large.

**2.2 Usual Challenges of Sentiment Analysis and Opinion Mining**

Sentiment analysis proves itself useful in a variety of real world applications. For instance, in marketing and management fields, it helps managers grasp the progress of their products and advertisements in terms of its success in fulfilling the consumers' demand or to be able to distinguish what the customers like and what they do not like. People are intended to develop a system that can identify and classify opinion or sentiment as represented in an electronic text. An accurate method for predicting sentiments could enable us, to extract opinions from the internet and predict online customer’s preferences, which could prove valuable for economic or marketing research (Vinodhini, G. & Chandrasekaran, R.M. 2012).

Plain text processing of two different statements with only small differences does not matter as much. But in the case of sentiment analysis, the statements 'We had a romantic and beautiful night in this hotel' and 'We had a more romantic and beautiful night at the other hotel' make notable contrasting declarations. Things get even more promising especially in handling highly unstructured data from informal social mediums. The more frequent users combine several opinions on several matters in the same sentence, the more difficult it is for the computer to parse everything stated as compared to using pure human intuition. The researchers of this study implies that there arises several challenges in sentiment analysis such as: An opinion perceived as positive in one perspective may be sensed as negative in another perspective, and conclusively, people always have different ways of expressing their thoughts and opinions through giving sentiments. Aside from the noticeable challenges, the paper also presented the key methods and techniques in sentiment analysis.

**2.3 Approaches for Sentiment Analysis on Twitter: A State-of-Art Study**

Harsh Thakkar and Dhiren Patel conducted a study on “Approaches for Sentiment Analysis on Twitter: A State-of-Art Study”. These days, microblogging has been the trend for everyone in the world. People nowadays spend their time sharing their thoughts, opinions and sentiments about the society, industry, their emotions etc. on social media applications like Twitter, Facebook, Instagram etc.

It said in the study that Twitter’s opinion mining is working on its post by means of three techniques. Lexical Analysis, Machine learning based analysis and Hybrid/Combined Analysis. In Lexical Analysis, they stated that it usually guided by a dictionary consisting pre-tagged lexicons. Lexicons, meaning the vocabulary of one person. Machine learning based analysis is a way of analyzing sentiments by using a machine and it usually undergoes 3 stages which are Data Collection, Training of Data and Classifications and Plotting results. Lastly, Hybrid or Combined Analysis is the combination of the two said techniques. It shows the accuracy of machine learning based analysis and the speed of Lexical Analysis.

The authors concluded that sentiment analysis really plays a big part when it comes to social media. Lexical Analysis, Machine learning based analysis and Hybrid Analysis will always vary depending on the application. But the authors said that Lexical approach is easy to use and straight to the point. They also stated that Machine learning based analysis will require a lot of work but will produce exquisite output. Lastly, they stated that Hybrid approach has displayed positive sentiments as far as performance is concerned.

**2.4 Cross-Lingual Information to the Rescue in Keyword Extraction**

Methods for extracting keywords in a bilingual context have been tested. As mentioned by the study ‘the method involves estimating keyword preferences, word-aligning parallel articles and bridging language-specific word statistics using PageRank.’ The results of conducted tests show that the method can identify more keywords and rank them higher in the candidate list than monolingual KEA’s (Keyword Extraction algorithm).

**III. Technical Background**

**3.1 Java Programming Language**

According to the study called “Comparative Studies of 10 Programming Languages within 10 Diverse Criteria” which was conducted by four people from Corcordia University (2010), Java, a high-level, computer programming language, is one of the most popular language used by programmers nowadays. Many software developers and students in the programming community have been relying upon Java because, compared to other programming languages, it has a highly understandable syntax, and it is best known for its portability and object-oriented manner. Other than that, its creator had managed it to run through any operating systems, such as Mac OS, Windows, and Linux by also developing Java virtual machines for each OS’s, and that was why it outranked all other programming languages.

Java has wide variety of usable libraries which are collections of pre-existing, open-source codes made by professional Java programmers; jsoup is one of those libraries, and it is extremely essential for extracting data from internet websites. Technically, websites are simply made of HTML (Hyper Text Markup Language) files which, aside from a programming language, is a mark-up language that only displays either texts or its other elements in a website. The programmer of the jsoup library, Jonathan Hedley (2009), had shared the documentation to the community, so that anyone would know how to properly use and fully integrate jsoup into their own software.

**3.2 Database**

A collection of data is called a Database. Basically, a Database can consist of many variables like schemas, tables, queries, reports and other objects. Based the book of Christopher J. Date (2003), in order to manage a data in a database, it will need to have a Database Management System (DBMS). DBMS can help the user interact with the database itself and it can also help them to handle and capture the data.

Normally, the data can be organized into rows, column, and tables. We can relate this to our project because first of all, we will need a database for our algorithm to work. Our project will need a set of words, sentences, phrases, and other thoughts to help us sort all the given data. We will need to have a so called “Bag of Words” that we will get from the database for our algorithm. The “Bag of Words” that we are planning to make contains the words that are either positive or negative thoughts.

**3.3 Sorting Algorithm**

According to Art Kagel (November 1985), the so called “Sorting Algorithm” is an algorithm that puts elements of a certain list in a certain order. There can be two desired outputs in the Sorting Algorithm. One can be that the output will be in a non-decreasing order, or it can be in a permutation. Since data computing has been a trend nowadays, sorting has played a big part in this topic since we need a lot of sorting in data manipulation. More likely in this algorithm, the data will be arranged in an array which allows it to random access, rather than in a list which only allows sequential access.

You can relate this to our project because we need to make an algorithm that detects positive, negative and neutral comments in surveys. Since Sorting Algorithm is related to this, this can help by arranging all the data that we can gather in a list so that we can easily identify the data. The data that we will be gathering will come from different websites that provides reviews for the hotel. It will be easier for us to arrange the data gathered if it is sorted.

**3.4 Djikstra’s Algorithm**

Based on the works of Robert B. Dial in 1956, This algorithm is for finding the shortest paths in a graph from point to its destination. This algorithm is usually presented in a form of a graph. Normally, this graph is like a tree, so many branches but only one source. The graph will always have a source node, and it will find its way to the end in the shortest way possible.

An example problem that can be related to Dijkstras’s Algorithm is when a user tries to find themselves out in a city they’re not familiar with. They will enter the starting point and the destination, then the algorithm will find the shortest path for them to get to that destination. However, this has also a weakness, its relative slowness in some topologies. If you’ll relate this to our project, it would help us by finding the data the shortest way. The proponents of the research will be gathering a lot of data and it will need to sort it all but it also needs to find the shortest and fastest way to do it.

