A Final Project Final Report on

XSA: Sentiment Analysis using SVM

Submitted in Partial Fulfillment of the Requirements

for the Degree of **Computer Engineering**

Under **Pokhara University**

Submitted by:

Aashrit Thapa, 191303

Ashish Ayer, 191312

Siddhant Raj Nath, 191349

Unique Pradhan, 191346

Under the supervision of

Mr. Simanta Kasaju

Date:

August 10, 2024

Department of Computer Engineering

NEPAL COLLEGE OF INFORMATION TECHNOLOGY

Balkumari, Lalitpur, Nepal

SUPERVISOR'S APPROVAL

This is to certify that the major project entitled "XSA: Sentiment Analysis using SVM" undertaken and demonstrated by Unique Pradhan, Siddhant Raj Nath, Aashrit Thapa and Ashish Ayer has been successfully completed under my supervision as a partial fulfilment of the requirements for the degree of Bachelors of Engineering in Computer Engineering under Pokhara University. I, henceforth, approve this project to be awarded the certificate by the concerned authority.

During supervision, I found students hardworking, skilled and ready to undertake any professional work related to this field in future.

Mr. Simanta Kasaju

Supervisor

Date: 22nd August, 2024

EXAMINER'S ACCEPTANCE

This is to certify that the major project entitled "XSA: Sentiment Analysis using SVM" presented by Unique Pradhan, Siddhant Raj Nath, Aashrit Thapa and Ashish Ayer as a partial fulfilment of the requirements for the degree of Bachelors of Engineering in Computer Engineering under Pokhara University has been examined and accepted by the following panel of experts. We, henceforth, recommend this project to be awarded by the certificate from the concerned authority.

We extend all the best wishes to the students for their future careers.

Tulasi Dahal

Examiner

Date: 22nd August, 2024

Suman Dahal

Examiner

Date: 22nd August, 2024

CERTIFICATE

Following the Supervisor's Approval and Examiners' Acceptance, the major project entitled "XSA: Sentiment Analysis using SVM" submitted by Unique Pradhan, Siddhant Raj Nath, Aashrit Thapa and Ashish Ayer as a partial fulfilment of the requirements for the degree of Bachelors of Engineering in Computer Engineering under Pokhara University, has been officially awarded by this certificate.

I wish the students all the best for their future endeavours.

Mrs. Resha Deo

Head, Department of Computer Engineering

Date: 22nd August, 2024

DECLARATION

We Unique Pradhan, Siddhant Raj Nath, Aashrit Thapa and Ashish Ayer, students of Bachelors of Engineering in Computer Engineering, Nepal College of Information Technology affiliated to Pokhara University, hereby declare that the work undertaken in this major project entitled "XSA: Sentiment Analysis using SVM" is the outcome of our own effort and is correct to the best of our knowledge. This work has been accomplished by obeying the engineering ethics; and it contains neither materials published earlier or written by another person/people nor materials which has been accepted for the award of any other degree or diploma of the university or other institution, except where due acknowledgement has been made in the document.

Unique Pradhan

Student

Date: 21st August, 2024

Siddhant Raj Nath

Student

Date: 21st August, 2024

Aashrit Thapa

Student

Date: 21st August, 2024

Ashish Ayer

Student

Date: 21st August, 2024

Acknowledgement

The development of this project would not have been possible without the joint efforts of many connected individuals. It has been a pleasure for us to acknowledge the assistance and contributions that were very essential and supportive throughout the project. We would like to extend our sincere thanks to all of them. We owe a special sense of gratitude towards a number of people who have devoted much of their time, opinion and expertise without which it would have been very difficult for us to implement our project.

We owe a great deal of gratitude to our project supervisor, Mr. Simanta Kasaju sir for his insightful counsel during the project's development phase and for his technical support and recommendations, which enabled our project to advance to a degree we never would have imagined in such a short amount of time.

Finally, but just as importantly, we would like to express our sincere gratitude to our teachers and colleagues who, whether consciously or not, have contributed to this project and shared their opinions and support throughout its whole evolution.

Unique Pradhan

Siddhant Raj Nath

Aashrit Thapa

Ashish Ayer

Nepal College of Information Technology,

Balkumari, Lalitpur, Nepal

Abstract

The rapid rise of internet and its aspects such as social media sites (X, Facebook), forum sites (Reddit) and ecommerce services in recent years has drawn a lot of attention to sentiment analysis where offers a plethora of real-time textual data, including user thoughts and attitudes on a range of issues. The goal of this project is to create a sentiment analysis system designed especially for such data collection. The project analyzes and categorizes statements into positive, negative, and neutral sentiment categories using machine learning algorithms and Natural Language Processing (NLP) approaches. Preprocessing techniques like stemming, tokenization, and stop word removal are used to improve the quality of the incoming data. To extract pertinent information from the sentences, feature extraction techniques like stemming and tokenization are applied. The sentiment analysis model is trained and assessed using a labeled dataset made up of statements with sentiment annotations and parameters. To examine the accuracy and efficiency of various classification methods, including VADER and TextBlob, Support Vector Machine (SVM) is used. The outcomes show how well and precisely the system that was designed could classify the feelings that were stated in the data. The results emphasize how crucial it is to take into account the distinctive features, like the usage of jargons, hashtags, and acronyms, in order to increase sentiment analysis accuracy. In summary, this project advances the science of sentiment analysis by offering perspectives on the difficulties and possibilities associated with sentiment analysis on and by suggesting a sentiment analysis system that works well on such particular social platforms.

