Team Name: Big Data Enthusiasts

Team Members:

1. **Asmita Sonavane (as20428)**
2. **Vishwajeet Kulkarni (vk2630)**
3. **Sarang P. Kadakia (sk11634) (Team Lead)**

Big Data-Powered Trend and Sentiment Analysis for Code-Mixed Hindi-English Tweets

Abstract

Understanding trends and sentiment in Hindi-English code-mixed tweets is challenging due to language mixing, spelling variations, and transliteration differences. To solve this, we use Big Data technologies like Apache HBase for data storage, Spark NLP for text preprocessing, and PySpark MLlib for structured data preprocessingOur system integrates BM25 for information retrieval and uses IndicBERT fine-tuned with LoRA to detect trends and sentiment accurately in large-scale Twitter data. IndicGLUE is used to evaluate performance on monolingual and multilingual texts. This approach ensures fast, scalable, and accurate analysis of social media discussions.

Keywords: Apache HBase, SparkNLP, PySpark MLlib, Codemixed (Hindi+English) Tweets, IndicBERT, LoRA

1. Problem Statement and Objectives

With the rise of Hindi-English code-mixed content on Twitter, understanding trends and sentiment in such posts has become challenging. Users frequently switch between Hindi and English, use different spellings, and write in Romanized Hindi, making traditional NLP models ineffective.

For example:
a. *"Election results aa gaye! Modi ji ki speech dekhi?"*
b. *"Cricket ka match bohot intense tha! Kohli ne kya performance diya!"*

Challenges include:

1. **Code-Mixing Complexity** – Tweets contain varying proportions of Hindi and English, making language processing difficult.
2. **Transliteration Issues** – Words like *"jeet liya"* and *"jit gaye"* mean the same but are written differently, affecting tokenization and meaning extraction.
3. **Spelling Variations** – No standard way to write names (e.g., *"modiji"* vs. *"Modi ji"*), impacting search accuracy.
4. **Search Limitations** – Traditional keyword-based methods (BM25) fail to capture meaning variations in mixed-language text, requiring BM25 for better results.

This project builds a Big Data-powered pipeline to:

a. Preprocess and clean large-scale tweets using Apache HBase for data storage, Spark NLP for text preprocessing, and PySpark MLlib for handling missing values, feature scaling, and vectorization.
b. Index and search tweets efficiently using BM25.
c. Detects trends and sentiment using LLMs (IndicBERT fine-tuned with LoRA).

d. Ensure accurate results at scale for multilingual social media analysis.
e. Evaluate model performance using IndicGLUE for monolingual and multilingual accuracy.

2. Methodology

Firstly, there are two biggest reasons for using Big Data Technologies over scalable data solutions

a. Massive Data Volume & Real-Time Processing

Twitter generates millions of tweets daily, requiring distributed storage (Hadoop HDFS, HBase) and real-time processing (Spark NLP and PySpark MLLib) to detect trends instantly. Traditional scalable databases fail under such high data velocity and volume.

b. Advanced Search & Semantic Understanding

BM25 enables trend detection across linguistically diverse, code-mixed tweets. Traditional keyword-based search cannot link phonetically different but semantically similar tweets, making it ineffective for multilingual trend detection.

