SENTIMENT ANALYSIS ON REAL TIME REDDIT DATA

Presented by Group No. 48

Sanskar Dorwal, Somya Agarwal, Vanshika Verma

Sentiment Analysis

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique that involves determining the sentiment or emotion expressed in a piece of text. The goal is to understand whether the text conveys a positive, negative, or neutral sentiment. This analysis is commonly applied to social media comments, reviews, and other forms of textual data to gauge public opinion, customer feedback, and overall sentiment towards a particular topic, product, or service.

Reddit

Reddit is a social media platform and online community where registered users can submit content, such as text posts, links, images, and videos, and engage in discussions with other users. The platform is organized into "subreddits," which are individual communities or forums centered around specific topics of interest. These topics can range from technology and science to entertainment, news, hobbies, and more. Users on Reddit can upvote or downvote content and comments, influencing their visibility and popularity. The platform is known for its diverse and passionate user base, and discussions can be informative, humorous, or serious, depending on the subreddit's focus.

Overview of Project

The Multilingual Sentiment Analysis Project on Real-Time Reddit Data aims to provide a versatile and inclusive sentiment analysis tool by leveraging real-time data from Reddit. Using the Python Reddit API Wrapper (PRAW), the system collects posts and comments in multiple languages. The project incorporates advanced sentiment analysis models, such as RoBERTa for English and Hindi SentiWordNet for Hindi, ensuring a nuanced understanding of sentiments. To enhance inclusivity, language classification is implemented, with a focus on both English and Hindi, including Romanized Hindi. The web application, built with Flask, facilitates user interaction, allowing users to input topics of interest for real-time sentiment insights.

Objectives

01

To implement real-time data retrieval from Reddit for up-to-date sentiment analysis.

02

To classify the data based on languages and pre-process for implementing sentiment analysis models.

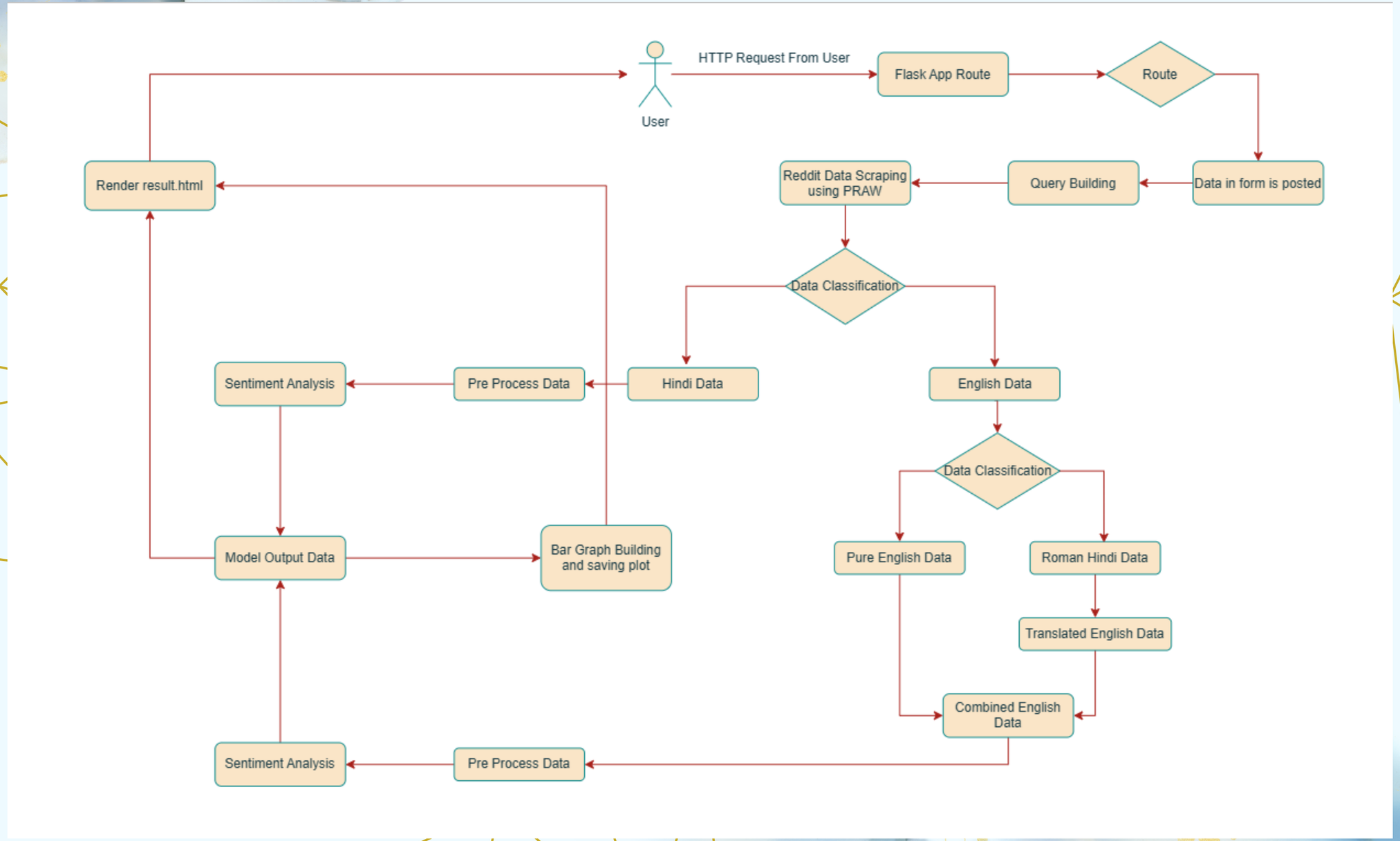
03

To extend sentiment analysis to support multiple languages.

