Propensity Life Tracker – Ariadne Project

Final Proposal

15-12-2021

Team members:

Abhinandan Umakant Jawalekar

Jayarani Rajesh

Karthikeyan Suresh Kumar

Table of Contents

Our Team

Abhinandan Umakant Jawalekar

I am a graduate with Bachelors in Computer Applications. I have good experience in dealing with computers and various programming languages. I also have completed a Pro degree in Data Science which helped me enhance my analysis skills. Later I had worked on a project with a bank which helped me develop my skills like documenting, analyzing, etc.

Jayarani Rajesh

I'm a B.Sc Statistics graduate who completed the program in 2020. I have excellent statistical and mathematical skills required for Data Science. I have good knowledge in statistical analysis and data cleaning of the datasets. I am good with the Univariate, Multivariate, Inferential techniques which are primarily used in the data cleaning process.

Karthikeyan Suresh Kumar

Bachelor of Engineering, Electronics and Communication.

Pursuing a career in the field of Data Science.

I have 3+ years of experience as a Software Developer in HTC Global Services. The major chunk of my work is focused on the coding, testing, and execution of web applications for various banks such as RBI, ICICI Bank, Axis Bank, etc. I have expertise in programming languages – Java and python, Frameworks – Spring and Hibernate, SQL, Webservices – REST, and SOAP.

Introduction

Propensity-Life Tracker is a project idea by Teresa Woods Snelgrove and she is the project lead of this innovative project. Teresa is a successful serial entrepreneur who has had success in building numerous businesses. Teresa is accompanied by John Pickard, who is an exceptional entrepreneur, whose expertise and inputs would bring additional value and insights to the project. Teresa's vision of creating this project would in turn changes the lives of many individuals who are suffering from mental illness and stress. This project is a tool that would have a wide reach of audiences as it is can be used from monitoring the well-being of the kids from school, monitoring old people in homes, in hospitals for recovering patients, and so on.

Project Description

Propensity is a tool that helps a user capture a succession of thoughts or memories through dialogues/conversations with an intelligent agent (Like Alexa or Siri). Our agent is called "Ariadne". Through this tool, people can have conversations with Ariadne, which in turn analyses those conversations and converts them into meaningful reports which highlight the sentiment, subject, mood analysis of a person.

The tool could have alternative uses: A mood monitor, mental health monitor, sentiment analysis, or bias analysis, etc. It could be offered as a wearable device, a smartphone app, etc.

Objective

Further development of the project, from the end of phase 1. Phase 1 of the development of the project largely focused on the initial research and the design of the Machine Learning (ML) models. The team has developed two ML models: one for the classification of the Taxonomy and the other model is for the Sentiment Classification.

The Taxonomy classification model is based on the Naïve Bayes algorithm and this model has a prediction accuracy of 25%. The Sentiment classification model is based on the VADER (Valence Aware Dictionary for Sentiment Reasoning) model and it has a prediction accuracy rate of 70%.

Phase 2 of the project involves improving the performance of the working ML models for the Taxonomy and Sentiment classification, developing new models for Subject and Tense classification, creating datasets for training and testing the models, connecting the models with the database to store data, creating visualizations for the output data.
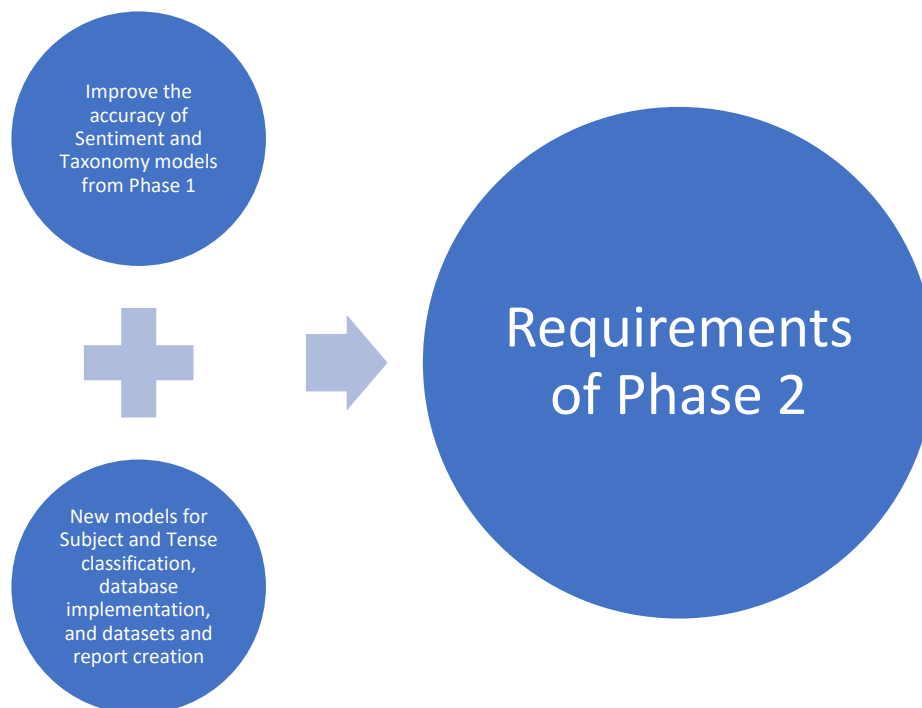
Rationale