Keywords

Twitter, Facebook, Reddit, sentiment, Natural Language Processing (NLP), tokenization, VADER, TextBlob, Support Vector Machine (SVM), jargons, hashtags

Table of Contents

List of Figures

List of Tables

1 Introduction

Internet and networking sites have completely changed how individuals worldwide express their thoughts and feelings. It becomes difficult to decipher the thoughts underlying the millions of text messages that are generated every day by users. Opinion mining, another name for sentiment analysis, has become a potent tool for automatically identifying and categorizing the sentiments included in such data.

Sentiment analysis is quite valuable for many applications. Companies can use it to learn more about what customers think and feel about their goods and services. To create campaigns that are specifically targeted, marketing experts can recognize trends and feelings. Individuals can assess the general sentiment surrounding particular events or topics, and researchers can examine popular perspectives on social and political concerns. The goal of this project is to design an efficient sentiment analysis system in order to overcome these issues. Our goal is to automatically categorize message statements into positive, negative, and neutral sentiment categories using machine learning algorithms and natural language processing (NLP) [1] approaches. An extensive labeled dataset of statements with sentiment parameters will be used to train the algorithm.

The project will investigate the usage of sentiment lexicons and domain-specific lexicons to increase the precision and contextual understanding of sentiment analysis. These resources offer more sentiment data on particular sectors, subjects, or frequently used terms on platform. Through comprehensive experimentation and analysis, we aim to contribute insights and advancements to the field of sentiment analysis.

1.1 Problem Statement

Social media sites have grown in popularity as a means for people to voice their thoughts and feelings in the current digital era. But one of the biggest challenges is gleaning meaningful insights from the vast amounts of textual data produced. Businesses, researchers, and individuals cannot fully utilize the power of sentiment analysis to comprehend user feelings and make data-driven decisions if there is a lack of an efficient system designed.

- The existing sentiment analysis techniques and tools may not adequately capture the nuances and characteristics unique to social platforms.

- Factors such as specific jargon, hashtags, or emoticons require specialized approaches for accurate sentiment classification.

- The sheer volume and velocity of data should necessitate comprehensive sentiment analysis capabilities to keep up with the dynamic nature of user sentiments.

1.2 Problem Objectives

The major objectives of our project are as listed below:

- To develop a sentiment analysis model that accurately classifies the textual data into positive, negative, or neutral sentiments.

- To evaluate the performance of the analysis model using appropriate metrics and compare it with existing approaches and models.

1.3 Significance of the Study

The sentiment analysis project holds substantial significance due to its potential impact on various stakeholders and its ability to unlock valuable insights from the textual data. It can provide businesses with crucial insights into customer sentiments, opinions, and feedback about their products, services, or brand. By accurately classifying sentiments, businesses can identify areas of improvement, adapt their marketing strategies, and make informed decisions to enhance customer satisfaction and loyalty.

Understanding user sentiments can contribute to enhancing the overall user experience. By analyzing sentiments expressed in user feedback, comments, or reviews, we can identify pain points, address user concerns, and optimize their platform to better align with user expectations, resulting in improved satisfaction and engagement. It also facilitates the analysis of public opinions on various social, political, or cultural issues. Researchers can gain valuable insights into public sentiment trends, identify emerging topics, and analyze the impact of events or policies on public sentiment.

1.4 Scope

The project phases must be designed and carried out in such a manner that it fulfills all of the objectives. The scope of the project is listed below:

- Development of an efficient sentiment analysis program to ascertain whether the general public's attitude toward the topic of interest is neutral, negative, or favorable.

- Collection of substantial diverse and representative dataset of textual data from various platforms covering different topics, user demographics, and time periods to ensure comprehensive sentiment analysis.

- Evaluation and research of marketing, political, branding and workforce analytics to comprehend meaningful insights.

1.5 Limitation

The realized limitations of the project are as follows:

- Biases in the datasets, such as overrepresentation of certain keywords or topics has led to skewed sentiments to be produced.

- The analyzer model struggles to capture the sentiment associated with the evolving language due to emergent of many new words, slangs and phrases.

- User input textual data may include emojis, sarcasm, irony, or other forms of figurative language that has led to difficulty to accurate analysis, misinterpretation and misclassification.

2 Literature Review

Opinion mining, another name for sentiment analysis, is a popular topic of study that examines and categorizes sentiments found in textual data. With an emphasis on sentiment analysis this literature review attempts to give a summary of pertinent papers and methods in the field.

2.1 Related Theory

2.1.1 Machine Learning Standards

Sentiment analysis makes extensive use of machine learning methods including supervised and unsupervised learning. Labeled data is used to train supervised learning models, in which text is annotated by humans to determine its sentiment. These algorithms categorize attitudes in unseen text by using patterns and attributes that they discover from the labeled data. In supervised learning, neural networks, pre-trained models [2], and Support Vector Machines (SVM) [3] are frequently used. Supervised learning is commonly used for classification tasks, where the goal is to assign an input data point to one of several predefined categories or classes. Examples include spam email detection, image recognition, and medical diagnosis.