**3.5 Support Vector Machine**

Support vector machines are supervised learning models utilized for analyzing data and for its classification and regression analysis using learning and sorting algorithms infused to it. It falls under a subfield of computer science which is known as the machine learning, which is responsible to bestowing computers their own brain and perhaps, the will to do and learn tasks without requiring too much aid from human programming. SVM’s deploy its capability to create predictions and forecasts based of the given set of data presented. Thus, its resolve falls on one main objective: to make corresponding decisions driven by data inputs from its built model. Moreover, it constructs a hyperplane or sets of hyperplanes which is used for specific tasks such as data classification, regression analysis etc.

Putting everything into technical terms, SVMs used for sentiment analysis of opinion mining is directly focused on two things: classifying and predicting. The model represents a vast dimensional plane where inputs are represented by points in that plane. Then, the samples are mapped according to the respective categories they fall into such that a clear gap would be wide enough to separate each classification. Combined with its machine learning skills, the support vector machine will be able to identify new inputs to the model and then designate each input to its rightful category just as how it processes previous ones. The application of SVM towards to real world scenarios ranges from text categorization, image segmentation, hand-written character recognition, permutation and many more to mention.

**3.6 Text Mining**

As written by Hotho, Nurnberger, and Paab (2005), Text mining aims at disclosing the concealed information by means of methods which on the one hand are able to cope with the large number of words and structures in natural language and on the other hand allow to handle vagueness, uncertainty and fuzziness. Text mining is the automated process of searching forms of texts that are related to the query of the user, and it is also responsible for the extraction of the data of the document which is known to be related to the query.

Text mining helps us to easily get any written works that are published in the web that we are eventually going to need when formulating a research or an idea. The most, common application of text mining is used mostly for search engines, such as Google and Bing. It can also be applied for filtering spams in online mails. Studying text mining will be a great help for understanding how texts from the web are being extracted, and it will be important for the creation of software projects that have a feature of text extraction from other resources.

**3.7 Understanding How Bag-of-Words Works**

Bag of words refers to the framework model of words. In programming, it contains strings which refer to as either words or numbers that cannot be computed. Dictionary, for instance, is one of the examples of bag of words for us humans, but the only difference is that dictionary contains tons of words that come with different descriptions, whereas the bag-of-words model contains only the words. Bag of words is simply a storage for words that are going to be relevant for a certain program. Bag of words uses algorithms to be able to detect word occurrences in a specified context or chunk of words. Using bag of words as a method to identify word frequencies can be applied, for instance, in assessing reviews. For example, the word “bad” is the most frequently used word in the bag of words model made for hotel and restaurants feedback system.

As said by Masumoto, Takamura and Okumura (2005), not only a bag-of-words but also word order and syntactic relations between words in a sentence are intuitively important and useful for sentiment classification. Thus, there appears to remain considerable room for improvement by incorporating such information. Knowing that it occurred as much as any other word among the feedback collected form the customers, it can be easily inferred that there is something wrong with their service and the word “bad” justifies the conclusion made.

**3.8 Natural Language Processing**

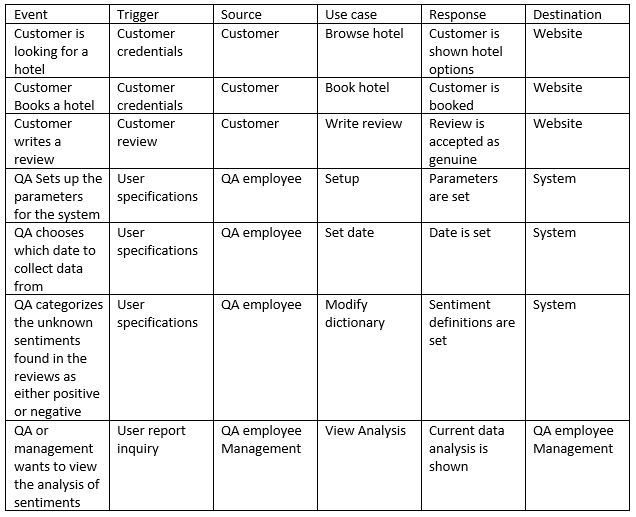
According to techtarget.com, NLP is the ability of a computer program to understand human speech as it is spoken. It is used to enable computer programs that are focused on data gathering to be able to comprehend ambiguous human speech and be converted to usable data. This can be found in web scraping tools -tools or programs that are designed to gather large amounts of data from websites- that are used to obtain data from social networking sites as the language used there is human.

It is considered to be as a component of artificial intelligence. Since the human language is often read within context an NLP program must have the ability to learn. The adaptability of a NLP capable program can be seen in its accuracy from before and after it has been given several sample data’s as references. Not only does NLP allow a program to understand vague human speech, it is also enabling it to learn and improve its understanding of human speech making it a self-learning program.