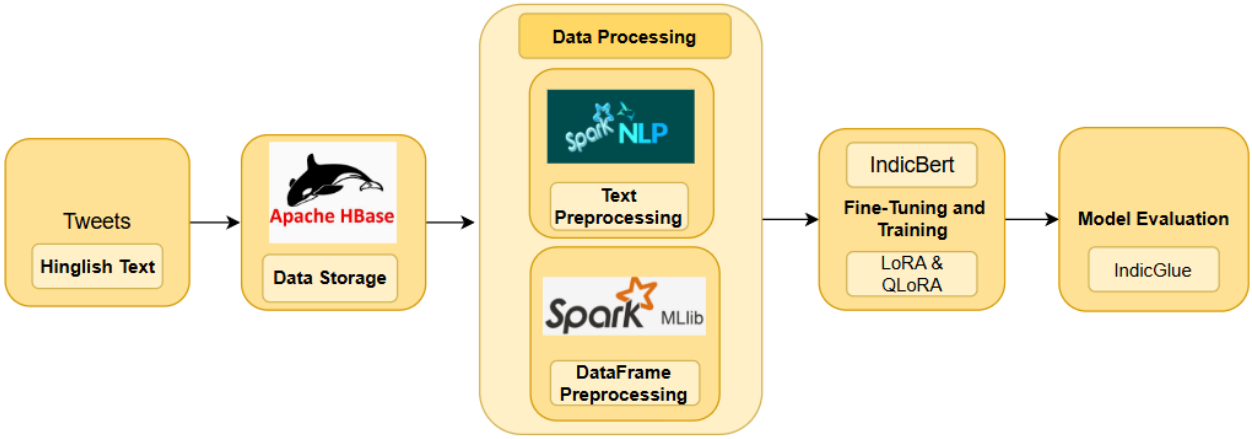
2.1. About Dataset

The dataset used for this project is sourced from the **AI4Bharat Twitter dataset** (https://huggingface.co/datasets/ai4bharat/Mann-ki-Baat), which contains large-scale Hindi-English code-mixed tweets. We are also exploring additional data sources and planning to integrate them into our custom database. The current dataset provides real-world examples of language switching, transliteration variations, and sentiment-rich content, making it ideal for training and evaluating trend and sentiment analysis models. It comprises tweets collected over time across diverse topics such as politics, sports, entertainment, and social issues. The dataset's linguistic diversity, script variations, and sentiment polarity present a challenging yet valuable resource for accurate trend detection and sentiment analysis.

2.2. Technologies Used

a. **Apache HBase –** Stores and manages structured tweet metadata for efficient retrieval.
b. **Spark NLP –** Preprocesses tweets by handling tokenization, normalization, and transliteration of Romanized Hindi.
c. **PySpark MLlib** – Performs large-scale feature extraction, vectorization, and transformation of structured tweet data for efficient modeling.
d. **BM25 –** Indexes tweets and improves keyword-based search ranking.
e. **IndicBERT (LoRA Fine-Tuned) –** Detects trends and classifies sentiment in multilingual content.
f. **IndicGLUE –** Evaluates model performance across monolingual and multilingual datasets.

2.3. Flow



2.3.1 Data Collection & Storage

The dataset used in this project comes from AI4Bharat, which provides a large-scale multilingual dataset including Hindi-English code-mixed tweets. The data is gathered from the Twitter API, which streams real-time tweets covering various topics such as politics, sports, and public sentiment. Since these tweets contain diverse linguistic structures, spelling variations, and transliteration differences, they require specialized handling for accurate analysis.

Once collected, the data is stored in Apache HBase, which provides a structured format for tweet metadata, enabling fast retrieval and scalable storage. To manage high-velocity data streams, Hadoop HDFS is used as a distributed storage system, ensuring efficient handling of massive datasets.

2.3.2 Data Preprocessing

Since the dataset consists of code-mixed text, extensive preprocessing is required to prepare it for trend prediction and sentiment analysis. Spark NLP is used for tokenization, normalization, and transliteration to standardize mixed-script text, ensuring consistency in spelling variations. PySpark MLlib acts as a scalable alternative to traditional frameworks like Pandas and NumPy by handling large-scale data transformations and vectorization. It enables feature extraction, stopword removal, and TF-IDF computation, converting text into structured numerical representations suitable for machine learning tasks.

For example, a tweet like "Election results aa gaye! Modi ji ki speech dekhi?" undergoes Spark NLP preprocessing, where it is first tokenized into individual words, then transliterated into a standardized Hindi format: "इलेक्शन रिजल्ट्स आ गए! मोदी जी की स्पीच देखी?". PySpark MLlib then processes the cleaned text, removing stopwords such as "ki" and "ji", extracting key features, and applying TF-IDF vectorization to transform the words into numerical embeddings. These embeddings are then used for sentiment classification and trend detection, ensuring that the model can efficiently analyze social media discussions in real-time.

2.3.3 Indexing & Searching

Once preprocessed, the data is indexed using BM25, which enables efficient retrieval of relevant tweets by applying a probabilistic ranking function. BM25 ensures that keyword-based searches prioritize the most relevant discussions by considering term frequency, document length, and inverse document frequency. This method enhances information retrieval by ranking tweets based on their contextual relevance, improving trend detection and sentiment analysis in code-mixed social media data.

2.3.4 Trend Prediction & Sentiment Analysis

After indexing, trend prediction and sentiment classification are performed using IndicBERT fine-tuned with LoRA. The fine-tuned model is trained on the preprocessed dataset, allowing it to accurately detect trending topics and classify sentiment as positive, negative, or neutral.

For instance:

a. *"Petrol price badh gaya, bohot dikkat ho rahi hai!"* → Negative Sentiment
b. *"Kohli ne shandar batting ki!"* → Positive Sentiment

The trend detection module identifies emerging discussions and evolving topics over time, providing real-time insights into public sentiment and trending conversations.