2.1.2 Social Networking Analytics

The structure and dynamics of social relationships and interactions are the main subjects of analysis in social network analysis theory. Social network analysis can be used in sentiment analysis to take into account how social connections—like friends, followers, or communities—affect how sentiment is expressed. Sentiment analysis can capture sentiment polarization inside social networks, identify prominent users, and find sentiment diffusion patterns by using social network analysis.

2.2 Review of related works and research

Nearly all pertinent research studies share a common methodology that involves gathering data via the API, preprocessing and filtering the data, and then applying techniques for feature extraction, classification, and pattern analysis to distinguish the results. Furthermore, they employed a variety of algorithms and techniques to ascertain the impact of an active entity on the tweet patterns of people displaying particular emotions. Instead, then mining the entirety of user-posted data, they only mined the textual data at the entity level, i.e., brand, product, and celebrity aspects. Their methodology, which employed algorithms to extract traits and monitor their impact and influence, set their study apart from previous research in the field. Following preprocessing, the feature extraction procedure involved building n grams and using POS taggers to handle the negation portion and increase classification accuracy.

- In sentiment analysis, feature extraction is essential. Scholars have utilized diverse methodologies to extract pertinent attributes from textual data, such as topic modeling, n-grams, bag-of-words models, and word embeddings (like Word2Vec and GloVe). These methods make it easier to represent text data and find patterns linked to sentiment [4].

- Researchers have explored techniques such as domain-specific lexicons, sentiment-specific word embedding, and domain adaptation methods to improve sentiment classification accuracy for specific domains. There has been development of domain-specific sentiment analysis model for Twitter, (for example [5]) incorporating domain-specific lexicons and slang dictionaries to improve sentiment classification

performance.

- Deep learning models have also shown promising results in sentiment analysis tasks, capturing complex patterns and contextual information (for example [6]).

3 Proposed Methodology

This section explains the methodology that is proposed to be followed during the development of the project.

3.1 Proposed Software Development Life Cycle

In order to fulfil the requirements of our project, Sentiment Analysis, we planned to work using these approaches for the application of knowledge, skills, tools, and techniques to a wide range of activities. This section provides comprehensive details regarding our project's methodology, software development process, and tool selection. We have intended to use the incremental approach (as shown in Figure. 1) as the framework for this project's development. This model combines the iterative prototype approach with the linear sequential model. As each iteration is developed, new features will be added. The linear sequential model consists of four phases: analysis, design, coding, and testing. Iteratively, the program goes through these stages again and again before delivering an increment with progressively more modifications.

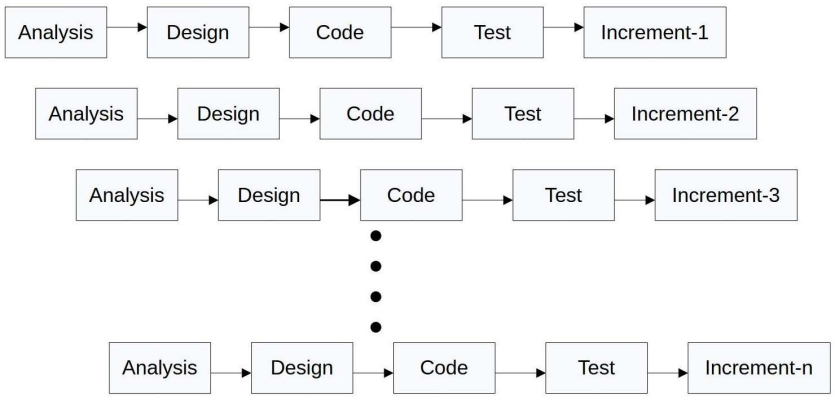

Figure 1: Incremental Model

The project is achieved through the accomplishment of the following increments:

3.1.1 Increment I: Model building and data collection

An analysis model is developed with the help of Python libraries and programs such as Pandas AND NumPy which the model will firstly split text into individual words or sub words through tokenization, reduce words to their base or root forms via Stemming and Lemmatization and then convert text data into numerical representations that machine learning models can process. The model then judges with evaluation metrics.