With technology advancing in all aspects of life, it should be advancing in the field of mental health too, which is one of the reasons for the birth of the Project Propensity. The mental health issues vary from people of any age, from a school kid to old people. Everyone might have experienced mental health-related issues at least once in their lifetime. Also, in the post-

pandemic world, everyone around the world might have been through some mental stress: they might have lost a loved one, they would have lost their jobs, they might have been stuck wherever they are unable to reach home, etc. Everyone would have gone through something which would have affected them mentally. The Project Propensity might be a one-point diagnosis solution for that, the Ariadne agent listens to our conversations, analysis the data, and provides a report which would help to know their wellbeing from time to time.

Requirements

- Find possible ways to improve and enhance the performance of the current working models.
- Create new datasets/monologues to train the model.
- Design and develop a database to save the input data to be consumed and the output data generated by the machine learning model.
- Research on machine learning models which would be the best fit for the Tense and Subject classification.
- Implement the finalized machine learning algorithm for the Tense and Subject classification.
- Develop interactive reports based on the data from the ML models.

Improve the accuracy of Sentiment and Taxonomy models from Phase 1

New models for Subject and Tense classification, database implementation, and datasets and report creation

Requirements of Phase 2

Research Questions

The following research questions are the key challenges in Phase 2 and researching and analyzing those paves the way for the project.

- How to improve the accuracy of the Taxonomy model. Should we change the current ML model or find techniques to improve the accuracy.
- What models would be the best fit for Subject and Tense classification?
- How can we create/find data that matches or closely corresponds to real-life conversations?
- What database can be used to store the processed data? And how to design it?
- There are many visualization tools in the market. Which tool would be apt for our analysis visualization?

Scope of Project

The scope of the project in Phase 2 is as follows

- Performance enhancement of Taxonomy and Sentiment Analysis models.
- Research and implementation of new models for Tense and Subject classification.
- Database design and development for storing the input and output data.
- Creating new datasets or finding new data libraries for training and testing the ML models.

Out of Scope:

The plans of the project include:

- Developing a chatbot
- Making an Android/IOS app

Outcomes and Benefits

Upon the completion of Phase 2, the core part of the project would be almost over. The outcomes and benefits upon completion of Phase 2 would be the following:

- Closely following with Phase 1, Phase 2 of the project also adds two more models for Subject and Tense classification.
- More monologue datasets will be available.
- Database will be implemented so that the team can view and analyze the past data.
- As the data is stored, the visuals developed can be across a wide time frame.
- The core part of the project will be completed, so that the next phase of the project which includes building UI, developing app/wearables tech can proceed.
- A SharePoint site will be created containing all the necessary research documents, analysis documents, workflow documents, source code, sample datasets, etc.

- Clients would have an idea of what has been completed so far and the next necessary steps to carry forward the project.

Deliverables

Upon the completion of Phase 2, the following components will be delivered to the clients.