**I.V. Design and Methodology**

**4.1 Requirement Analysis**

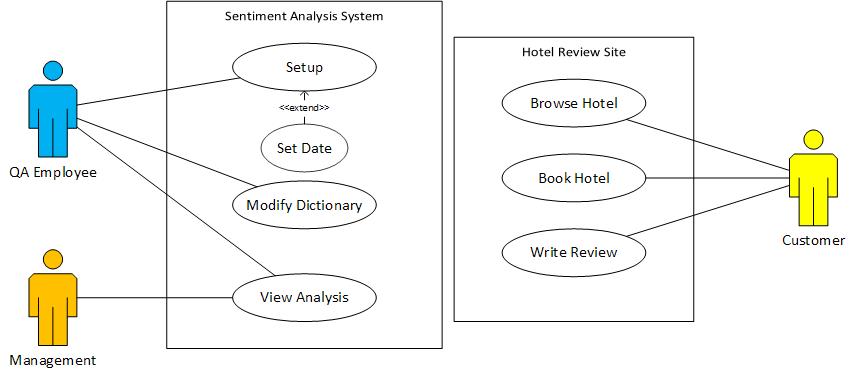
**4.1.1 Event Table**

****

**4.1.1 Context Diagram**

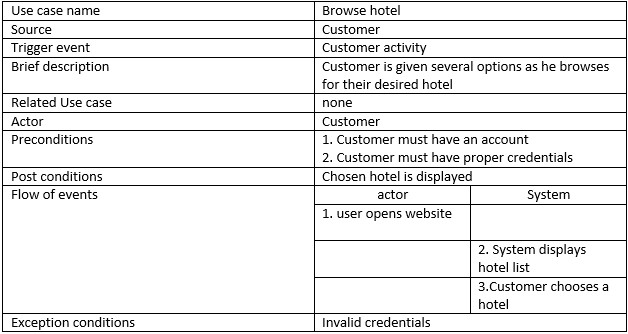
**4.2 Requirements Documentation**

**4.2.1 Use Case Diagram**

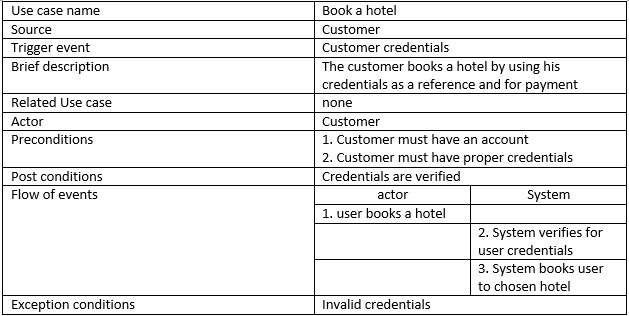
****

**4.2.2 Use Case Description**

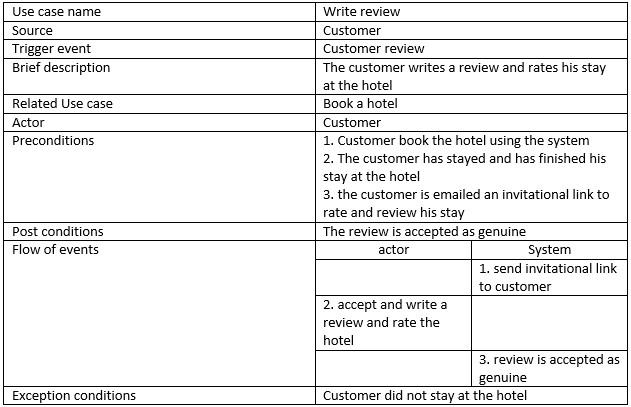
**4.2.2.1 Browse Hotel**

****

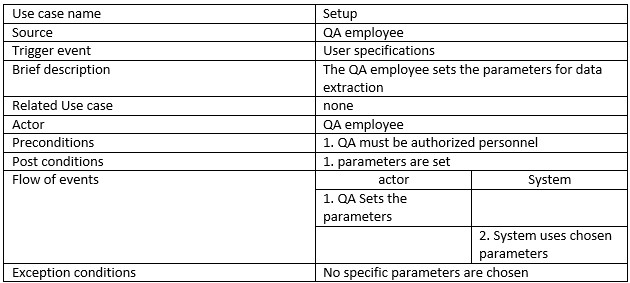
**4.2.2.2 Book a Hotel**

****

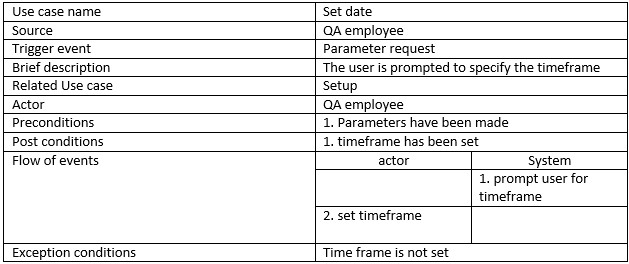
**4.2.2.3 Write Review**

****

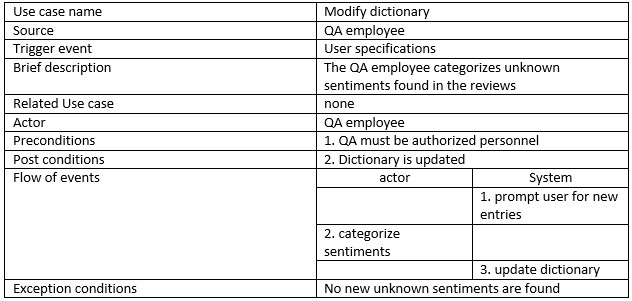
**4.2.2.4 Setup**

****

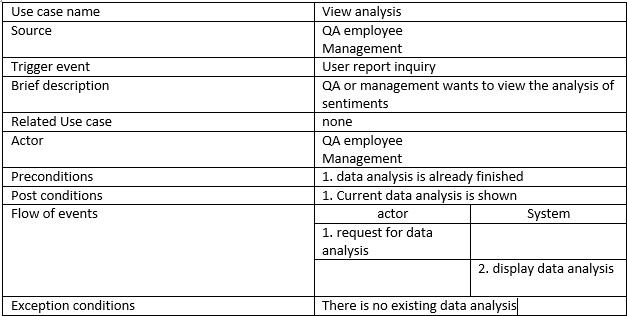
**4.2.2.5 Set Date**

****

**4.2.2.6 Modify Dictionary**

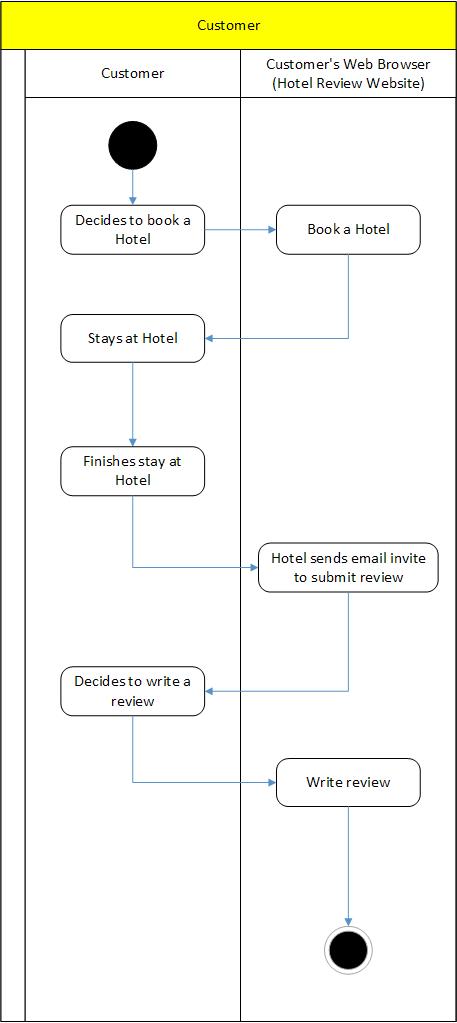
****

**4.2.2.7 View Analysis**

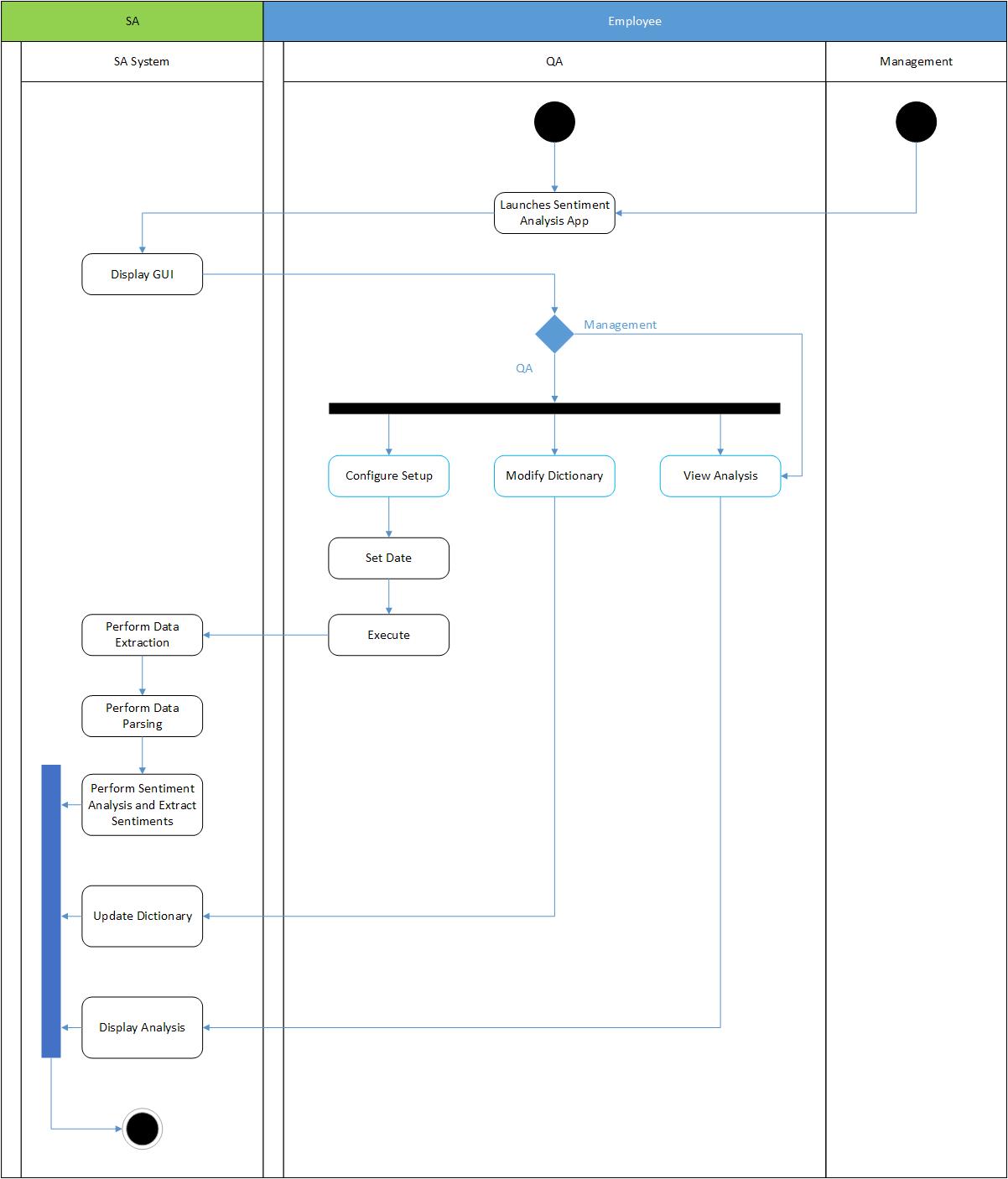
****

**4.2.3 Activity Diagram**

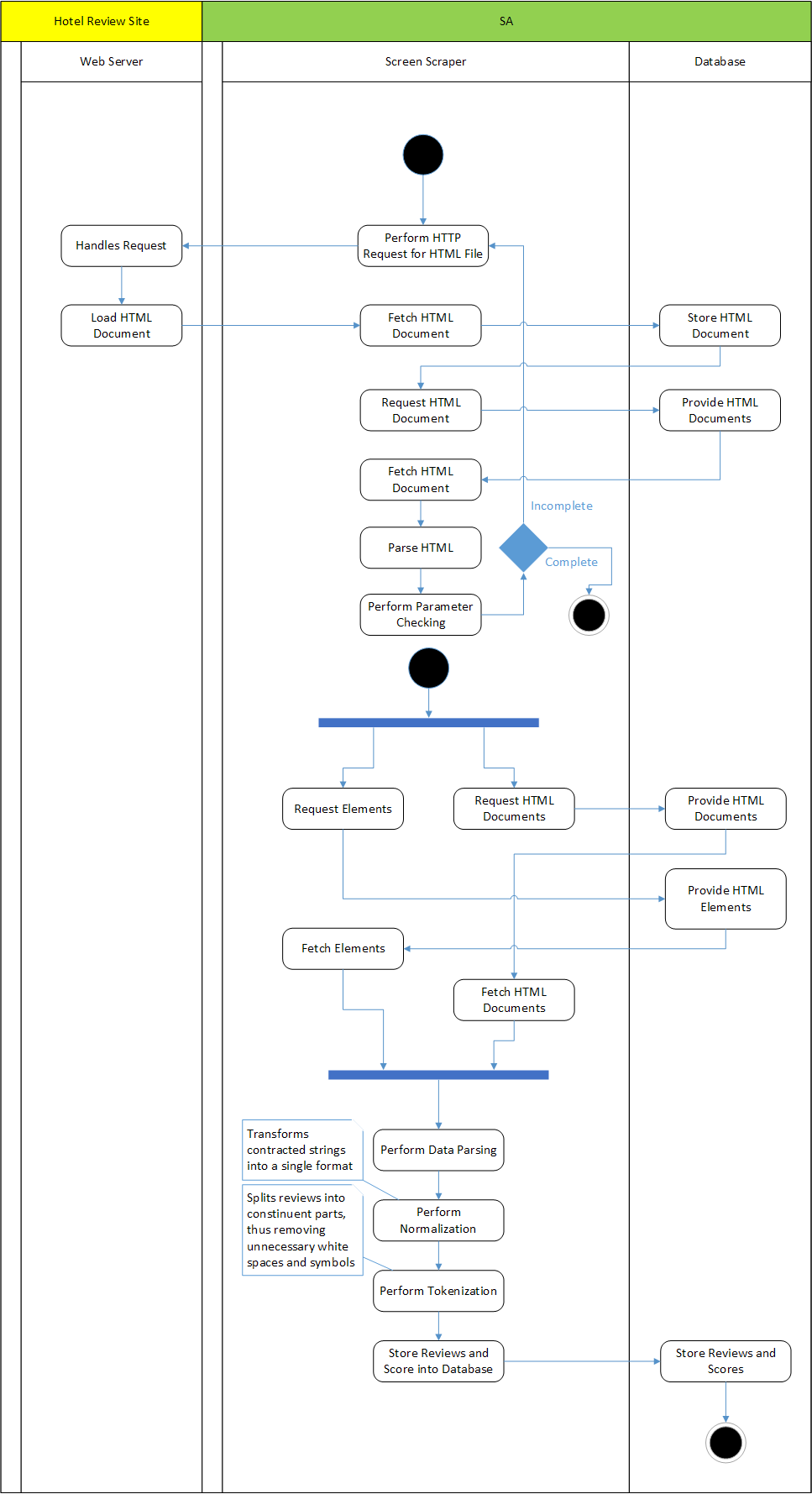
**4.2.3.1**

****

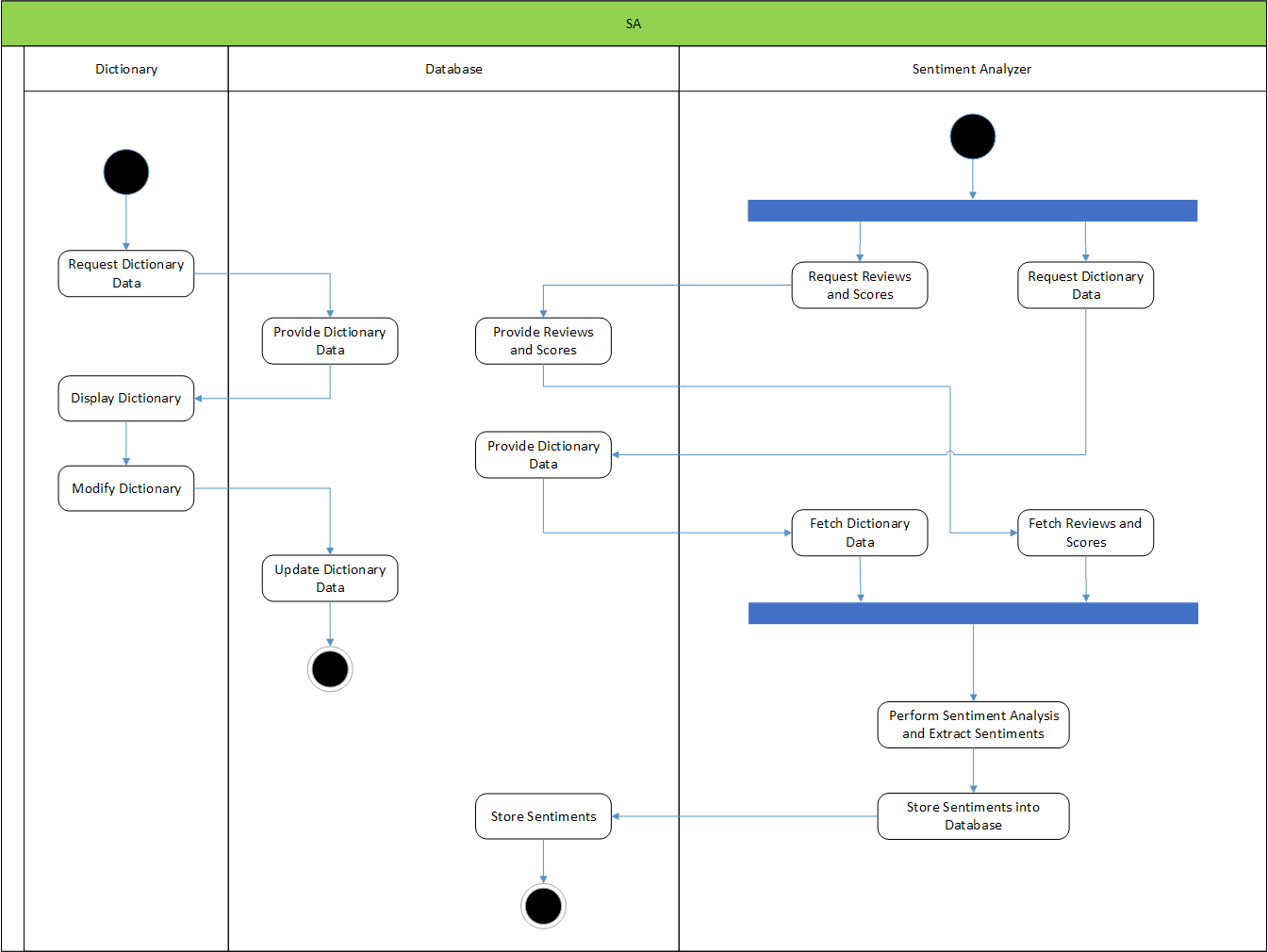
**4.2.3.2**

****

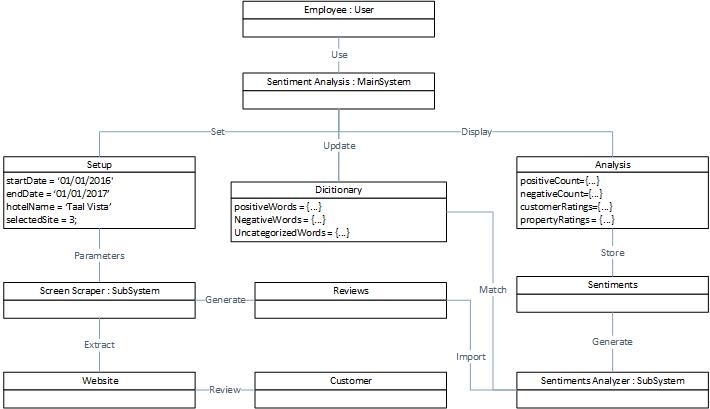
**4.2.3.3**

****

**4.2.3.4**

****

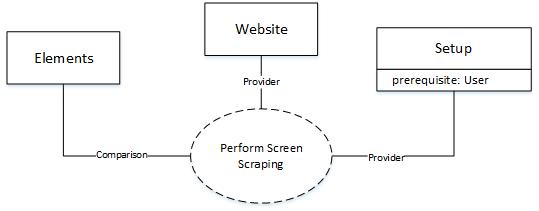