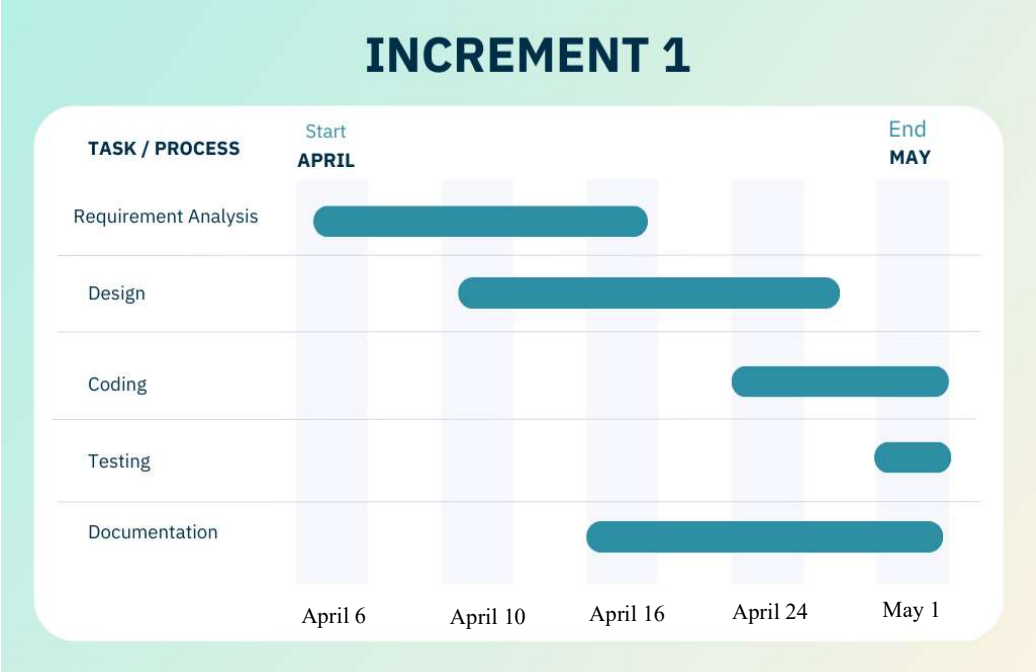

Figure 2: Gantt chart for Increment 1

3.1.2 Increment II: Model Enhancement

We have enhanced the performance of the model by normalizing and standardizing the datasets, which ensures that the features are on a similar scale. Considering techniques like Min-Max scaling, undersampling and oversampling help in dealing with unbalanced datasets.

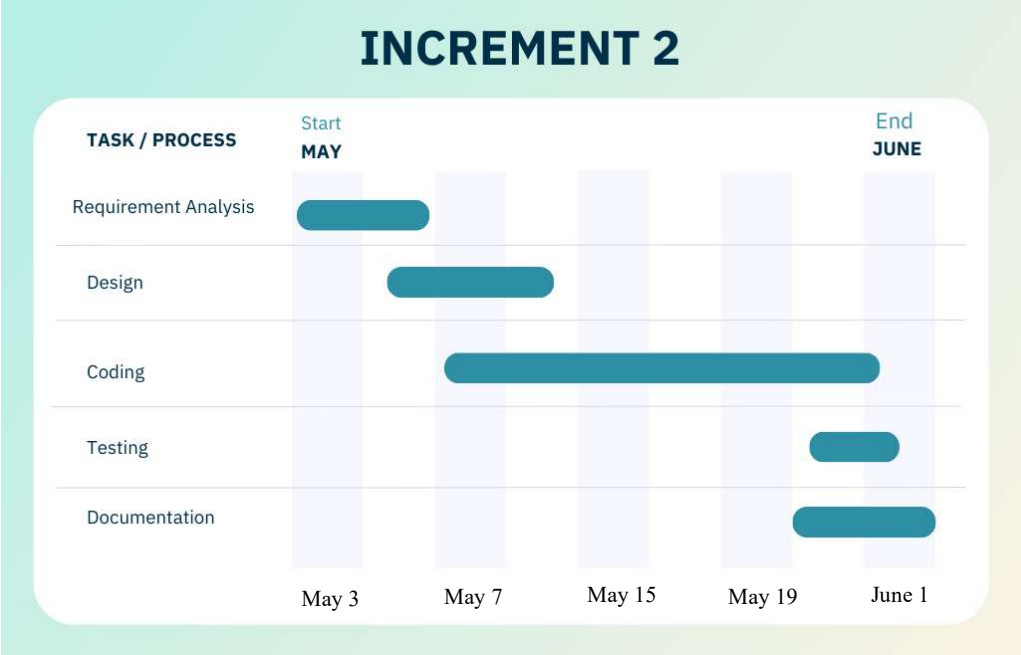

Figure 3: Gantt chart for Increment 2

3.1.3 Increment III: Analysis and Evaluation

The errors made by the enhanced model is identified for common patterns or challenging cases and then these errors are addressed by further refining the preprocessing steps or incorporating additional features. Methods like error analysis, ensemble learning, or active learning are enforced to improve the model's performance. The result produced by the model is then compared with the results of other pre-trained models.

