Informatics Institute of Technology

In Collaboration With

UNIVERSITY OF WESTMINSTER, UK

Reviewly

Applying sentiment analysis for Sinhala-English mixed reviews

A Project Proposal by

Mr. Sashminda Iranga Withanage

w1790117 / 2019586

Supervised by

Mr. Deshan Sumanathilaka

November 2022

Submitted in partial fulfilment of the requirements for the

BSc (Hons) Software Engineering degree at the University of Westminster.

Table of Contents

List of Figures

List of Tables

Acronyms

ANN Artificial Neural Network. 1

GPU Graphics Processing Unit. 17, 20
GUI Graphical User Interface. 10

LSTM Long Short Term Memeory. 4–6

ME Maximum Entropy. 4
ML Machine Learning. 7, 10, 18, 19

NB Naive Bayes. 4, 6
NLP Natural Language Processing. i, 1–3, 7, 8, 12, 13, 18–20

RF Random Forest. 4, 6

SSADM Structured Systems Analysis And Design Method. 12
SVM Support Vector Machine. 4, 6

TF-IDF term frequency–inverse document frequency. 4

UI User Interface. 18

1 INTRODUCTION

In this research project, the author tries to apply current trends in Natural Language Processing (NLP) specifically sentiment analysis in a Sinhala-English mixed environment. Although review analysis systems are widely available all of them are mainly focused on the English language. When it comes to Sinhala English mixed reviews such as customer reviews found on online shopping sites such as Daraz and Kapruka there is no distinct system to analyse them. The proposed system will provide the user with a summary of available reviews of a product unless otherwise which will be needed to do manually.

This proposal defines the problem domain, research gap, research contributions and research questions. Research objectives are mapped accordingly to the learning outcomes and methodology is explained alongside with the deliverables expected.

1.1 Problem domain

1.1.1 Natural language processing

Natural language processing can be taken as a secondary field in computer linguistics. NLP is mainly relied on data and computations such as machine learning, probability and statistics (Otter, Medina, and Kalita, 2021). With the help of the recent improvements in computational power, NLP has leveraged the power of modern Artificial Neural Network (ANN)'s (ibid.). Deep learning models have been popular in completing NLP tasks over other methods. It is proven that a simple deep learning network will outperform almost all state-of-the-art approaches in NLP (Young et al., 2017).

1.1.2 Sentiment analysis

Sentiment analysis is often performed on textual data to get various outputs such as customer needs, help businesses to monitor brands and to summarise customer feedback. Applying deep neural network models to sentiment analysis have proven to be successful in the past few years (Basiri et al., 2021).

There are multiple steps when performing sentiment analysis in a set of textual data. These steps include tokenization , stop word removal, stemming etc. For example when data cleaning it is important to be careful with prefixes. A simple ”NOT” will convey the opposite of a meaning (Chauhan et al., 2017). The above mentioned data cleaning methods still applies to Sinhala-English mixed code data. Special attention should be given when applying deep learning techniques to low resource languages such as Sinhala. Multiple research has proven that translation models and language detection models tend to struggle when faced with code mixed data (Tennage et al., 2018). It will be challenging to apply these concepts in a Sinhala-Snglish mixed text environment.

1.1.3 Review analysis systems

There are plenty of review analysis systems done using Ebay, Amazon related datasets. A recent research was done on Amazon mobile dataset (Alqahtani, 2021) had used multiple machine learning models as well as pre-trained models. Although there were many studies done on customer review analysis there was no available research for Sinhala-English mixed review analysis.

Only a handful of research is available related to Sinhala language. According to a survey done on publicly available Sinhala NLP resources shows that there are four currently available resources for Sinhala-English mixed scenarios (Silva, 2019). Most of the recent research on Sinhala language are done on hate speech detection on social media and language identification. Similar research done on other languages such as Hindi-English can be also leveraged here. A research which was done in 2015 on Hindi-English code mixed data had taken a dictionary based approach to correct misspelled words (Singh and Wassan, 2015). This approach should be evaluated further whether it can be applied in Sinhala-English code mixed data too.

1.2 Problem definition

sentiment analysis has improved a lot with the help of the research done in the past few years. Although it has improved a lot for a low resource language such as Sinhala it still can be challenging. Currently there is no model or system to analyse sentences which contain Sinhala-English mixed text (Shanmugalingam, 2019). There are only handful of research done in the area of Sinhala-English code mixing.