1. Improved ML models for Taxonomy and Sentimental Analysis
2. Model for Tense classification
3. Model for Subject classification
4. Datasets used for training/testing the models
5. Database design architecture and scripts
6. Visualization of the analysis results
7. Project-related documents
8. Reference documents

Execution

Resources

The resources readily available for the project are as follows:

- Source code of the project
- Documentation of the work done by the Phase 1 project.
- Required tools
- Research documents
- Sample datasets

Process

The process to be followed in Phase 2 of the project is as follows:

1. Research
2. Design
3. Development
4. Testing
5. Validation
6. Result Analysis
7. Documentation

Process Visualization

Research:

This is the key stage of the project as it involves analyzing the various machine learning algorithms which would prove to be competent for the type of classification chosen. The

advantages and disadvantages of the models are analyzed and their efficiencies are compared to choose the right model which would fit our classification problem.

Design:

The Design part of the model should be concentrated because only an efficient design would lead to an efficient model. The Design part involves how the data is handled and how the model is being fed with input data and what would be the expected output data. It also covers the design of the database.

Development:

The development process closely follows the design process and follows the steps as per the design process in building the model. All the logic and the functions intended to be followed will be developed with the help of the IDEs and the supporting python libraries.

Testing:

Once the development of the model is complete, the testing should be done to check whether the model is working as expected or is having any issues/errors. The datasets to train and test the model must be either created or curated from websites.

Validation:

After testing with the training data, the results of the model must be validated to check whether the model is generating the results as expected. If the efficiency of the model is not satisfactory, then ways to optimize the model's efficiency should be explored and implemented.
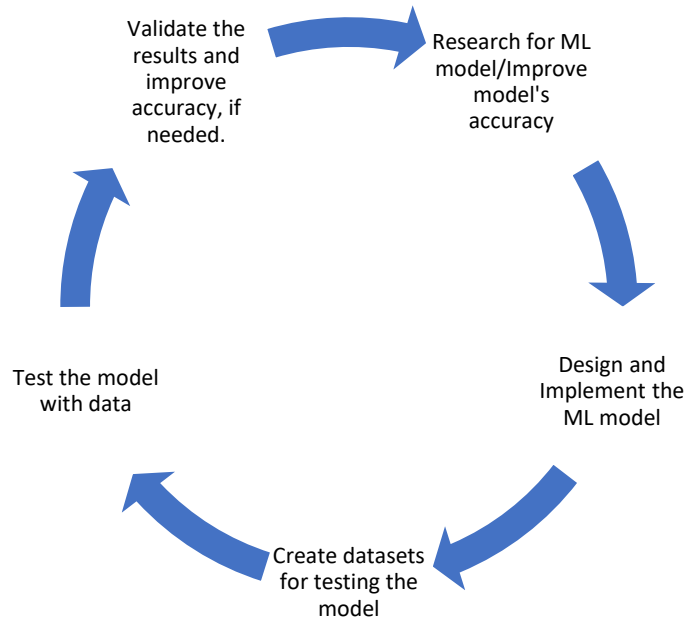
Result Analysis:

After all the above process is completed successfully and the expected results are obtained, various visualization software must be used to create reports, dashboards, charts, etc. Only with the proper analysis, it can be ensured that the model is created as per the requirements and the expectations.

Documentation:

This is a crucial process and must be carried out in all the stages of development. Documenting all the processes, requirements, steps, analysis, and results of the project provides an insight to the clients as well as helps anyone or any team to work on the project with ease. Also, the documents are a part of the deliverable, so it is necessary that it is being followed in each process.

Validate the results and improve accuracy, if needed.

Research for ML model/Improve model's accuracy

Test the model with data

Design and Implement the ML model

Create datasets for testing the model

Tools and Software

The following tools have been used in the project and the same will be used for further development.

- Python 3.x
- Python Libraries: Numpy, Pandas, NLTK, Vader, Text blob, sci-kit learn, etc.
- Tableau 2021.2.4
- PyCharm
- Visual Studio Code
- MS-Office Suite

Python 3.x: Python is an interpreted high-level general-purpose programming language. It is the most used programming language for building machine learning models, artificial intelligence, data science, and even web application development. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs, as well as its object-oriented approach, aim to help programmers write clear, logical code for small and large-scale projects

Python Libraries:

Various python libraries like Numpy, Pandas, Scikit-learn, NLTK, Vader, Text blob, etc., are the core libraries used for building the machine learning models. The major libraries used in the project are as follows:

Numpy: NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

Pandas: Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

NLTK: The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language.