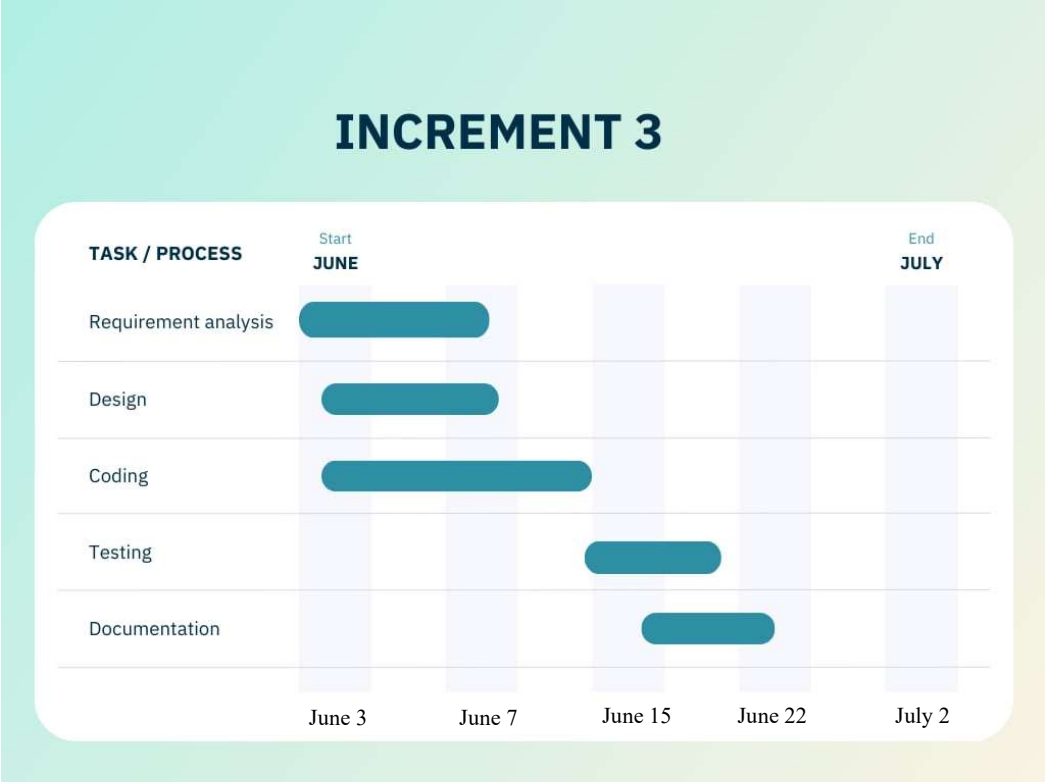

Figure 4: Gantt chart for Increment 3

3.1.4 Increment IV: Web App and Data Visualization

The analysis SVM model (analyzer and vectorizer) is then pickled which is then implemented through a web app developed in Flask. The model can analyze multiple sentences at once and represent the result graphically while comparing its derived results with the results obtained by a pre-trained model.

3.2 System Flow Diagram

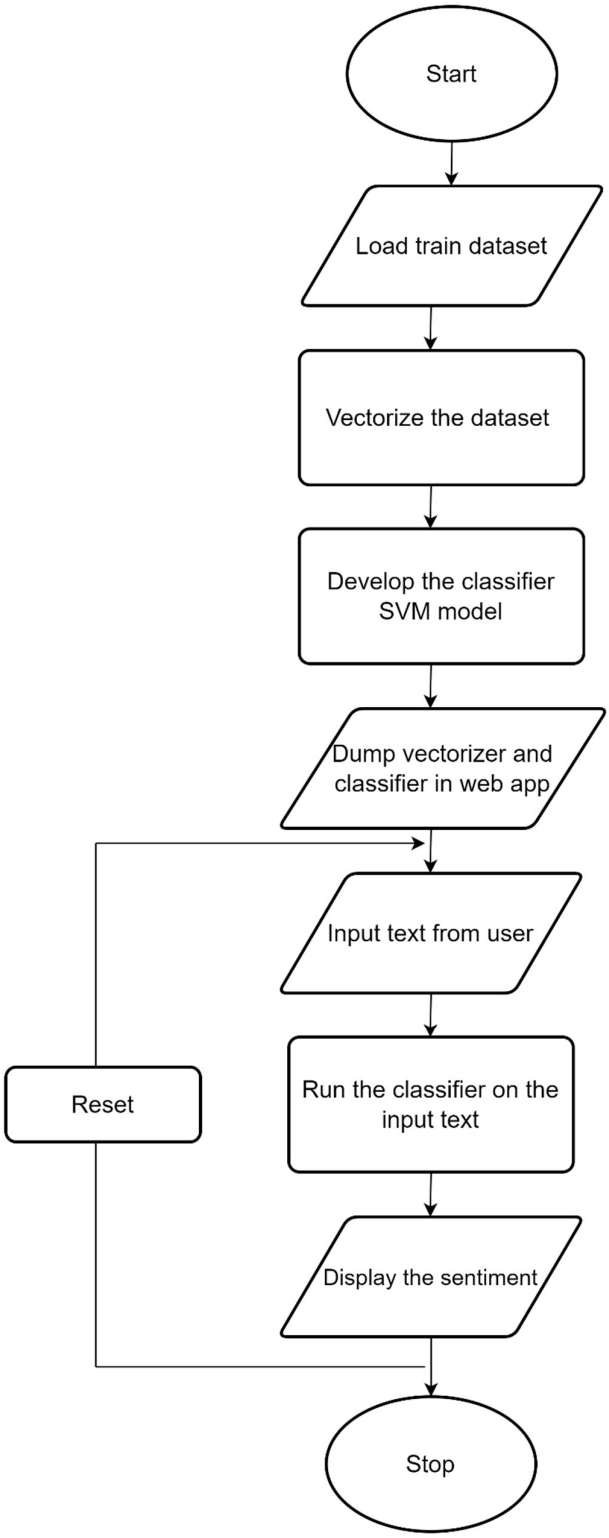

Figure 5: System Flow Diagram

3.3 Use Case Diagram

Figure 6: Use-case Diagram

3.3 Proposed Technologies

TOOLS	PURPOSE
Python	High level programming language, general-purpose programming language
NLTK and Pandas	Symbolic and statistical natural language processing
Numpy	Natural Language Processing
Flask	Web app development
TextBlob	Result Comparison

Table 1: Proposed Technologies

4 Tasks Done so far

This section explains about the tasks that our team have been able to accomplish till date.

