Social Media Data Mining using Hadoop Framework

Student Name : Bathula Veera Mahesh

Roll Number : 22JK1A0525

College Name : KITS AKSHAR INSTITUTE OF TECHNOLOGY

Domain Name : Data Science And Big Data Analysis

Title of the project : Social Media Data Mining using Hadoop Framework

Submitted to : Black Bucks

Date Of Submission : 17/07/2025

Abstract: Social Media Data Mining Using Hadoop Framework

In today's digital age, social media platforms such as Twitter, Facebook, and Instagram have become significant sources of user-generated data. With billions of posts, interactions, and engagements occurring daily, the potential to extract meaningful insights from this data is immense. However, the sheer volume, velocity, and variety of social media content pose challenges for traditional data processing systems. To address this issue, the "Social Media Data Mining Using Hadoop Framework" project presents a scalable and efficient Big Data solution designed to collect, store, process, and analyze large-scale social media content using the robust Hadoop ecosystem.

Problem Statement & Overview

Organizations and researchers increasingly rely on real-time trends, behavioral analytics, and sentiment insights derived from social platforms to inform marketing strategies, brand reputation management, and socio-political forecasting. Yet, the majority of social media data is unstructured, complex, and difficult to process without powerful distributed systems. This project aims to overcome the limitations of conventional analytics by leveraging Hadoop-based distributed storage and processing for scalable data mining. The primary objective is to uncover patterns in user behavior, trending hashtags and mentions, sentiment distributions, influencer identification, and engagement metrics across multiple platforms.

Tools and Applications Used

The project utilizes various components of the Hadoop ecosystem and complementary tools to streamline the data mining pipeline:

- **HDFS (Hadoop Distributed File System):** For distributed storage of raw and processed social media data.
- **MapReduce (Java):** To extract tokens such as hashtags, mentions, and URLs through parallel text parsing.
- **Hive / Pig:** For structured querying, aggregation, and transformation of processed data.
- **Apache Flume / Kafka:** To ingest real-time data streams from APIs like Twitter Streaming API.
- **Sqoop:** For exporting analytical results to relational databases such as MySQL or PostgreSQL.
- **YARN:** To manage cluster resources and schedule jobs efficiently.
- **Python with TextBlob/NLTK/spaCy:** For performing sentiment analysis and optional topic modeling.
- **Visualization Tools:** Data insights are presented using dashboards in Tableau, Power BI, or Kibana for interactive exploration. **Submodules Description**

The project comprises several well-defined submodules:

1. **Data Ingestion Module:** Collects real-time social media feeds using Flume or Kafka and ingests them into HDFS.
2. **Preprocessing Module:** Cleanses raw data by removing noise, tokenizing texts, and extracting relevant features (hashtags, mentions, URLs).
3. **Mining Module (MapReduce):** Identifies trends, counts token frequencies, and extracts patterns.
4. **Querying and Aggregation Module (Hive):** Enables querying of engagement metrics, sentiment, and topic clusters.
5. **Sentiment Analysis Module (Python):** Analyzes post sentiments using NLP libraries for classification into positive, negative, or neutral.
6. **Export and Visualization Module:** Migrates results to external databases and visualizes them using BI dashboards.

Design Flow

The overall workflow initiates with API streaming into HDFS, followed by a MapReduce job for text parsing. Cleaned data is then stored in Hive tables for querying. Sentiment analysis is done in parallel with NLP scripts. Final reports and visualizations are generated using Tableau/Power BI. YARN coordinates all executions across the Hadoop cluster.

Conclusion / Expected Output

This project demonstrates how Big Data technologies can be effectively applied to mine actionable insights from unstructured social media data. The expected outputs include trending hashtag reports, sentiment breakdowns, user activity heatmaps, influencer rankings, and engagement statistics across platforms. These insights can drive strategic decisions in marketing, brand reputation, and public engagement.

By integrating Hadoop and supporting tools, this solution ensures scalability for both batch and real-time processing, flexibility in analysis, and robustness in dealing with diverse social media content.