As a occurrence of above scenario it is not possible to analyse the reviews left by customers in online shopping sites such as Daraz, Kapruka or Takas since they contain both Sinhala, English and both mixed comments. Making decisions based of currently available reviews can be difficult task for a monolingual user. Even for a bilingual customer it can be tedious to read all available comments and make a decision based off them. By automating analyzing customer reviewing process it can help the future customers to make their decisions wisely.

1.2.1 Problem statement

A model to analyse user reviews of websites such as Daraz will be helpful to stakeholders, Since making decisions based off from Sinhala-English code mixed customer reviews can be difficult *and time consuming for a monolingual user.*

As elaborated above Sinhala-English code mixed previous customer reviews can be found widely on websites such as Daraz, Kapruka. Making decisions based off from these reviews can be difficult since these reviews contain multiple languages. Although Sinhala can be considered as a low resource language it will be a good research contribution to try and apply current trends of sentiment analysis to this scenario.

1.3 Research motivation

Making decisions based on customer reviews from online shopping platforms like Daraz can be tough since user reviews contain Sinhala English code mixed data. Currently there is no system to analyse such user reviews.

Customer reviews should be taken into great consideration since it directly affects the decisions of future customers (Cernian, Sgarciu, and Martin, 2015). There are various types of review analysis systems for English language. The lack of a review analysis system for a low resource language such as Sinhala is an issue. A system that gives a summary of current customer reviews at a glance would be helpful in making decisions for future customers. Therefore developing a generalized model for both Sinhala and English review comments would impact the domain positively.

1.4 Related work

With the help of research in recent years NLP has improved a lot as a domain. Since the author of the research is considering applying sentiment analysis to a low resource language it is important to consider the existing work in both English and Sinhala languages. Already existing work on the Sinhala language on NLP area will be critically evaluated. There was no research found on sentiment analysis on Sinhala-English code mixed user review data to the best of authors' knowledge. Because of above mentioned reasons customer review analysis done in the Amazon platform will be considered.

1.4.1 Sentiment analysis in English language

Table 1.1: Existing work in English language

Citation	Summary	Technologies Used	Contributions	Limitations
(Rathor, Agarwal, and Dimri, 2018)	After extracting data from Amazon API naive bayes, support vector machine and maximum entropy was used for classification.	Support Vector Machine (SVM), Naive Bayes (NB), Maximum Entropy (ME)	Using weighted unigrams and applying them have increased accuracy.	-
(Alqahtani, 2021)	This research analyses Amazon reviews dataset with different machine learning approaches. Reviews were transformed into vector representations and analysed and finally best performing model was retrained on binary classification.	Bag of words, term frequency–inverse document frequency (TF-IDF) were used to transform reviews to vector representation. NB,Random Forest (RF), Bert, Long Short Term Memeory (LSTM) as models	Different approach on feature extraction was suggested.	Implement word2vec for feature extraction with these models. This can be extended to amazon reviews in general.

(Basiri et al., 2021)	An attention based bi-directional deep learning model was created. Convolution and pooling mechanisms were utilized.	Neural networks,LSTM	New two layers were added to model (Gru and bi-lstm).	Rating predictions and helpfulness predictions. This can be also extended to other languages.

1.4.2 Sentiment analysis in Sinhala language

Table 1.2: Existing work in Sinhala- English language

Citation	Summary	Technologies Used	Contributions	Limitations
(Kugathasan and Sumathipala, 2020)	Sinhala-English code mixed social comments were analysed. A dictionary was created using analysed data. Few normalisation techniques were used to reduce noisy data.	Normalisation methods to reduce noise.	A dictionary was created where Sinhala alphabet letters are mapped to Singlish text.	A model which would translate Singlish text to English using the developed dictionary is proposed as future work.

(Jayaweera, Senanayake, and Haddela, 2019)	This research mainly focuses on creating a stopword removal list for Sinhala language using the mentioned technologies. More than 90000 documents were used for this scenario.	SVM,NB, RF, Newton's iteration method	A list of Sinhala stop words and a optimization algorithm was developed.	Stopword list can be further improved.
(Kugathasan and Sumathipala, 2021)	Proposing a Neural machine translation model to translate Sinhala-English code mixed data to Sinhala language.	LSTM	A parallel corpus to help with future research.	Further works should be focused on entity extraction and sentiment analysis using the parallel corpus created.

1.5 Research gap

It is common among most Asian countries to use code mixed languages (Kugathasan and Sumathipala, 2020). These code mixed data can be widely found in social media and product reviews. Most bilingual people prefer to mix their native language with English to express their thoughts (Shanmugalingam, 2019).