- Creation of an SVM model, which has been trained on a bulk dataset consisting of reviews, tweets, comments and statements along with their sentiment polarity.

- Implementation of pre-trained models (VADER and TextBlob) for the purpose of comparison and evaluation of the model result.

- Deployment of a simple web app, developed through Flask for statement input and displaying the result.

5 Results and Discussion

The Sentiment Analysis app is now at the initial phase having most of the basic functionalities. All the modules have been working properly after integrating and are ready for the demo. As the features adding up the level of complexity has been increasing as well. However, it is not complete with the ideas we have put through and might need more improvisation in the coming days as well.

- The developed model still delivers results with relatively less accuracy as compared to the pre-trained models due to the model being trained only within a limited dataset.

- The web app has not been fully developed which makes it very basic and needs more workings in it.

6 Tasks Remaining

Due to the Sentiment Analysis app having a larger scope and potential in the current networking industry there can be a lot of improvements and future extension carried out within the project. Some of such future extensions and improvements that our team have thought of are as follows:

- To improved and enhance the web app with better UI experience.
- To enhance the SVM model with the help of a larger dataset consisting of larger variations of textual data.

7 Project Deliverable

The XSA: Sentiment Analysis project aims to bring forth a comprehensive and fluid solution for judging and classifying the opinions and sentiments of the local public people towards certain world topics and products aligning with the defined architecture and project objectives. The proposed deliverables encompass various components and functionalities of the sentiment analyzer:

7.1 Web App

A web portal is developed as key deliverables for the project. This platform serves as the primary interface for user interaction and facilitating engagement.

7.2 Analyzer Model

Through the implementation of various Natural Language Processing (NLP) libraries and programs such as NLTK, NumPy and Pandas, an SVM model is developed which is capable of interpretation of bulk of data and then judging the sentiment.

7.3 Result Comparison and Evaluation

The analyzed result is compared and evaluated with the results of the Textblob, then a ‘Sentiment Score’ is assigned to the analyzed result.

1 Methodology for Performance Analysis

For the purpose of making the Sentiment Analysis (XSA) project as much effective and robust as possible we will be making sure that the project model is trained from numerous training datasets and rigorously tested on countless test datasets and then validating it with various validation schemes and benchmarks tests.

An analysis and evaluation of the system's performance can be conducted using various metrics, such as accuracy, precision, and F1 score. Confusion Matrix, which is shown in the table below, can also be used to measure real-time performance.

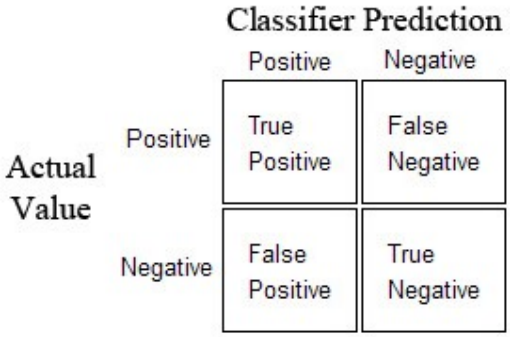

Figure 6: Confusion Matrix for Sentiments

To guarantee generalizability and reduce over-fitting, the models are validated. The system's performance is compared to the outcomes of other pre-trained models to establish the project's benchmark. These proposed methodology and validation scheme will provide a structured approach in evaluating the sentiment system’s performance, ensuring robustness and accuracy in its implementation.

Annex:

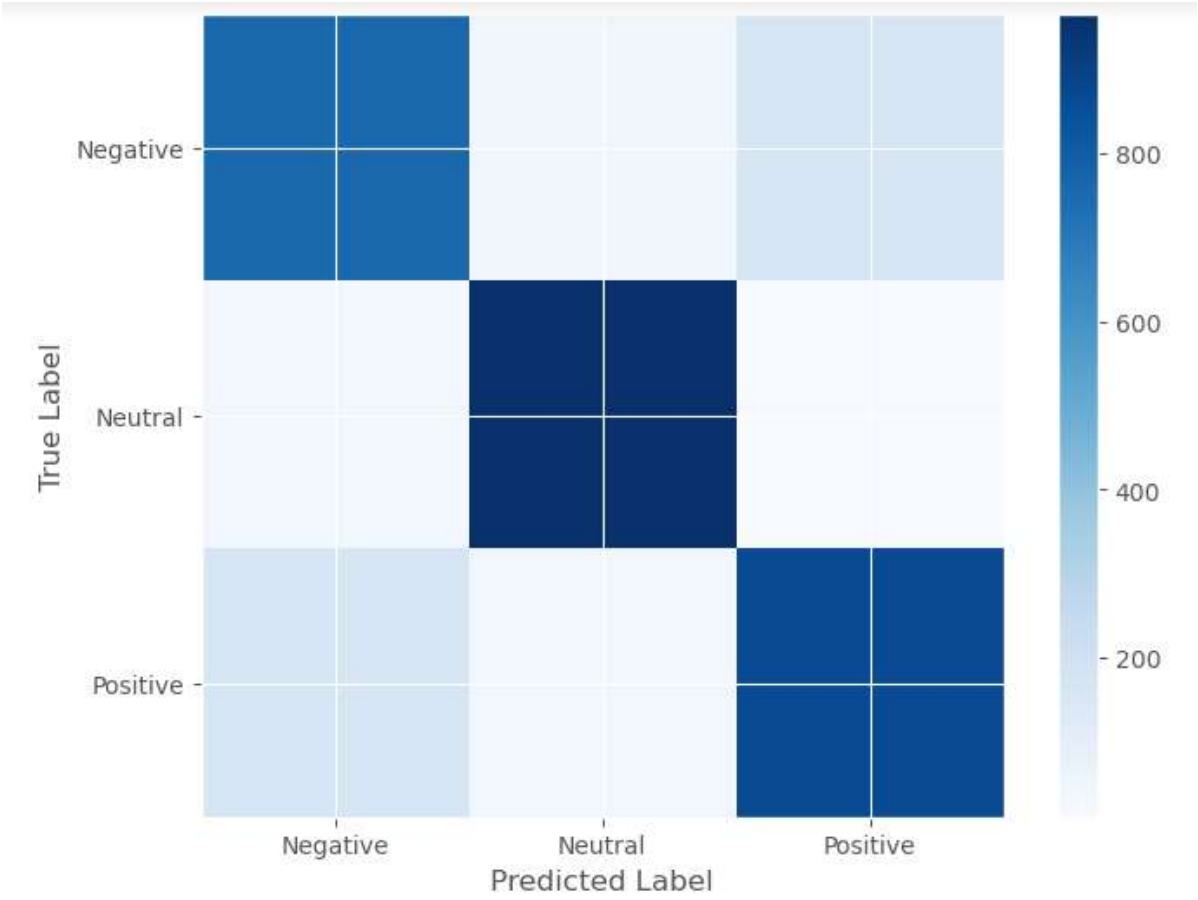

Figure 7: Confusion Matrix for XSA (SVM) Model with F1-Score of 85

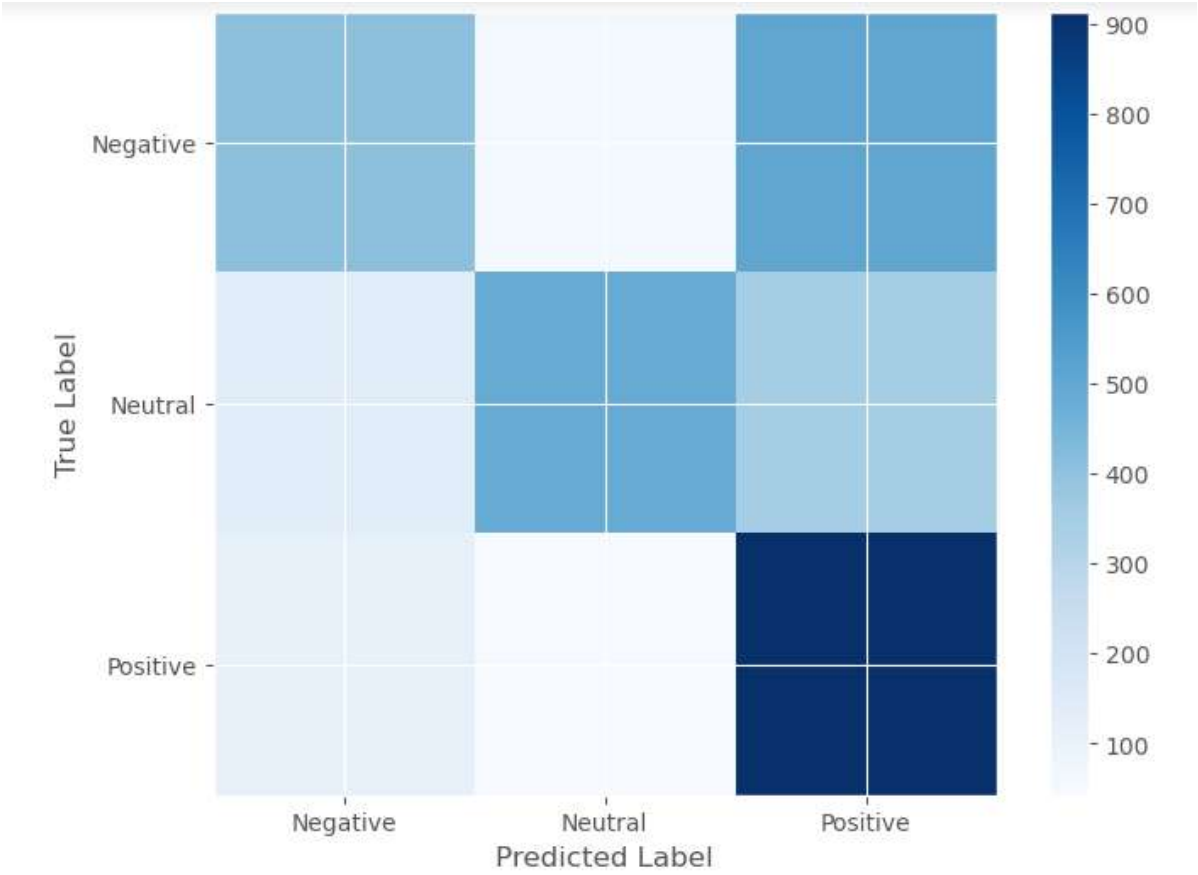

Figure 8: Confusion Matrix for VADER Model with F1-Score of 60

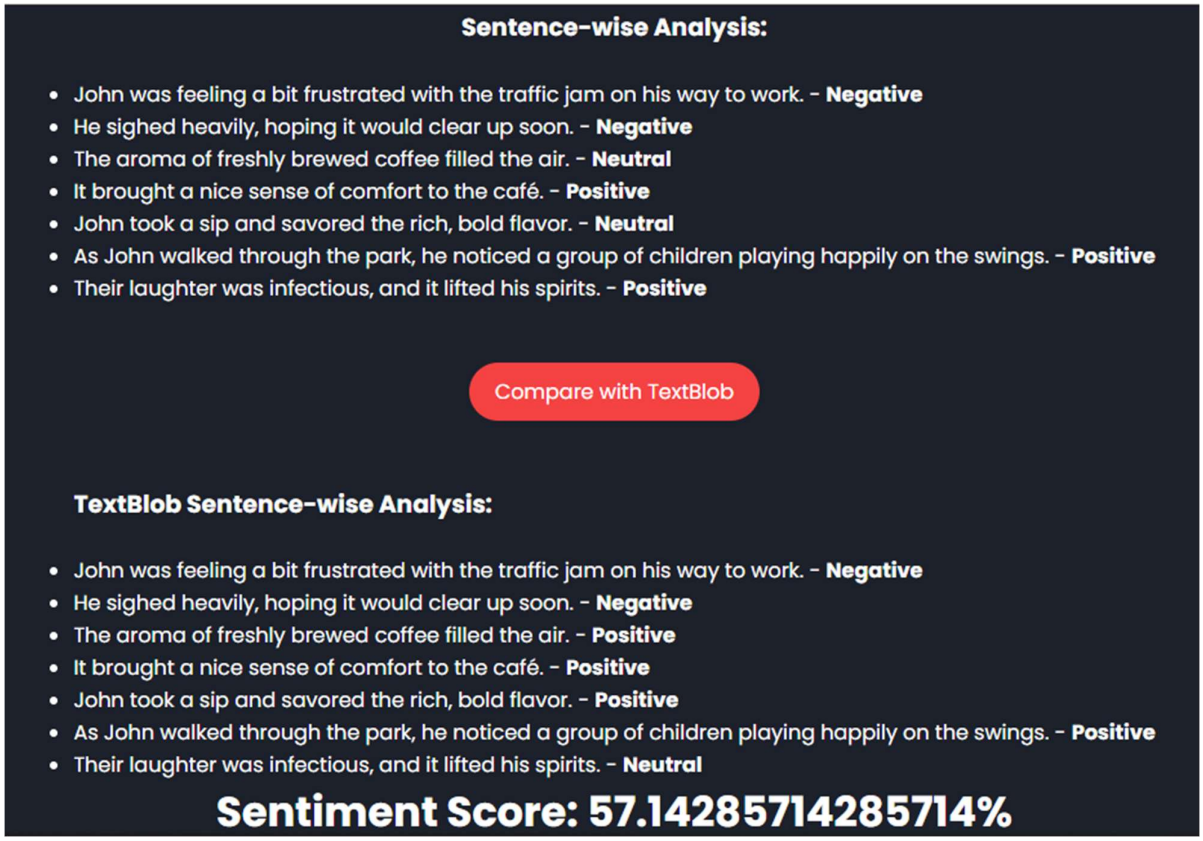

Figure 9: Comparison with TextBlob

2 Team Members and Work Division

For the efficient accomplishment of the set increments of the project, our team members have taken upon the following work division:

TEAM MEMBER	ASSIGNED ROLE/S
Siddhant Raj Nath	System Workflow Designing, Pre-trained model/s implementation.
Unique Pradhan	Model Designing and Development, Project Management
Aashrit Thapa	Web App Development, UI Designing and Development
Ashish Ayer	User-End Documentation, Model Testing and Evaluation

Table 2: Team Members and Work Division

References

[1] "Natural Language Processing: State of the art, current trends and challenges", Diksha Khurana, Aditya Koli, Kiran Khatter, 2022

[2] "Encyclopedia of Machine Learning", Claude Sammut, Geoffery I. Webb, 2017

[3] Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, 2000

[4] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space.

[5] Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural networks. Expert Systems with Applications, 40(16), 6266-6282.

[6] Kim, Y. (2014). Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language