SENTIMENT ANALYSIS ON REAL TIME REDDIT DATA

PROJECT SYNOPSIS

OF MINOR PROJECT

BACHELOR OF TECHNOLOGY

COMPUTER SCIENCE AND ENGINEERING

(2021-2025)

SUBMITTED BY

Sanskar Dorwal	Somya Agarwal	Vanshika Verma
2115129	2115140	2115158
2104183	2104197	2104214

UNDER THE GUIDANCE OF

ER. KULJEET KAUR

GURU NANAK DEV ENGINEERING COLLEGE,

LUDHIANA

TABLE OF CONTENTS

Content	Page No.
1. Introduction	3
2. Rationale	4
3. Objectives	5
4. Literature Review	6
5. Feasibility Study	7
6. Methodology/Planning of Work	8
7. Facilities Required	10
8. Expected Outcomes	11
9. References	12

INTRODUCTION

The project revolves around empowering users to explore and analyse sentiments associated with specific topics on Reddit. Users will be able to input their topic of interest, and the interface will dynamically fetch and present relevant data from Reddit threads. This ensures a personalized and focused approach to sentiment analysis.

The core technologies driving this project include RoBERTa - a state-of-the-art natural language processing model, SentiWordNet - a lexical resource for sentiment analysis, and mBERT - a multilingual BERT variant. These advanced NLP models enable the system to comprehend and analyze textual data across multiple languages, providing a comprehensive and inclusive sentiment evaluation.

This research-oriented project delves into the field of sentiment analysis, a domain critical for understanding the pulse of online communities. The application of sentiment analysis in real-time on Reddit data allows researchers to gain insights into public sentiment, opinions, and reactions related to various topics.

Special Technical Terms:

- RoBERTa (Robustly optimized BERT approach): A transformer-based deep learning model specifically designed for natural language understanding tasks, known for its superior performance in a wide range of NLP applications.

- SentiWordNet: A lexical resource that assigns sentiment scores to words, allowing for a more nuanced analysis of sentiment in text by considering the polarity of individual words.

RATIONALE

This project focuses on real-time sentiment analysis for Reddit data, aiming to understand public sentiments on the diverse platform. Leveraging Reddit's vast data, the project employs personalized analysis and advanced NLP models, including RoBERTa and SentiWordNet, for accurate and up-to-date sentiment evaluations. Multilingual considerations address inclusivity, catering to global perspectives. The user-friendly interface democratizes sentiment analysis tools, making them accessible to a broader audience. The real-time nature of sentiment insights extends practical applications to decision-making processes across sectors. Overall, the project contributes a comprehensive, multilingual, and user-friendly tool, meeting the increasing demand for insights into public sentiments in the evolving digital landscape. The inclusion of graphical representations in the user interface enhances accessibility, contributing to the project's goal of inclusivity and versatility across linguistic and cultural contexts.

OBJECTIVES

The objectives of this project are:

- To implement real-time data retrieval from Reddit for up-to-date sentiment analysis.
- To classify the data based on languages and pre-process for implementing sentiment analysis models.
- To extend sentiment analysis to support multiple languages.
- To develop a user interface that receives user input and dynamically display results through graphs and charts.

LITERATURE REVIEW

Shanghao Li et al. [1] (2023) proposed a model that aims sentiment analysis and topic modelling regarding online classes on the reddit platform. They investigated the sentiments, concerns, and temporal variations in opinions towards online classes among learners and educators on reddit. It Revealed polarized sentiments in online class discussions on Reddit, emphasizing learners' negative emotions, highlighting the pervasive issue of cheating with proposed solutions, and recommending extended tracking periods and increased educator engagement for community building.

G. Karuna et al. [2] (2023) proposed a model which conducts feasible sentiment analysis for real time twitter data. They developed a robust and efficient system that can perform sentiment analysis on real-time twitter data. It combines the power of the RoBERTa model for sentiment analysis and web scraping techniques using snscrape to collect real-time Twitter data.

