Methodology

# **Introduction**

Software development methodologies play a vital role in developing a system software. The basic purpose of these methodologies is to provide smooth software development according to the project requirements. Software development methodology ca be defined as a framework that is used to structure, plan, and control the process of developing an information system. This chapter look into the methodologies chosen for the research, development and implementation of this project.

# **Research Methodologies.**

Research is a careful consideration of the study using scientific methods for a particular concern or problem. According to the American Sociologist Earl Robert Babbie, “Research is a systematic inquiry to describe, explain, predict, and control the observed phenomenon. Research involves inductive and deductive methods.” Methodologies are methods method used in a common area of study or operation. Combination of these two words forms research methodologies that is defined as the basic methods or techniques used to define, pick, process and analyse information about a library sentiment.

There are two basic categories of research methods in software development and they are qualitative and quantitative. They are applicable to numerical and non-numerical projects. Quantitative research methods support experiments and testing by measuring variables to verify or falsify theories and hypothesis must be evaluable and answerable. The method requires large datasets and use statistics to test the hypothesis and make the research system valid. The qualitative research method is concerned with understanding meanings, opinions and behaviours to reach tentative hypotheses and theories or develop computer systems, artefacts and inventions. The qualitative model mainly uses smaller datasets that are sufficient enough to reach reliable results, where the data collection continues until saturation is reached (Ghauri nad Gronhaug, 2010).

Research methodologies define what activities constitutes research, how to proceed, how to measure and what constitutes the success of a project. There are generally five types of these methodologies which are Formal, Experimental, Build, Model and Process methodologies.

* 1. **Formal Methodologies.**

This type of methodology is used to prove facts about algorithms and systems. These formal methodologies are very popularly applied in theoretical computer science. In this project there is use of Support Vector Machine hence we can employ the use of the formal methodology to assess the algorithm. Formal methodologies will expose the complexity and strength of the SVM algorithm that will be applied in sentiment analysis of libraries. This helps determine the overall complexity of the operations in the system.

* 1. **Build Methodology.**

Build methodology comprise much of building an artefact, either a software system or a physical artefact that will be used to demonstrate that it is possible to develop. In the situation or case of this project a build methodology that will be used must demonstrate that the software system can be built. To be considered a research, the construction that will be undertaken must demonstrate new artefacts features that haven’t been demonstrated in any other artefacts before. Whenever a research question leads to the building of a software system, there is a certain of good practices that should be considered which include designing the software, component reuse, choosing an adequate programing languages and testing the system all the time during the development phase.

* 1. **Experiment Methodology.**

Experiment methodologies are used much more in the computer science field due to their ability to evaluate new solutions. Once there is a new solution there is need for it to be thoroughly evaluated to see the impacts of the new system and also if it’s feasible. The favouritism of a project is much based on its feasibility, that is the cost and the benefits it brings to the organisation as a whole.

Experimental evaluation are divide into two phases normally. The phases are exploratory phase and the evaluation phase. The earlier mentioned phase –exploratory phase\_ the researcher undertakes some measures that will help in identifying the possible questions that may be asked about the software as a whole. The later phase that is the evaluation phase will attempt to answer these questions identified in the exploratory phases. For an experiment to be labelled as well designed it has to start with a well-structured list of questions that the experiment must answer. The main aim of experimental methodologies is to show the experiments that will occur with aim of extracting results from the real world implementations and they can be used to test a variety of theories.

1. **Software Development Methodology.**

During literature review, several types of software development methodologies identified which can be implemented in the sentiment analysis system to be developed. The literature has reviewed a number of methods that were used during different sentiment analysis researches and projects. To select the most suitable methodology, a proper analysis has to be done in the context of this project. Modern day methodologies are structured to minimise risk by developing software in short time boxes referred to as iterations which typically last one to four weeks. The iterations are sub-projects which include all the tasks necessary to release the increment of new functionality or for a different module of the system that exerts different actions to the final solution.

There are factors that were considered for selecting the suitable methodologies. The major factors in the context of this project’s set up include the following:

1. **Complexity of the system**.