**4.2.4 Object Diagram**

****

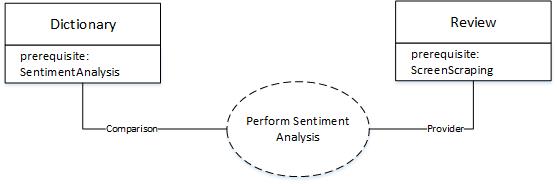
**4.2.5 Class Diagram**

**4.2.6 Composite Diagram**

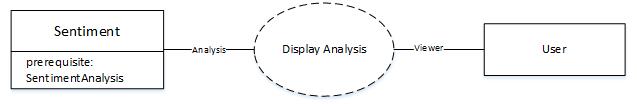
**4.2.6.1**

****

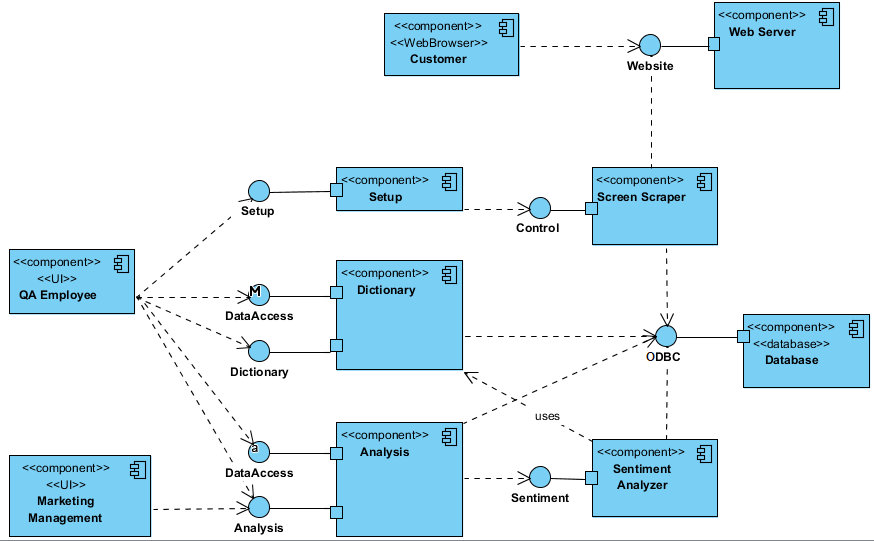
**4.2.6.2**

****

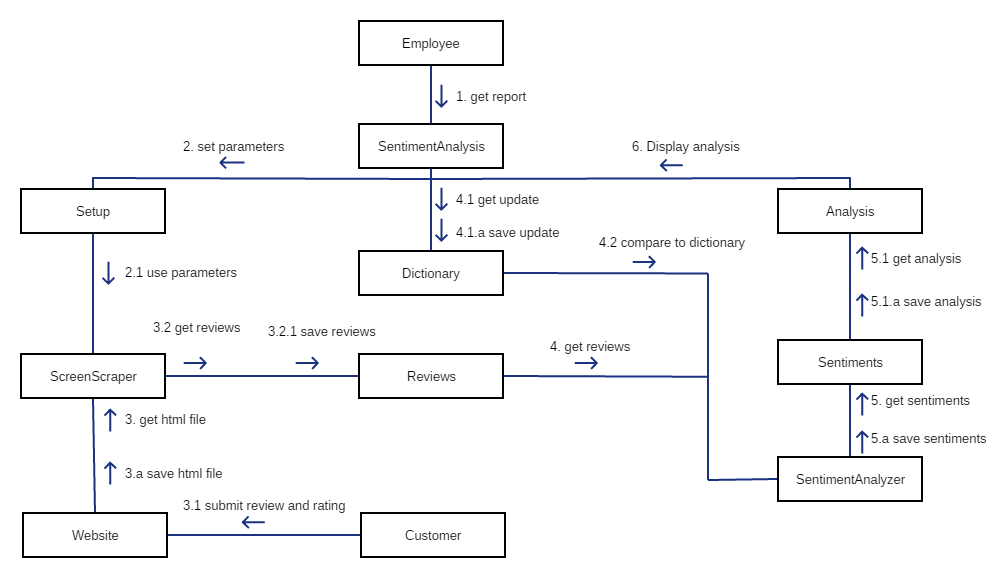
**4.2.6.3**

****

**4.2.7 Component Diagram**

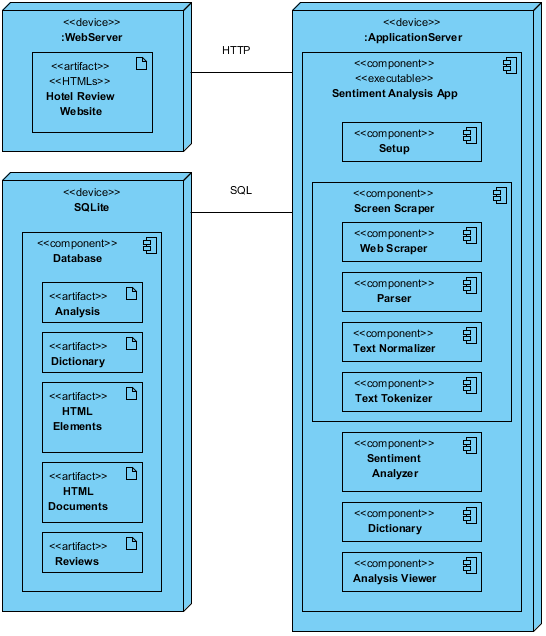
****

**4.2.8 Communication Diagram**

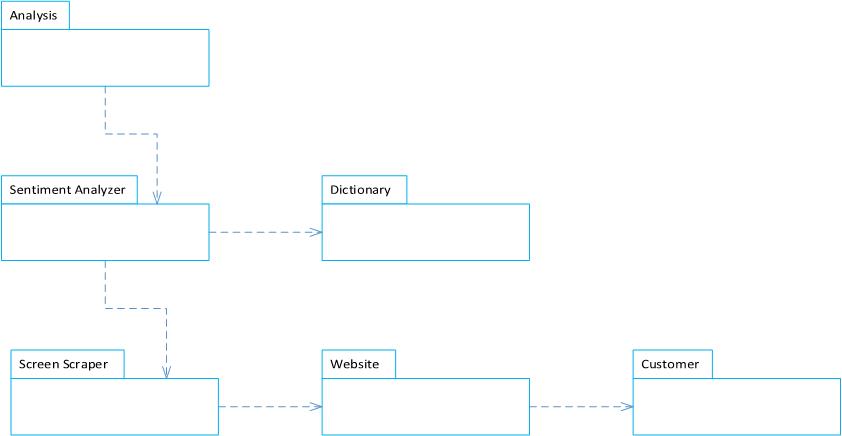
****

**4.2.9 State Diagram**

**4.2.10 Deployment Diagram**

****

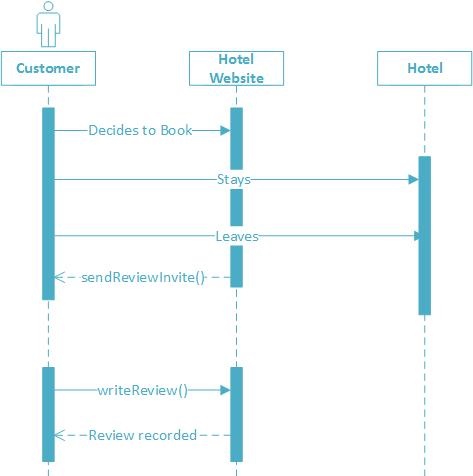
**4.2.11 Package Diagram**

****

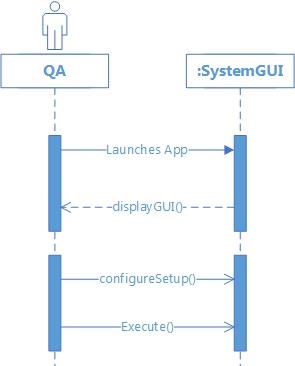
**4.2.12 Timing Diagram**

**4.2.13 Sequence Diagram**

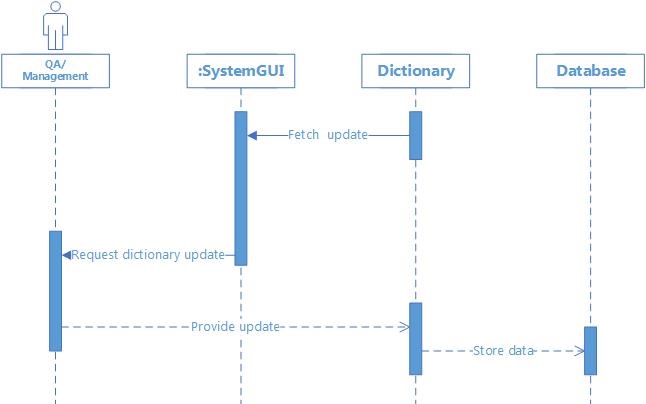
**4.2.13.1 Customer**

****

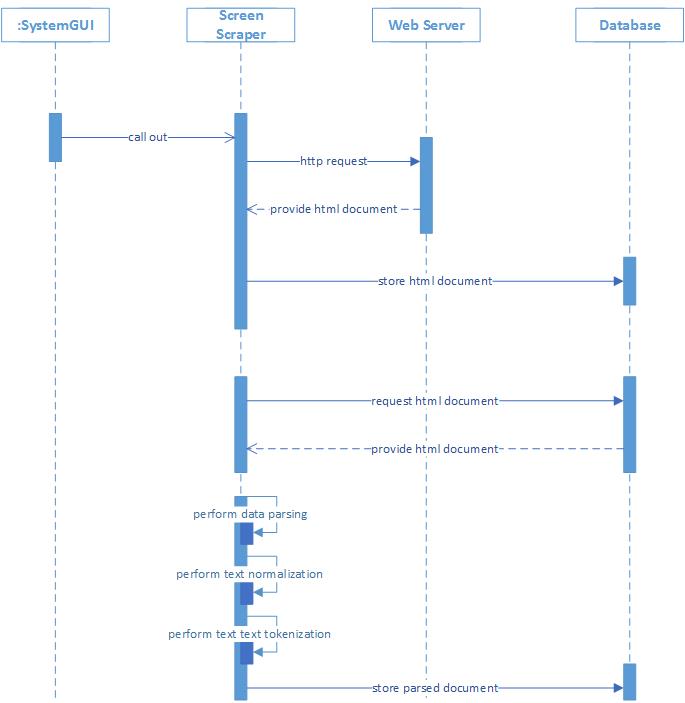
**4.2.13.2 Setup**

****

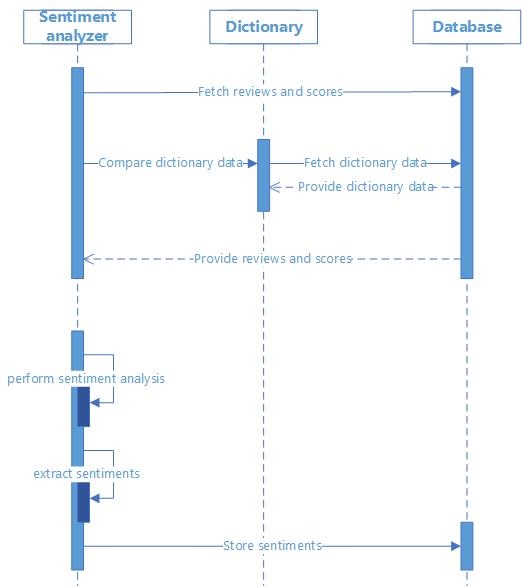
**4.2.13.3 Update Dictionary**

****

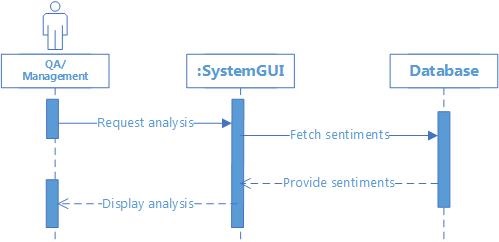
**4.2.13.4 Screen Scraper**

****

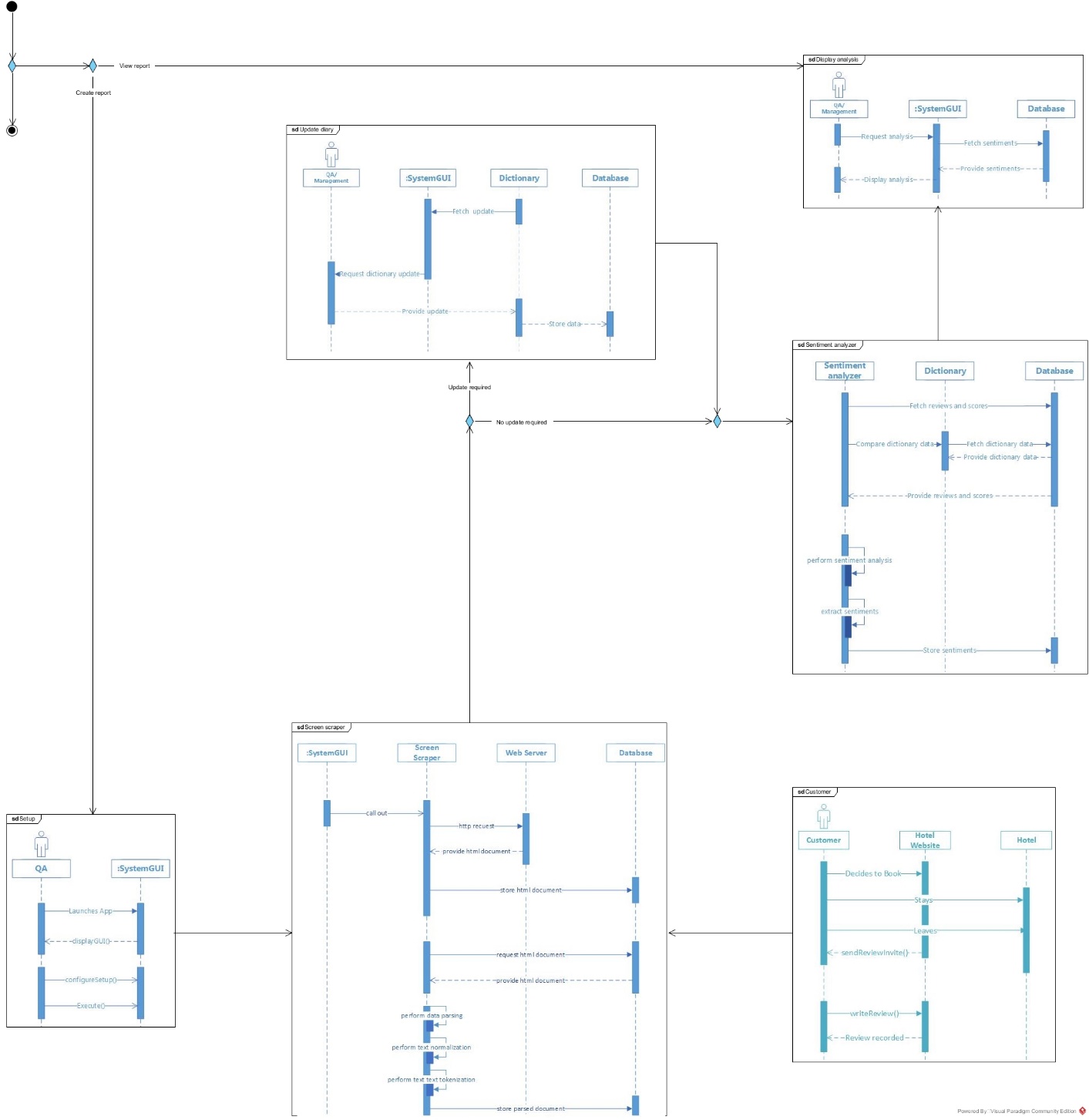