Scikit-learn: Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means, and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

VADER: VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments.

Text Blob: TextBlob is a Python library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, etc.

Tableau: Tableau is a visual analytics platform transforming the way we use data to solve problems—empowering people and organizations to make the most of their data.

PyCharm: PyCharm is an integrated development environment (IDE) used in computer programming, specifically for the Python language. It is developed by the Czech company JetBrains (formerly known as IntelliJ). It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django as well as data science with Anaconda.

Visual Studio Code: Visual Studio Code is a source-code editor made by Microsoft for Windows, Linux, and macOS. Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git.

Microsoft Office Suite: Microsoft Office, or simply Office, is a family of client software, server software, and services developed by Microsoft. It consists of Microsoft Office, Microsoft Excel, Microsoft PowerPoint, etc., bundled together.

Potential Challenges

Although all the requirements have been captured during the Requirement analysis phase, some obstacles may cause delays and challenges in the project. The identification of those obstacles and their risk management is very much needed to ensure that the project proceeds smoothly without any hindrance to the delivery.

Challenges	Description	Response Plan
Lack of Knowledge	The tasks completed in the previous phase must be conveyed to the new team accurately to ensure a smooth start to the next phase.	Collect details of all the performed tasks done by the Phase 1 team and the necessary documents. Also, arrange a meeting with the Phase 1 team for knowledge transfer.
Data Issues	The Phase 1 team has hinted that the data for training/testing the application is very limited.	Gather the existing data used for training/testing of the model and have knowledge of data created for the project.
Time Frame	Time frame is the time targeted for the completion of a task.	If there is an unavoidable delay in the execution of the project, the project manager must take a call on how to provide a workaround/solution for that situation.
Model Selection	Selecting a machine learning model, for the classification of Tense and Subject will be based on extensive research on what model would be a perfect fit for the classification topic.	Perform extensive research on which machine learning model will be the best fit for this research.
Client Unavailability	The clients are the driving and the success factor for each project. They are the ones who take key decisions regarding the project.	In case the clients are unavailable due to certain circumstances, the project manager should be able to advise the team on the decisions.
Delayed Client Responses	Like the previous challenge if the client provides delayed responses which compromises the delivery date.	The project manager should be able to communicate with the client in case of delayed responses.

Code of Conduct

We will ensure ethical practices are observed demonstrating respect, honesty, and dignity. We will thoroughly follow the directives of the client and analyze the results as directed by them without deviation from the objectives of the research. Details of our analysis will be provided to the client. We will be transparent in all our activities. Steps will be taken to circumvent unethical behavior such as exaggerating the results of our research, misrepresenting the data to validate our findings, or changing or omitting details in favor of making our analysis fit our hypothesis. Moreover, we will check and recheck our research for quality, accuracy, completeness, and integrity before presenting it to maintain the validity and credibility of the results. Participants in the team have no conflict of interest and will not benefit from this analysis as this is a volunteer project.

Milestones

Activity	Completion Date
Kick off meeting with clients	19/Nov/2021
Preparation of first draft proposal document	22/Nov/2021
First transition meeting with Phase 1 team	25/Nov/2021
Requirements document preparation	29/Nov/2021
Preparation of second draft proposal document	29/Nov/2021
Second transition meeting with Phase 1 team	02/Dec/2021
Final MRP proposal documentation	06/Dec/2021
MRP final project charter preparation	13/Dec/2021
Initial Analysis and exploration of Phase 1 materials	20/Jan/2022
Preparation of the datasets for the models	27/Jan/2022
Research which database would be a good fit to store data and its implementation with the models.	10/Feb/2022
Exploring different ways and improving the performance of the Taxonomy and Sentiment models	03/Mar/2022
Research which ML model would be best fit for the classification of Tense and Subject and the implementation of these models.	31/Mar/2022
Collating the work done and preparing a report covering all the aspects of the project and making the visualizations.	14/Apr/2022
Final meeting with the client and the delivery due of the project	21/Apr/2022

Conclusion

For our team, it would be a great pleasure and an excellent opportunity to work on a project which would revolutionize the mental health industry. We have a hope that the product that we are building would distinguish itself from its peers in the industry.