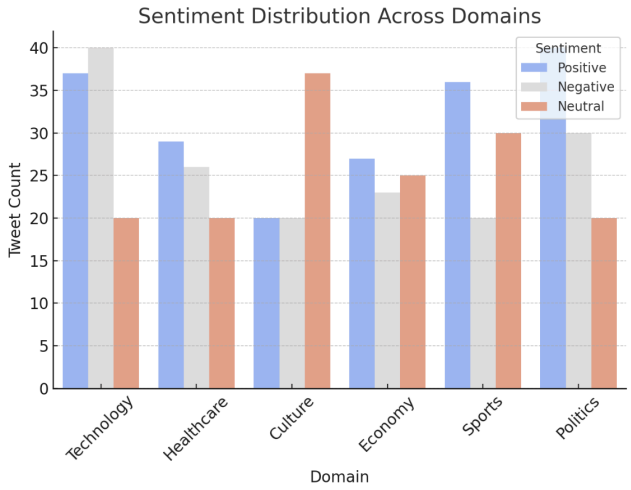
2.3.5 Model Evaluation

To ensure the model's accuracy and generalizability, IndicGLUE is used for evaluation. It benchmarks performance across monolingual and multilingual datasets, ensuring robustness in handling Hindi-English code-mixed content. Evaluation metrics such as accuracy, recall, and F1-score validate the effectiveness of the system in detecting trends and analyzing sentiment.

3. Analysis

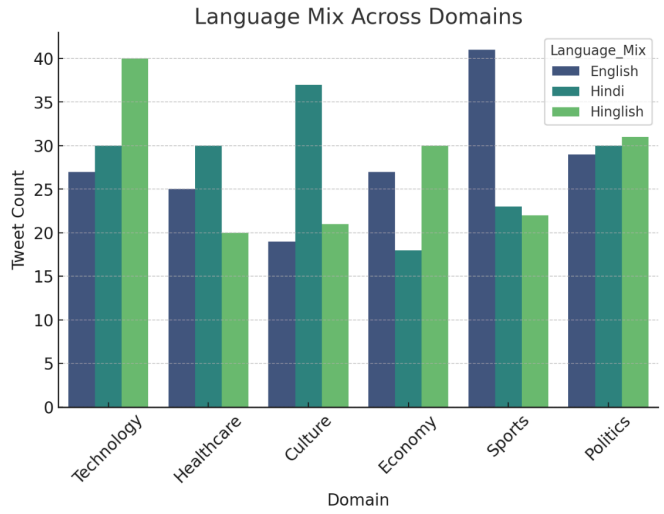
3.1. Sentiment Distribution Across Domains

a. The number of positive, negative, and neutral tweets across different subdomains (Politics, Healthcare, Technology, Culture, Economy, Sports).
b. Also it helps us see which topics receive more criticism and which ones have a positive perception.



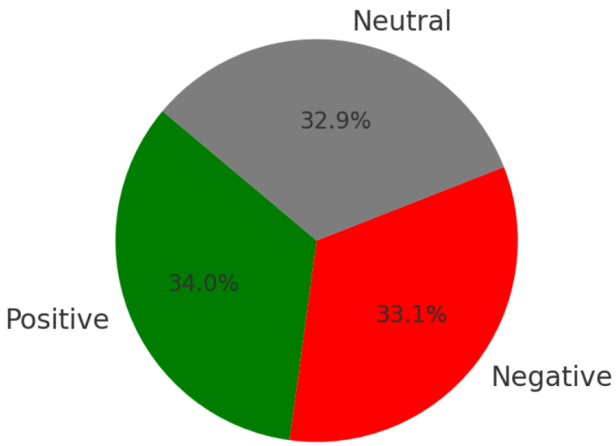
3.2. Language Mix Across Domains

a. Some domains (e.g., Politics & Technology) might have more Hinglish tweets, while others (e.g., Healthcare & Sports) may have more English.
b. This helps in training better models for code-mixed data processing as it is commonly used in customer interactions.