**4.2.13.5 Sentiment Analyzer**

****

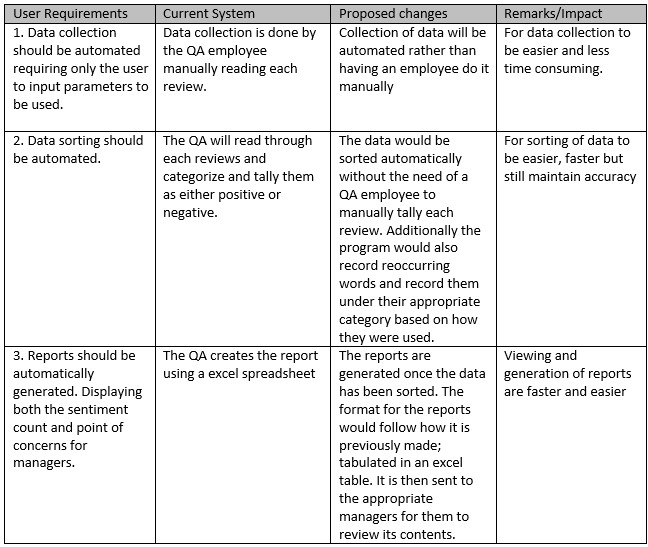
**4.2.13.6 View Analysis**

****

**4.2.14 Interaction Overview Diagram**

****

**4.3 Gap Analysis**



**4.4 Design of Software, Systems, Product, and/or Processes**

The system is designed to automatically collect data from several review sites that features S.M. hotels. The collected data is then filtered and prepped for sentiment analysis. During which the reviews from various customers would be identified as either a positive or negative feedback by matching a given set of words to the ones used in the reviews. In the event that a new word has been found the user (QA) is then prompted to categorize them and update the dictionary of the system. The sentiments are then tallied as well as reoccurring words found and used to generate a report file which is stored in the database. These reports can then be viewed and sent to the respective management for points of concern.

**4.5 Development and Testing, where applicable**

The prototype was created using Python 3.