With the work related to the customer review analysis systems, it is clear that most of the currently available systems mainly focused on the English language. When it comes to Sinhala, English mixed review comments there is no proper way of analyzing them. However, with the help of available limited resources it will be beneficial to apply current trends of deep learning techniques to Sinhala English mixed review data. It will undoubtedly aid the customers as well as the progression of NLP in Sinhala and English languages.

1.6 Research contribution

The author's contribution to the technological and domain areas have been explained below.

1.6.1 Technological contribution

Similar works will be studied when proposing the system to analyse Sinhala English mixed code data. There is currently no system is available to perform the mentioned task above. Available similar approaches will be analysed and a new system will be provided to automate analysing user reviews. The new developed model can be a huge contribution to the NLP in code mixing in low resource languages such as Sinhala. Since this model can be incorporated with further research such as user review star detection and customer feedback it can be identified as a major contribution to NLP.

1.6.2 Domain contribution

The proposed system will help stakeholders to make decisions easily by automating the analyzing process. This can be identified as the main contribution to the domain.

1.7 Research challenge

A lot of resources related to Natural language processing can be identified. When it comes to applying already available methods to code mixed data it can be challenging.

As stated above Machine Learning (ML) models tend to struggle when analysing code mixed data. Since there is only limited amount of research done in Sinhala-English code mixing it can be highly challenging to apply NLP techniques in a code mixed environment. As the author of this research elaborated above there is currently no system architecture to analyse code mixed user reviews. Due to these constraints similar products already available in English language were critically evaluated. Furthermore obtaining a dataset adhering to these conditions can be challenging as well.

As elaborated above most non English speakers tend to mix their native languages with English. Since a hybrid method is considered in this research machine translation should also be taken into consideration. Spelling errors, improper usage of discourse can be identified as challenges in neural machine translation (Kugathasan and Sumathipala, 2021). Applying these various NLP techniques and building a model which is capable of analysing Sinhala-English code mixed reviews will be challenging with above criteria.

1.8 Research questions

RQ1: What are the most suitable algorithms to analyse Sinhala and English mixed user reviews ?

RQ2: How can recent enhancements in sentiment analysis can be applied to analyse Sinhala comments ?

RQ3: What improvements needed to be done to already available models to effectively produce a summary about a product using Sinhala-English mixed reviews ?

1.9 Research aim

The aim of this research is to design, develop and evaluate a system that analyse given user reviews and provide customers with a summary which will help in decision making in Sinhala English mixed scenarios.

Further elaborating on this aim this research project will produce a system that is capable of analyzing customer reviews of a certain product in a Sinhala-English mixed code cases. The planned system will allow users to add previous user review or reviews in bulk and get a summary as a output based on those reviews. In order to achieve this aim NLP techniques, deep learning, data mining and available literature will be studied.

The required knowledge will be further studied using available literature and the performance of the product will be evaluated. The knowledge obtained will be used in developing the core components and achieving the research objectives. The author of this research will be publishing a research article based on the findings in the Sinhala-English code mixed data.

1.10 Research objectives

Research objectives are mapped to learning outcomes accordingly. These objectives must be full filled in order to conduct a successful research.

Table 1.3: Research objectives

Objective	Description	Learning Outcomes	Research Questions
Literature Review	A survey of already available work and their limitations. • **RO1:** Analyse similar existing review analysis systems. • **RO2:** A detailed study on machine learning and natural language processing techniques. • **RO3:** Finding out a feasible research gap using existing literature. • **RO4:** Finalizing programming languages, frameworks, datasets needed to conduct the research. • **RO5:** To identify methods and matrices that will be used in the evaluation phase.	LO1,LO4, LO5	RQ2, RQ1
Requirement Elicitation	Gather and analyse requirements needed in order to satisfy defined research questions and gap. • **RO1:** To identify software and hardware requirements through surveys. • **RO2:** To get insight from domain and technology experts.	LO3,LO4, LO6	RQ2
Design	Designing a architecture that is capable of solving the identified problems • **RO1:** To design a model that analyses Sinhala English code mixed user reviews. • **RO2:** To design a user interface to interact with the model.	LO2	

Implementation	Developing the proposed system. • **RO1:** Develop a model that can analyse Sinhala, English and both Sinhala and English mixed user reviews. • **RO2:** Develop a interface that can be used by customers in order to analyse reviews.	LO7, LO5	RQ3
Testing and evaluation	Conducting testing on the developed prototype and developed ML models. Functional, integration and usability testing will be carried out. • **RO1:** Creating a test plan and conducting the tests mentioned. • **RO2:** Getting feedback from domain experts. • **RO3:** Testing the prototype in simulations.	LO8	RQ3

1.11 Project scope

Scope of this research project is defined below according to research objectives and after analysing existing similar products on the market.