Simran Siddhu et al. [3] (2018) proposed a critical review on sentiment analysis of hindi language text. They comprehensively reviewed and analysed the existing approaches in Hindi sentiment analysis. The literature review establishes a foundation for future research in sentiment analysis and opinion mining specifically tailored for the Hindi language.

P. V Veena et al. [4] (2018) proposed a system for character embedding for language identification in hindi-english code-mixed social media text. They proposed a novel technique for identifying the language of Hindi-English code-mixed data. The research successfully classified Hindi-English code-mixed data into distinct categories using popular word embedding features and introduced a novel character-based embedding approach.

Om Trivedi et al. [5] (2023) conducted sentiment analysis on twitter data using different models. Their aim was to compare them and find out the best model for performing sentiment analysis. They found that RoBERTa model's accuracy was highest.

FEASIBILITY STUDY

1. Technical Feasibility:

Existing Technologies: The project leverages well-established natural language processing (NLP) technologies such as RoBERTa, SentiWordNet, and mBERT, which have proven success in sentiment analysis tasks.

2. Economic Feasibility:

Cost-Benefit Analysis: The benefits derived from real-time sentiment analysis, including research insights, decision support, and personalized analysis, outweigh the implementation and maintenance costs.

Resource Utilization: Efficient utilization of open-source NLP models and tools minimizes the economic burden, making the project economically viable.

3. Operational Feasibility:

User-Friendly Interface: The development of a user-friendly interface ensures operational feasibility by making the tool accessible to a diverse range of users, including researchers, analysts, and decision-makers.

Scalability: The system is designed to scale with the growing volume of Reddit data, ensuring continued operational efficiency as the user base expands.

4. Schedule Feasibility:

Parallel Development: By dividing the project into manageable modules, parallel development can be undertaken, speeding up the overall implementation process.

METHODOLOGY

1. Data Collection using PRAW (Python Reddit API Wrapper):

   - Utilize PRAW to fetch real-time data from Reddit, including posts, comments.

2. Language Classification:

   - Employ language identification tools or libraries (e.g., langid.py) to classify the text data into two main categories: English and Hindi.
   - Further categorize English text into Romanized Hindi and pure English based on language patterns.

3. Translation of Romanized Hindi:

   - Use translation libraries or services to convert Romanized Hindi text to English for uniform sentiment analysis.
   - Ensure proper handling of transliteration nuances during translation.

4. Data Preprocessing: Perform standard text preprocessing steps, including:

   - Lowercasing.
   - Removing stop words, punctuation, and special characters.
   - Tokenization.
   - Handling emojis and emoticons.

5. Sentiment Analysis Models:

   - Apply RoBERTa for sentiment analysis on English texts, considering its advanced capabilities in understanding contextual nuances.
   - Utilize Hindi SentiWordNet for sentiment analysis on Hindi texts, incorporating the lexicon's sentiment scores for better accuracy.

6. Graphical Representation:

    - Use Matplotlib and Seaborn libraries to create visualizations such as bar charts, pie charts, and line graphs.

7. Web Application Development using Flask:

    - Implement a Flask-based web application to showcase the real-time sentiment analysis results.
    - Design a user-friendly interface for inputting topics and navigating through sentiment insights.