The proponents plan for the expected users (Sales Department) to test the prototype. The proponents will populate the prototype with the company’s data from 2014-2016. These data will be used to generate reports and forecasts for the first 6 months of the year 2016. The users will then validate if the values forecasted by the system match the actual data found in their existing data for 2017.

**4.6 Description of Prototype**

The prototype was created using Python 3. The prototype uses a WebDriver from Selenium to perform automatic web data extraction. The data needed come from the three hotel websites which contain the customers’ reviews that are essential for the sentiment analysis. During the data extraction, the data would undergo text parsing to clean up the data and to determine the needed elements, and those would then be consecutively stored into the SQLite database. After storing the data, the prototype would now perform a simple, sentiment analysis. The sentiment analyzer has two dictionaries; each contains a list of words that is classified as positive or negative; the analyzer uses these dictionaries to determine whether a review is positive or negative.

**V. Appendices**

**5.1 References**

Black Jr, James E., and Uri Zernik. "Method for natural language data processing using morphological and part-of-speech information." U.S. Patent No. 5,331,556. 19 Jul. 1994.

Church, K. W. & Mercer, R. L. (1993). Introduction to the special issue on computational linguistics using large corpora. MIT Press Cambridge, MA, USA

Date, C. J. (2008). An introduction to database systems. Boston, MA: Pearson Addison Wesley.

Dial, Robert B. (1969). Algorithm 360: Shortest-path forest with topological ordering. Communications of the ACM: 632–633. doi:10.1145/363269.363610

Hedley, J. (n.d.). Jsoup: Java HTML Parser. Retrieved from <https://jsoup.org/>

Hotho, A., Nurnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining (Unpublished master's thesis). University of Kassel.

Kagel, Art (November 1985). Unshuffle, Not Quite a Sort. Computer Language.