1.11.1 In-scope

The following elements can be identified as in-scope.

- **System to analyse user reviews** - A system that is capable of analyzing Sinhala English code mixed data and output sentiment score and a summary of current reviews.
- **Graphical User Interface (GUI) that allows user to interact with the model** - The initial prototype will be released as a web application.

1.11.2 Out-scope

The following are considered as exclusions from the project.

- **Only Sinhala English code mixed reviews** - With the time constraints only Sinhala and English or mixed reviews will be considered. Other languages such as Tamil can be added as future enhancements.

- **Real time fetching of reviews** - A web application that only support text as input and does not auto fetch reviews in real time.

1.12 Prototype diagram

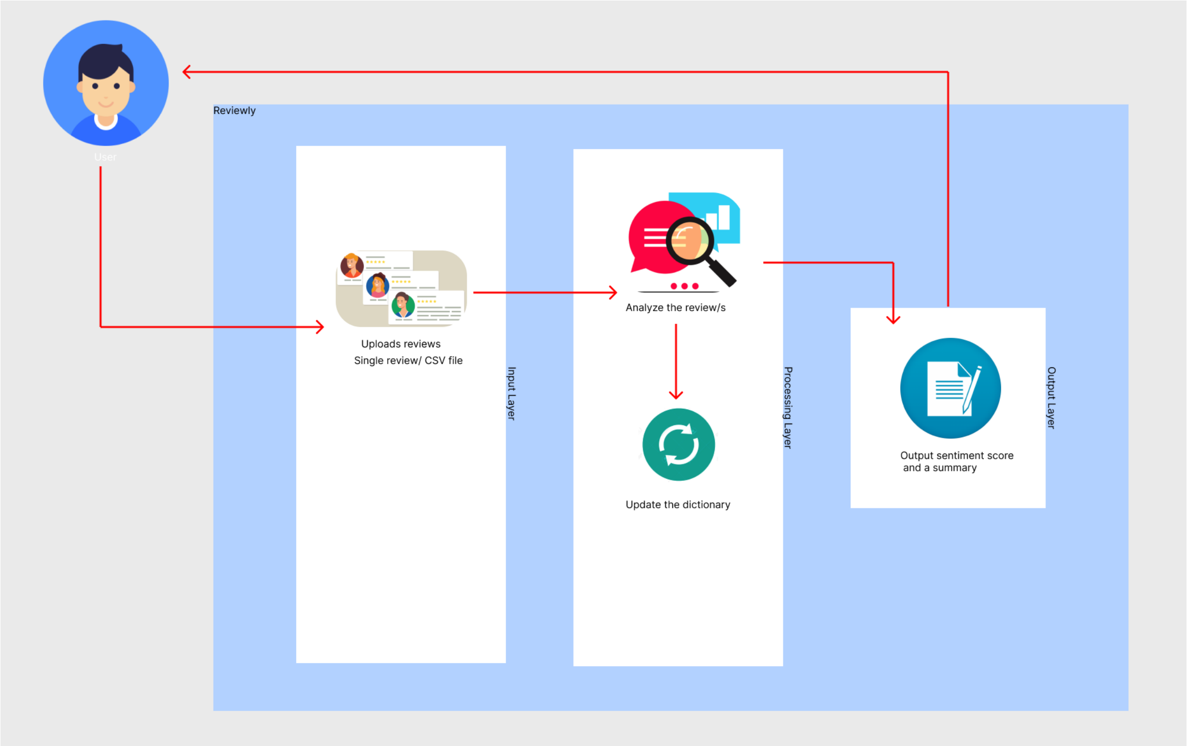

Figure 1.1: Prototype diagram *(self-composed)*

2 METHODOLOGY

2.1 Research methodology

In a research project, its quality is determined by cost, time and scope. The methodologies used in this project are listed down below. Saunders's research onion model can be used to explain different stages in a research. The following table includes the selections made and justification for each choice made by the author

Table 2.1: Research methodology

Research philosophy	**Pragmatism** was chosen as the research philosophy over positivism, interpretivism and realism. Surveys including close ended questions and interviews are used to conduct this research. This methodology suits this research best due to its nature of containing both qualitative and quantitative data.