The complexity of the project doesn’t allow or doesn’t work very well with traditional methodologies. Traditional methodologies are good at managing management system where there is really need for requirements elicitation. Traditional methods can’t be used in sentiment analysis (machine learning).

1. **System Objectives.**

The objectives of the project are not yet certainly specified, there is room of changing the objectives of the system in due course. The project goals and objectives were specified with a single human being (the student) that leaves the objectives at risk of being changed as time moves on after analysis from the project supervisor.

1. **The availability of expertise.**

The only expertise at hand are two humans being a naïve programmer (Student) and management (supervisor). The methodology that will be used must take note of the above specifications.

1. **Cost involved in the project.**

The project has no budget hence we would require the methodology that is not costly to implement

1. **Risks associated with the project.**

Project failure will result in the student obtaining a poor grade or even failing but apart from that no other risks are associated with the project. A methodology that reduces the extent of risks can be used so that the student can get a good grade for the project.

1. **The availability of resources.**

In the student’s hands there is a single laptop and a cell phone that will be the device to use during project development hence methodologies that employ the use of other tools are neglected.

* 1. **CRSIP DM**

CRISP DM stands for Cross-Industry Standard Process for Data Mining. CRISP DM is a methodology that is applied much in data mining but it is considered to be a standard methodology applied to the extraction of knowledge from big data sets (Sharda, Delen and Turban, 2016). Sentiment analysis works with datasets for training of systems either training the system using supervised learning algorithms or unsupervised learning algorithm. Due to the its ability to work with datasets, CRISP-DM was used by Nave, Rita and Guerreiro, (2018)

CRISP-DM has got six steps that are business understanding, data understanding, data preparation, modelling, evaluation and deployment (Chapman et al., 2000).

**Project Understanding**

**Evaluation**

**Final Presentation**

**Modelling**

**Dataset Understanding**

**Data**

**Data Preparation And Cleaning**

1. **Business Understanding (Understanding project goal)**

The stages start with understanding the business goals. There is need to understand what the project is exactly about and knowing the expected outcome of the project. After understanding the objectives of the project then the knowledge in transferred into machine learning objectives.

Situation assessment is also done at this stage to see the risk that may arise and providing contingency plan to the risks. Assumptions that will be used in the project are gathered and some constraints of project success are noted down

Development of a project plan is done at this stage this will lay the root that will be followed as the project proceed.

1. **Data Understanding**

This phases starts with data collection that is searching for possible datasets from reliable datasets. Analysis of the necessary procedures that are needed to have the datasets. Identification of data quality problems that will help to see if the data will help meet the project objectives or not basing on the field in which sentiment analysis is been applied.

After this process initial data collection report, data description report, data exploration report and data quality reports are expected to be produced.

1. **Data Preparation**

After the initial data was collected there in need now to produce a perfect data set that will be used in sentient analysis. The first step is to describe the dataset this gives a view to project members what the data set collected about and what it is about to be used. After description the data is selected and the cleaned to remove unnecessary data in the dataset. Then finally the data is constructed, integrated and formatted.

At the end of the phase there is data cleaning report, and formatted data.

1. **Modelling**

In this phase various techniques are selected and then the appropriate one is selected. The modelling assumptions are stated here in a try to perfect the environment and to create a ground that will reduce criticism of the model. The model test plan is set aside and the model is built. After the building process is done assessment is done to measure the success of the model.

1. **Evaluation**

The model or models produced are thoroughly evaluate to see if they went through the correctly stated procedures and see if all the assumptions stated during modelling were properly followed.

1. **Deployment**

The model is now functioning. The model is then presented to the intended guest or customers. The customers are taught how the model works.

After that the project is placed under monitoring to see how it will continue functioning.

* 1. **KDD**

The KDD process, as presented in (Fayyad et al, 1996) is the process of using DM methods to extractwhat is deemed knowledge according to the specification of measures and thresholds, using a database alongwith any required pre-processing, sub sampling, and transformation of the database. There are considered fivestages, presented below

1. **Selection**

This stage consists on creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.