Li, J., Liu, M., Rabah, S., & Lai, Y. (2010). Comparative Studies of 10 Programming Languages within 10 Diverse Criteria. University, Montreal: Concordia.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classiﬁcation using Machine Learning Techniques (Master's thesis, Department of Computer Science Cornell University Ithaca, NY 14853 USA, 2002) (pp. 1-8). Proceedings of EMNLP.

Prato, S. (2017, January 20). What is Text Mining? Retrieved April 01, 2017, from <https://ischool.syr.edu/infospace/2013/04/23/what-is-text-mining/>

Sentiment Analysis: How Does It Work? Why Should We Use It? (2016, August 16). Retrieved April 01, 2017, from <https://www.brandwatch.com/blog/understanding-sentiment-analysis/>

Sierra, K., & Bates, B. (2008). Head first Java: Beijing: O'Reilly.

Thakkar, H., & Patel, D. (n.d.). Approaches for Sentiment Analysis on Twitter: A State-of-Art study (Unpublished master's thesis). Department of Computer Engineering, National , Institute of Technology, Surat-395007, India.

Vinodhini, G. & Chandrasekaran, R.M. (2012) Sentiment Analysis and Opinion Mining: A Survey (Research paper, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar-608002 India, 2012) (Vol.2, Issue 6).