Research approach	This research tends to experiment with currently available NLP techniques and algorithms. The author will try to prove the proposed hypothesis by experimenting and applying current techniques. **deductive approach** was selected from available candidates due to the above mentioned reasons.
Research strategy	Research strategy mainly covers the data collection methods used in this research. Currently available **literature** were referred. **questionnaires** were used to gather requirements and **interviews** were conducted to back up the research. Furthermore, **experimenting** with available algorithms and techniques were used to support the hypothesis proposed.
Research choice	Out of the available candidates such as mono, mixed, multi, **mixed** method was selected. As explained above this research contains both qualitative and quantitative data mixed method suits best.
Time horizons	Refinements and testing need to be done to the system constantly via data gathering. The data will be from various sources, since the model need to be tested on various data components.**cross-sectional** was chosen as the time horizon due to the above reasons.
Techniques and procedures	For collecting data surveys, interviews are used. Furthermore reports, survey papers and literature of existing work are used.

2.2 Development methodology

From the available development methodologies **prototype** methodology will be used. Since this research project is based on iterative development with trial and error, prototyping methodology will be the best candidate. The author will be constantly refining the system based on supervisor and expert feedback, therefore using a methodology that allows iterative development is a must.

2.2.1 Design methodology

Structured Systems Analysis And Design Method (SSADM) was chosen as the design methodology for this research project. Since this project includes experimenting on a trial and error basis SSADM was selected. It is highly practical and offers reusability for components which can be helpful.

2.2.2 Requirement elicitation methodology

Requirement elicitation can be also described as requirement gathering from stakeholders. Methods such as surveys and interviews will be used to gather requirements. Prototyping of the product will be done using visual representations to gather requirements.

2.2.3 Testing methodology

Testing the developed architecture is the process of evaluating whether the system meets the requirements. Testing methodology has been broken into three main components.

2.2.3.1 Prototype testing

The developed prototype will be tested using multiple methods. Since the frontend of the system is presented as a web page created using React framework unit testing will be done using Jest. Integration testing, cross browser testing and visual regression testing will be done to further evaluate the frontend component.

2.2.3.2 Model testing

Validating the machine learning model will be done in several stages. Automated unit testing and integration testing will be done while the model also be evaluated according to the testing metrics available. The sentiment accuracy of the model will be tested using precision, recall, cross entropy and f-score metrics.

2.2.3.3 Benchmarking

Since there is no similar review analyzing products are available benchmarking cannot be done. However the model developed will be tested against other models which use Sinhala-English code mixed data.

2.2.4 Solution methodology

Solution methodology defines the process which a problem is analysed and implemented solutions. The following diagram summarises the main steps in a NLP system.

2.2.4.1 Data gathering

Data gathering stage includes collecting available data and getting them ready to feed to the model. This step should be conducted cautiously since the quality of collected data directly affects the model. Data set should not be biased and must be balanced with true negative and positive data.

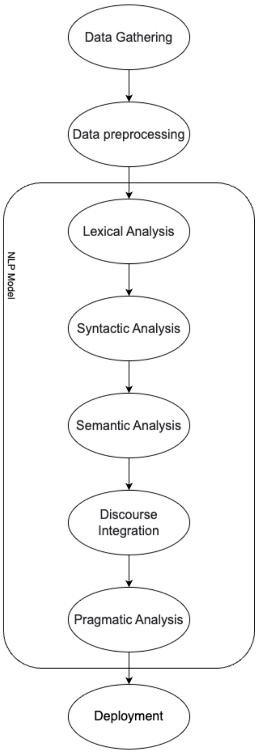

Figure 2.1: solution-methodology diagram *(self-composed)*

2.2.4.2 Data preprocessing

After the data gathering process, the textual data needs to be preprocessed before feeding to the machine learning model. Data preprocessing helps to mitigate the issues that can occur by Missing values, duplication, punctuations and stop words. Although different researchers tend to use different methods the main process of data preprocessing is given below.

- **Data cleaning** - Cleaning includes smoothing out noisy data, dealing with missing data and inconsistent data.
- **Data integration** - Detecting and removing redundant attributes, resolving data value conflicts are done in this stage.
- **Data transformation** - This step includes changing the structure of data. Normalisation, generalisation is mainly used to achieve it.
- **Data reduction** - The dataset acquired could be large in size. Methods such as data compression, numerosity reduction and discretization is used to reduce the size of the dataset.

After preprocessing dataset is divided into 60,20,20 ratios for training, validation and testing respectively. Although deep learning models tend to use 90,5,5 ratio (Alqahtani, 2021).