1. **Pre-processing**

This stage consists on the target data cleaning and pre-processing in order to obtain consistent data.

1. **Transformation**

This stage consists on the transformation of the data using dimensionality reduction or transformation methods.

1. **Data Mining**

This stage consists on the searching for patterns of interest in a particular representational form, depending on the data mining objective (usually, prediction)

1. **Interpretation/Evaluation**

This stage consists on the interpretation and evaluation of the mined patterns.

* 1. **SEMMA**

SEMMA methodology is a methodology that stands for Simple, Explore, Modify Model and Assess. This methodology is more of a data mining methodology. Its help it’s much in data mining applications development. It can be said it’s a logical organisation tool set by SAS Enterprise Miner with the intentions of undertaking core data mining tasks. This methodology was developed with much focus into data mining application (Enterprise Miner) and other methodologies (such as CRISP DM) where developed to suit a wide range of data mining applications. SEMMA has got five stages that also makes up the abbreviation. The stages are Simple, Modify, Model and Assess

SEMMA stages.

1. **Sample**.

The first phase of the SEMMA methodology is Sample. This phase is all about data sampling. This include processes like selecting the data set for modelling. The selection of a data set is a crucial process that must be treated with higher priority. The dataset that will be selected must be enough or big to provide sufficient information needed and small to be usable efficiently. Data partition functionality is dealt with at this phase of the SEMMA methodology.

1. **Explore**.

This phase focuses much on data understanding. By understanding data there is need to discover the unanticipated and anticipated relationships between the variables and their abnormalities. To understand this there is really need for data visualization.

1. **Modify**.

The selection, creation and transformation of variables in preparation for data modelling are done at this stage.

1. **Model**.

The phase where there is vast application of various different modelling techniques. The techniques are applied on the variables prepared in the modify phase. All this is done so as to create models that provide the desired software outcome.

1. **Assess**.

The last phase is of the SEMMA is Assess. The results of the process will be finally out and then there is need to assess them and see if they have met the objective.

The **Sample** and **Explore** stages of SEMMA they have got some sort of similarity with Data Understanding phase of CRISP-DM. **Modify** can be translated to the Data Preparation phase. **Model** likewise is the Modelling phase. **Assess** is parallel to the Evaluation phase of CRISP-DM. All these models are expected to go in a cyclical way rather that linearly.

1. **Chosen Development Methodology.**

For research purposes the Build methodology was adapted. The reason behind the selection of this methodology is much based on its ability to cater and support system design, selection of programming languages which makes it flexible for code reuse in the event of upgrading the system itself. The other feature of Build methodology that called for its selection is its ability and mechanism of continually testing of system during development other than testing only after the project is completed. The system that will be built need continues testing because of the little know how of the programmer. The type of dataset that will be used need some thorough cleaning and to see if the dataset is clean and meeting the required threshold there is need for continues testing.

For software development methodology the CRISP DM methodology is the one that is going to be used. The selection of this methodology is much based on the steps of the methodology that are followed. The steps of the methodology suits the type of project that is being undertaken with little changing efforts. Other that other methodologies that are much in Data Science projects.

1. **Development Tools.**

The project will be based much on python programming language for the back end. The front end will be based on python and some Gibrish. The text editor that will be of use is Visual Studio Code text editor. Django will be the frame work of choice for the full development of the system.

* 1. Python

Python is an interpreted, object oriented high level programming language that has semantic dynamic semantics. This programming language. This programming language is very easy to learn and easier to read code. This brings an advantage of easy to maintain code. It also has garbage collection features that reduces the labour of the programmer during coding. The programmer doesn't have to worry about when to and when not to release no longer being used memory.

* 1. **Visual Studio Code**

Visual Studio Code (VS Code) this is a source code editor or text editor developed combinatorial by macOS and Windows. It supports code [debugging](https://en.wikipedia.org/wiki/Debugging) and syntax highlighting. It also has embedded [Git](https://en.wikipedia.org/wiki/Git) control and [GitHub](https://en.wikipedia.org/wiki/GitHub). Moreover it has intelligent code completion, refactoring of code and snippets. It is customizable, enabling users to set the themes, keyboard shortcuts, preferences, and install additional feature extensions. All the functionalities and specialities of VS Code make it suit for these project.