04

To develop a user interface that receives user input and dynamically display results through graphs and charts.

Flowchart


User
HTTP Request From User
Flask App Route
Route
Data in form is posted
Query Building
Reddit Data Scraping using PRAW
Render result.html
Data Classification
Hindi Data
Pre Process Data
Sentiment Analysis
English Data
Data Classification
Pure English Data
Roman Hindi Data
Translated English Data
Combined English Data
Pre Process Data
Sentiment Analysis
Model Output Data
Bar Graph Building and saving plot

Methodology

The methodology involves user interaction through a web application developed using Flask. Users input topics on the website, triggering an HTTP request that passes through Flask routes. The data in the form is posted, and a query is constructed to facilitate the collection of relevant data from Reddit in real-time. This user-driven input mechanism enhances the dynamic nature of the sentiment analysis process, allowing for tailored insights based on user-specified topics.

Methodology Continued

Continuing forward, the methodology involves real-time data collection from Reddit using PRAW. Text data is then classified into English and Hindi, with further differentiation between Romanized Hindi and pure English. Romanized Hindi is translated to English, considering transliteration nuances. Standard text preprocessing steps are applied, including lowercasing, stop word removal, and tokenization. Sentiment analysis is performed using RoBERTa for English and Hindi SentiWordNet for Hindi. Graphical representations are generated using Matplotlib and Seaborn. Finally, a user-friendly web application is developed with Flask to display real-time sentiment analysis results and allow user interaction.

Tools to be Used

Data Collection and Classification

PRAW, langid, langdetect

Data Translation and Pre-processing

googletrans, re, NLTK

Sentiment Analysis and Graph Plotting

RoBERTa, SentiWordNet, Matplotlib, Seaborn

UI Designing

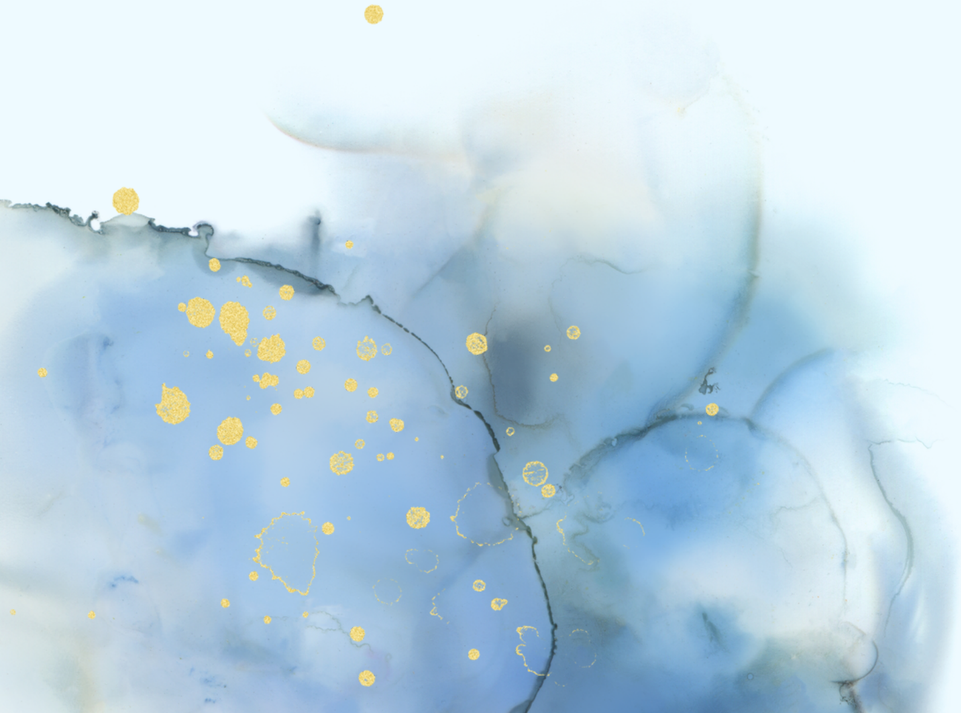
HTML, CSS, Flask

Enhancements

The Multilingual Sentiment Analysis Project exhibits a distinctive focus on inclusivity, employing language classification to accommodate diverse linguistic sources, particularly English and Hindi, including Romanized Hindi. Real-time data analysis using Python Reddit API Wrapper (PRAW) ensures up-to-date insights from Reddit. Notably, the project addresses translation nuances during Romanized Hindi to English translation, emphasizing precision in sentiment representation. This holistic approach combines multilingual capabilities, real-time data analysis, and nuanced translation handling to create a comprehensive and inclusive sentiment analysis tool.

Applications

1. Social Media Monitoring
2. Customer Feedback Analysis
3. Market Research
4. Predicting election outcomes



Future Scope

01

Increasing Model's Accuracy

02

Sarcasm Detection

03

Extension to other platforms

04

Increasing the System's Functionality

THANK YOU