2.2.4.3 NLP model

- **Lexical analysis** - Lexical analysis can be identified as the first step in the compiler. Lexical analyser breaks the text code into tokens by removing white space. analyser takes lexemes as inputs and produce tokens as output using regular expressions. A sentence will be broken down into identifiers, operators, punctuations, literals and keywords.
- **Syntactic analysis** - In this stage model starts to extract meaning from the sentences. Meaningfulness of sentences will be checked in this stage. These validations will be done while adhering to formal grammar rules.
- **Semantic analysis** - Making sure the data is clear and consistent with a clear meaning. Important information such as emotions, context and sentiment are extracted in this phase. Processes such as word sense disambiguation and relationship extraction are used in semantic analysis.
- **Discourse integration** -Discourse integration Understanding the sense of context is known as discourse integration. Words like ”this”, ”that” which needs prior context to understand is identified in this stage.
- **Pragmatic analysis** - Dealing with outside word knowledge and abstracting meaning from those situations is known as pragmatic analysis.It mainly handles the overall communicative and how it affects the interpretation.

2.2.4.4 Deployment

After the evaluation of the model is completed it can be deployed to a cloud based service. It is important to keep the model updated with constant refinements.

2.3 Project management methodology

Out of widely available project management methodologies **Agile prince 2** hybrid model was selected. Prince2 focuses on dividing a large task into manageable chunks while agile focuses on iterative development. Recent research done on prince2 identifies it has benefits such as being plan oriented and well documented (Pawar and Mahajan, 2017). Due to the nature of this research project, this hybrid methodology would be ideal because of the deliverable plan.