* 1. **Django**

Django is a Python Web platform at the highest level that facilitates rapid development and clean, practical architecture. Created by seasoned developers, it takes care of much of the web development problem, so you can focus on writing your app without reinventing the wheel. It's open source and free.

1. **Conclusion.**

The success of every project is much of in the hands of the methodology that is chosen. The methodology lays a way that the project team (no matter the size) must follow. Meeting of deadlines, coordination between team members and meeting objectives they all rely on the methodology of choice. This chapter has culminated in the choice of the methodologies that suits this project that is CRISP DM and Build Methodology. Methodologies give a guide to how the project as a whole is going to be undertaken. This methodology determines the success of the proposed system. The methodology adopted impacts directly the cost, the quality and time that will be taken to develop the project as a whole, this implies a direct influence on the success of the solution. A strong correlation is required between the research and development methodology as both are tools used to successfully create a solution.

**References**

[Nave, M., Rita, P. and Guerreiro, J. (2018) ‘A decision support system framework to track consumer sentiments in social media’,](Nave, M., Rita, P. and Guerreiro, J. (2018) 'A decision support system framework to track consumer sentiments in social media', Journal of Hospitality Marketing and Management. Routledge, 27(6), pp. 693–710. doi: 10.1080/19368623.2018.1435327.Sharda, R., Delen, D. and Turban, E. (2016) 'Business intelligence, analytics, and data science: a managerial perspective'. Available at: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Business+intelligence%2C+analytics+and+data+science%3A+A+managerial+perspective+%284%2FE%29.+Pearson.&btnG= (Accessed: 19 November 2019).) *[Journal of Hospitality Marketing and Management](Nave, M., Rita, P. and Guerreiro, J. (2018) 'A decision support system framework to track consumer sentiments in social media', Journal of Hospitality Marketing and Management. Routledge, 27(6), pp. 693–710. doi: 10.1080/19368623.2018.1435327.Sharda, R., Delen, D. and Turban, E. (2016) 'Business intelligence, analytics, and data science: a managerial perspective'. Available at: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Business+intelligence%2C+analytics+and+data+science%3A+A+managerial+perspective+%284%2FE%29.+Pearson.&btnG= (Accessed: 19 November 2019).)*[. Routledge, 27(6), pp. 693–710. doi: 10.1080/19368623.2018.1435327.](Nave, M., Rita, P. and Guerreiro, J. (2018) 'A decision support system framework to track consumer sentiments in social media', Journal of Hospitality Marketing and Management. Routledge, 27(6), pp. 693–710. doi: 10.1080/19368623.2018.1435327.Sharda, R., Delen, D. and Turban, E. (2016) 'Business intelligence, analytics, and data science: a managerial perspective'. Available at: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Business+intelligence%2C+analytics+and+data+science%3A+A+managerial+perspective+%284%2FE%29.+Pearson.&btnG= (Accessed: 19 November 2019).)

[Sharda, R., Delen, D. and Turban, E. (2016) ‘Business intelligence, analytics, and data science: a managerial perspective’. Available at: https://scholar.google.com/scholar?hl=en&as\_sdt=0%2C5&q=Business+intelligence%2C+analytics+and+data+science%3A+A+managerial+perspective+%284%2FE%29.+Pearson.&btnG= (Accessed: 19 November 2019).](Nave, M., Rita, P. and Guerreiro, J. (2018) 'A decision support system framework to track consumer sentiments in social media', Journal of Hospitality Marketing and Management. Routledge, 27(6), pp. 693–710. doi: 10.1080/19368623.2018.1435327.Sharda, R., Delen, D. and Turban, E. (2016) 'Business intelligence, analytics, and data science: a managerial perspective'. Available at: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Business+intelligence%2C+analytics+and+data+science%3A+A+managerial+perspective+%284%2FE%29.+Pearson.&btnG= (Accessed: 19 November 2019).)