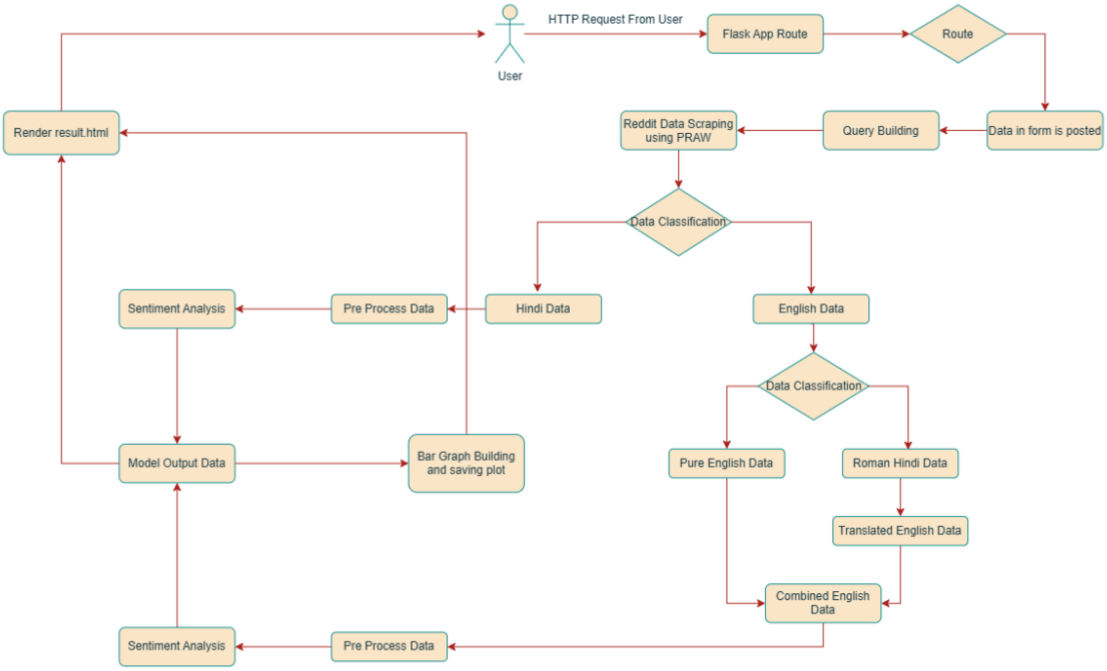

Fig 1: Flowchart of working

FACILITIES REQUIRED

Implementation will be done using the following tools and softwares:

- Necessary Python libraries, including PRAW for Reddit data extraction.
- Language identification tools or libraries for categorizing text into English and Hindi.
- Translation libraries for converting Romanized Hindi to English.
- Standard NLP preprocessing libraries for tasks such as lowercasing, tokenization, and stop word removal.
- Suitable sentiment analysis models for analyzing English and Hindi texts.
- Matplotlib and Seaborn for creating visualizations such as bar charts, pie charts, and line graphs.
- Flask framework for building the user interface and integrating sentiment analysis results into a web application.

EXPECTED OUTCOMES

This project aims the creation of a user-friendly interface for real-time sentiment analysis on Reddit data. The interface is designed to handle multilingual content, specifically in English and Hindi, providing users with sentiment classifications (positive, negative, neutral). Additionally, the interface will offer graphical representations, including bar graphs and pie charts, to visually convey sentiment trends. This outcome seeks to empower users with an efficient and insightful tool for comprehending the diverse sentiment landscape present in Reddit discussions, enhancing the overall user experience and analytical capabilities.

We will also provide a comprehensive project report and documentation for future reference and collaboration.

REFERENCES

[1] Li, Xie, K. W. Chiu and Kevin K. W. Ho "Sentiment Analysis and Topic Modeling Regarding Online Classes on the Reddit Platform: Educators versus Learners" Applied Science 2023, 13, 2250.

[2] G. Karuna, Anvesh, Sharath Singh, Reddy, Shah, Shankar "Feasible Sentiment Analysis of Real Time Twitter Data" E3S Web of Conferences, 010 (2023) ICMPC 2023

[3] P. V. Veena, M. Anand Kumar, K. P. Soman "Character Embedding for Language Identification in Hindi-English Code-mixed Social Media Text" Comp. y Sist. vol.22 no.1 Ciudad de México ene/mar. 2018

[4] Simran Sidhu, S. Khurana, Munish, Parvinder Singh, S. Bamber "Sentiment analysis of Hindi language text: a critical review" Multimedia Tools and Applications published online 11 November 2023

[5] Om, Harmeet "Sentiment Analysis of Twitter Data" Hans Shodh Sudha, VOL. 3, ISSUE 4, (2023), pp. 31-37

[6] Sathya, Thulasimani, Deva Surithi, "A STUDY ON TWITTER SENTIMENT ANALYSIS USING DEEP LEARNING"International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056