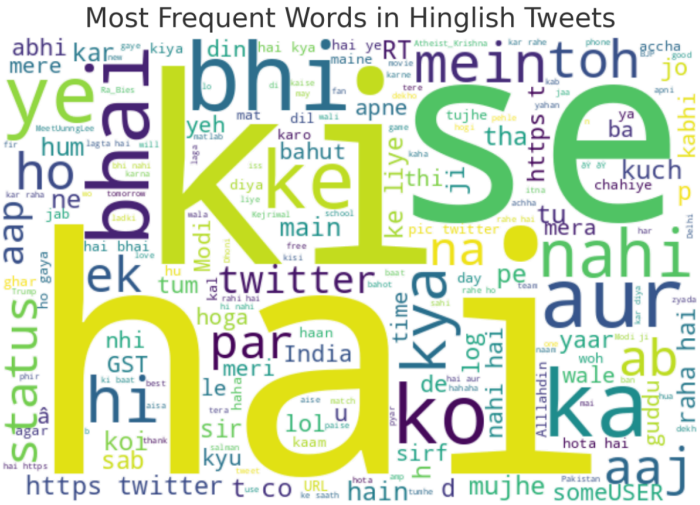
3.3. Sentiment Distribution in Code-Mixed Data

a. The proportion of positive, negative, and neutral sentiment in all tweets [Green = Positive, Red = Negative, Gray = Neutral].
b. Helps in detecting trends in social media conversations.



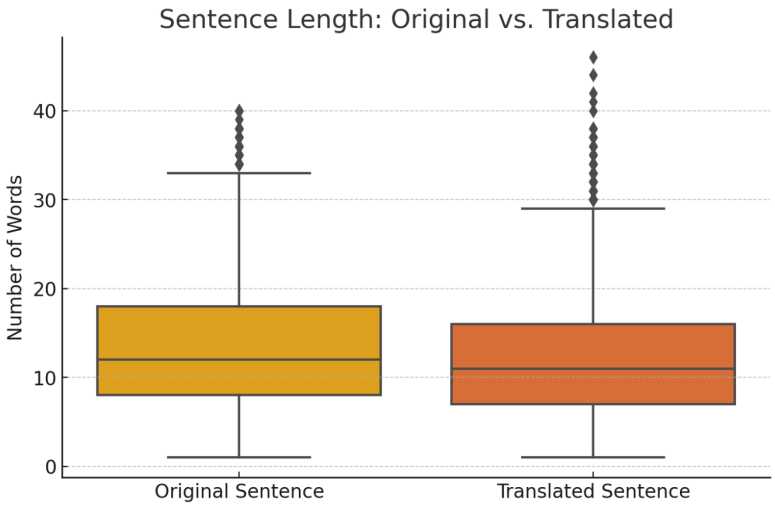
3.4. Most Frequent Words in Hinglish Tweets

a. Bigger words appear more frequently, showing popular terms and topics.
b. Helps in identifying trends and hot topics in Hinglish conversations and useful for enhancing search engine algorithms and social media monitoring tools.



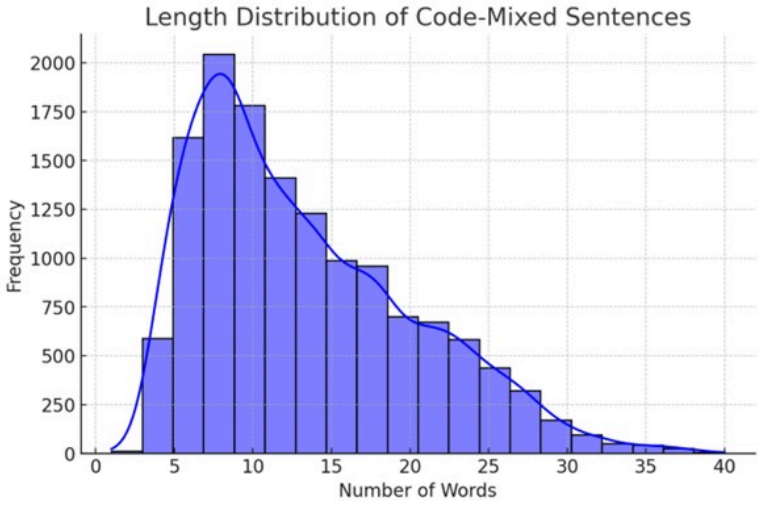
3.5. Sentence Length: Original vs. Translated

a. Detects whether translation accuracy is compromised when switching between languages.
b. It is useful for automating Hinglish-to-English translations and shows if translations expand or shorten sentences.



3.6. Length Distribution of Code-Mixed Sentences

a. The distribution of sentence lengths in Hinglish tweets.
b. Helps determine whether users write shorter or longer mixed-language messages.
c. Peaks in the graph show the most common tweet lengths and thus useful for text summarization.



4. Future Scope

4.1. Expansion to Multiple Social Media Platforms – The system can be extended beyond Twitter to analyze trends and sentiment on Facebook, Instagram, and YouTube, providing broader insights into multilingual social media discussions.

4.2. Expansion to More Code-Mixed Languages – The system can be enhanced to support other Indian languages like Tamil, Bengali, and Marathi, improving trend detection and sentiment analysis across diverse linguistic datasets.