2.3.1 Schedule

Gantt chart

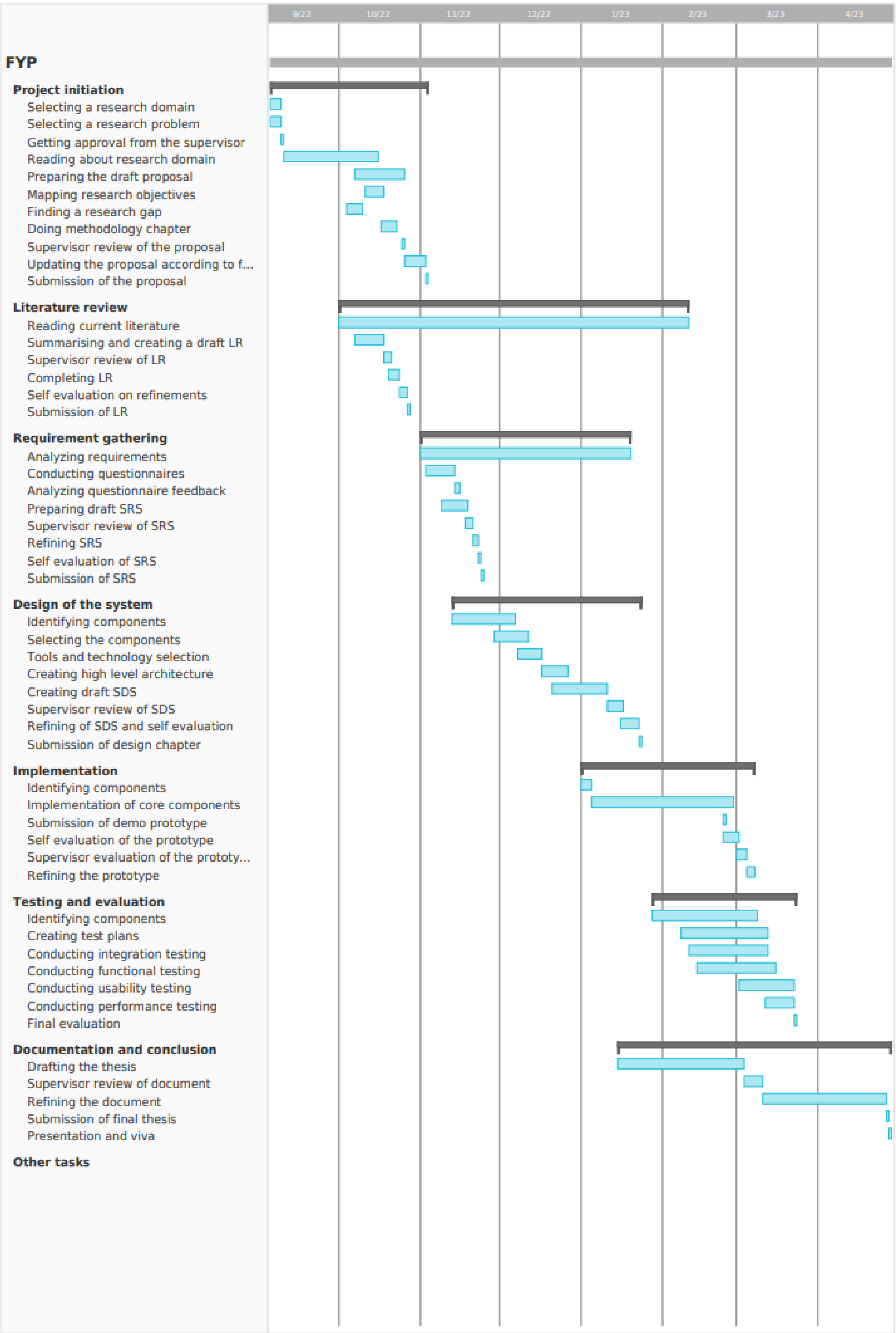

Figure 2.2: Gantt chart *(slef-composed)*

Deliverables and dates

Table 2.2: Deliverables

Deliverable component	Date
Project proposal Initial proposal for the research project	3rd November 2022
Literature review Evaluation of existing work and solutions	27th October 2022
Software requirement specification Requirements that needed to be completed in order to deliver a successful prototype	24th November 2022
System design The document which contains algorithms used and design diagrams	23rd January 2023
Prototype The initial prototype with all core features	23rd January 2023
Thesis Final document that elaborates the research process, findings and prototype	27th April 2023
Research paper A research paper that contains findings done in the project	27th April 2023

2.3.2 Resource requirements

Requirements needed to conduct the research are given below. They are given as software, hardware, data and skill requirements.

Software requirements

- **Operating system (Windows/ MacOSx/ Linux)** - Windows (x64 bit) operating system will be used to carry out the project since utilizing an external nvidia Graphics Processing Unit (GPU) is easier and faster when training models.
- **Python/ R** - Python is used as the preferable language to create the required models since it integrate well with any machine learning task and has more external libraries.
- **React/ Angular/ Ember** - React will be used to create the frontend of the system. Since react offers server-side rendering it will be faster than other frameworks when integrated

with a ML model.

- **Visual studio code/ Intellij/ Atom/ Jupyter** - VsCode with jupyter plugin installed will be used to develop the ML model. Vscode will be used because it is highly customizable and lightweight.
- **Git** - Git will be used as the version control system.
- **Mendeley reference manager/ Zotero** - Mendeley and Zotero both offer almost the same features. Mendeley will be used to manage all the research papers related to the project based on the personal preference of the author.
- **Overleaf** - Overleaf will be used to create the documentations since it's a cloud based LATEX editor.
- **Google drive/ iCloud/ Onedrive** - Google drive will be used to backup project and the documents. Google drive can be connected to Google colab which will make it easier to load datasets, save and train models.
- **Figma/ Adobe illustrator** - Figma will be used for the designing purposes of this research project since it's free.

Hardware requirements

- **Core i7 Processor(9th generation) or above** - To train NLP models efficiently.
- **8GB RAM or above** - To manage large number of datasets development enviroments (IDEs etc).
- **Nvidia GTX 1060ti GPU or above** - In order to reduce training time and to manage ML models.
- **at least 25GB of free hard disk** - To store data sets, trained models etc.

Skill requirements

- **Knowledge about ML and NLP** - A deep understanding about machine learning and natural language processing is needed to create and optimize models.
- **Programming knowledge** - Python is the most common language used in ML. A good understanding of python and other used languages are needed.
- **Research and academic writing skills** - Researching skills and creative writing skills is a must in order to carry out a complete research.
- **Designing** - A sound understanding about designing tools and concepts is needed for User Interface (UI) design.

Data requirements

- **Datasets to implement the NLP models** - Using Google dataset search and Kaggle.
- **Reviews to further test model** - Scraping reviews from websites.

2.3.3 Risks and mitigation

Table 2.3: Risk mitigation plan

Risk	Probability of Occurrence	Magnitude of the Loss	Mitigation Plan
Lack of knowledge about ML and NLP A deep understanding about machine learning concepts and mathematics is needed to complete the research project	5	5	Refer the limited number of available resources and reach out to domain experts if needed.
Inability to complete the project in time	3	5	Managing workflow and working on fixed deadlines
Hardware issues, Data loss Hardware corruption, physical damage to devices or an unpredictable hazards can lead to data loss	2	5	Keep all documents and code uploaded to GitHub and Google drive.
Fluctuation of requirements Requirements of the project could change at any given moment due to updates in technologies etc.	3	3	Prioritizing core features, constantly adapting to changing requirements can be challenging.

Insufficient hardware resources Developing NLP models and compiling need lot of hardware resources.	3	2	Using google colab or an external GPU to train heavier models.

References

Alqahtani, Arwa S M (2021). "PRODUCT SENTIMENT ANALYSIS FOR AMAZON REVIEWS". In: *International Journal of Computer Science Information Technology (IJCSIT)* 13 (3). DOI: 10.5121/ijcsit.2021.13302. URL: https://ssrn.com/abstract=3886135.

Basiri, Mohammad Ehsan et al. (Feb. 2021). "ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis". In: *Future Generation Computer Systems* 115, pp. 279–294. ISSN: 0167739X. DOI: 10.1016/J.FUTURE.2020.08.005.

Cernian, Alexandra, Valentin Sgarciu, and Bogdan Martin (Oct. 2015). "Sentiment analysis from product reviews using SentiWordNet as lexical resource". In: Institute of Electrical and Electronics Engineers Inc., pp. 15–18. ISBN: 9781467366465. DOI: 10.1109/ECAI.2015.7301224.

Chauhan, Shashank Kumar et al. (2017). "Research on product review analysis and Spam Review Detection". In: *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)*. DOI: 10.1109/spin.2017.8049980.

Jayaweera, A. A.V.A., Y. N. Senanayake, and Prasanna S. Haddela (Oct. 2019). "Dynamic Stopword Removal for Sinhala Language". In: *2019 National Information Technology Conference, NITC 2019*. DOI: 10.1109/NITC48475.2019.9114476.

Kugathasan, Archchana and Sagara Sumathipala (Mar. 2020). "Standardizing Sinhala Code-Mixed Text using Dictionary based Approach". In: Institute of Electrical and Electronics Engineers Inc. ISBN: 9781728165417. DOI: 10.1109/ICIP48927.2020.9367353.

— (2021). "Neural Machine Translation for Sinhala-English Code-Mixed Text". In: Incoma Ltd, pp. 718–726. ISBN: 9789544520724. DOI: 10.26615/978-954-452-072-4_082.

McCombes, Shona (Sept. 2022). *Developing strong research questions: Criteria amp; examples*. URL: https://www.scribbr.com/research-process/research-questions/.

Otter, Daniel W., Julian R. Medina, and Jugal K. Kalita (Feb. 2021). "A Survey of the Usages of Deep Learning for Natural Language Processing". In: *IEEE Transactions on Neural Networks and Learning Systems* 32 (2), pp. 604–624. ISSN: 21622388. DOI: 10.1109/TNNLS.2020.2979670.

Pawar, Rupali Pravinkumar and Kirti Nilesh Mahajan (Mar. 2017). "Benefits and Issues in Managing Project by PRINCE2 Methodology". In: *International Journal of Advanced Research in Computer Science and Software Engineering* 7 (3), pp. 190–195. ISSN: 22776451. DOI: 10.23956/ijarcsse/V7I3/0134. URL: http://ijarcsse.com/docs/papers/Volume_7/3_March2017/V7I3-0134.pdf.

Rathor, Abhilasha Singh, Amit Agarwal, and Preeti Dimri (2018). "Comparative Study of Machine Learning Approaches for Amazon Reviews". In: *Procedia Computer Science* 132, pp. 1552–1561. ISSN: 18770509. DOI: 10.1016/J.PROCS.2018.05.119.

Shanmugalingam, Kasthuri (2019). *Language identification at word level in Sinhala-English code-mixed social media text; Language identification at word level in Sinhala-English code-mixed social media text*.

Silva, Nisansa de (June 2019). "Survey on Publicly Available Sinhala Natural Language Processing Tools and Research". In: URL: http://arxiv.org/abs/1906.02358.

Singh, Jyoti Prakash and Alkhowaiter Wassan (2015). "Sentiment Analysis of Products' ReviewsContaining English and Hindi Texts". In: 9373. Ed. by Marijn Janssen et al. DOI: 10.1007/978-3-319-25013-7. URL: http://link.springer.com/10.1007/978-3-319-25013-7.

Tennage, Pasindu et al. (Feb. 2018). "Neural machine translation for Sinhala and Tamil languages". In: vol. 2018-January. Institute of Electrical and Electronics Engineers Inc., pp. 189–192. ISBN: 9781538619803. DOI: 10.1109/IALP.2017.8300576.

Young, Tom et al. (Aug. 2017). "Recent Trends in Deep Learning Based Natural Language Processing". In: URL: http://arxiv.org/abs/1708.02709.