**RevTrends**

**(Data Engineering Project)**

Center of Excellence

16-Aug-2023

Version 1.0

Copyright ©: 2023, Revature.

 Table of Contents

[Application Overview 3](#_Toc1555815592)

[Core Functional Requirements 3](#_Toc1194330873)

[Standard Functional Scope 3](#_Toc1763839213)

[Definition of Done 4](#_Toc2020375305)

[Competency-wise scoping 4](#_Toc1523109773)

[Non-Functional Expectations 6](#_Toc1436798500)

​​

​​

​

​

​

​

​

​​

​

​

​

​

​

​​

# Application Overview

RevTrends is a Social Media Sentiment Analysis and Trend Tracking Platform designed to provide comprehensive insights into public sentiment across various social media platforms. Leveraging real-time data streams, Trends aggregates, analyzes, and visualizes sentiment trends, enabling users to make informed decisions and gain a competitive edge.

# Core Functional Requirements

As a User, I want to:

**Data Mining:**

1. Access historical Twitter, Reddit, Facebook, and other social media data based on specified keywords so that I can.
2. Scrape relevant posts, tweets, and comments based on post IDs so I can analyze the sentiment on those posts.

**Data Aggregation:**

1. Aggregate social media data from multiple sources like Twitter, Facebook, and Reddit so I can have an integrated dataset.
2. Collect real-time streaming social media data to get the latest data.

**Data Analysis:**

1. Perform sentiment analysis on the text data to classify posts as positive, negative, or neutral.
2. Analyze trends in sentiment for ant text data like product name, person name, company name, etc. to track the perception over time.
3. Correlate changes in sentiment with marketing campaigns, product launches, and any other events to quantify impact.

**Data Visualization:**

1. View the charts describing the positive and negative trends for text data across multiple social medial platforms.
2. View the most trending topics and their negative and positive trends.

# Standard Functional Scope

1. Ingesting and processing various social media data feeds.
2. Transforming data into analytical models.
3. Providing interactive dashboards, queries, and reports for analysis.
4. Implementing administrative controls for user management and data governance.

This scope ensures efficient data handling, advanced analysis, and user-friendly tools, facilitating the seamless transformation of raw social media data into valuable insights like the sentiment trends towards a text.

# Definition of Done

1. Working application demonstration.
2. Sharing the associates’ code repo for technical review with:

* Architecture
* data models
* ETL documentation

# Competency wise scoping

|  |  |  |
| --- | --- | --- |
| **Competency** | **Project Type** | **Expectations** |
| Python, SQL  Scala, SQL | REST API service | **Framework Specific**   1. Ensure the appropriate APIs are used for any of the API calls. 2. Ensure the routing is centrally configured. 3. Best practices & design patterns are to be followed.     **Validation and Error Handling:**   1. Validate the inputs for their types and format. 2. Display functional-related user messages (either for input/error/output) - no system error codes or SQL error codes. 3. Handle the exceptions and errors gracefully.     **Logging:**   1. Ensure the application is using the proper logging framework and methods. 2. Ensure the application’s log level is configured using configuration files so that it can be changed without changing the code. 3. Also ensure that the application logging is configured to output to the mentioned log file.     **Testing**:   1. Ensure sufficient test cases are written using appropriate testing frameworks. 2. Ensure the code coverage close to be 80%     **Security**:   1. Ensure the SQL injection threat is taken care. 2. Ensure the CORS restriction is applied, if applicable. 3. Ensure that the secrets are stored as environment variables using secure credential storage.   **Coding Standard:**   1. Use the industry coding standards and conventions. 2. Modular based code development for better reusability. 3. Ensure proper usage of resource objects such as database connectivity objects to avoid resource leakages. 4. Ensure proper usage of design patterns and application layering (such as Business Service, DAO Layer etc.) wherever applicable. |
| Spark  Hadoop, Hive | Data Science | **Data Preprocessing:**   1. Load the dataset from a reliable source. 2. Handle missing data, outliers, and any data quality issues appropriately. 3. Perform data cleansing, normalization, and transformation to prepare it for analysis.   **Trend Tracking and Analysis:**   1. Build algorithms to analyze sentiment trends over user-specified time intervals for specific keywords. 2. Implement statistical methods to identify significant sentiment shifts and patterns. 3. Provide interactive visualizations to display sentiment trends graphically.   **Event Impact Assessment:**   1. Develop a module to correlate sentiment changes with events, such as marketing campaigns and product launches. 2. Design mechanisms to quantify the impact of events on sentiment, including statistical measures and visual representations.   **Data Visualization:**   1. Design interactive visualizations (line charts, bar charts, pie charts, etc.) to represent coal production trends, energy source distribution, and other relevant insights. 2. Ensure the visualizations are user-friendly, allowing users to explore and interact with the data to gain deeper understanding.   **Reporting and Presentation:**   1. Generate detailed reports summarizing the findings from the data analysis and mining processes. 2. Create presentations with clear explanations of insights, trends, and correlations for stakeholders' understanding.   **User Interaction and Exploration:**   1. Provide user interfaces or dashboards that allow stakeholders to interact with the visualizations and customize views based on their interests. 2. Enable filtering, sorting, and comparison functionalities within the visualizations to facilitate data exploration. |
| Azure  AWS  GCP | ETL Pipeline | **Deployment artifacts:**   1. The deployment artifacts should be minified and obfuscated if required.   **Pipeline Orchestration**   1. Scheduled ETL jobs monitor, orchestrate and restart flows. 2. Job monitoring and alerts for data quality and SLAs 3. Scalable compute resources based on workload.   **Scalability and Performance:**   1. Design the ETL pipeline to handle large volumes of data efficiently. 2. Optimize data processing and transformation to minimize execution time.   **Monitoring and Alerts:**   1. Set up monitoring for ETL job status, performance metrics, and data quality. 2. Configure alerts to notify stakeholders about failures or anomalies in the ETL process. |

# Non-Functional Expectations

* Application development should use version control systems (e.g., Git) to manage the project codebase and facilitate collaboration.
* Application development is supposed to follow the Scrum process.

# Source Data Location

The data can be acquired from Kaggle using:

* Link to the csv files [Technology - Social Media Sentiment - All Documents (sharepoint.com)](https://revature0.sharepoint.com/sites/trainers/Shared%20Documents/Forms/AllItems.aspx?id=%2Fsites%2Ftrainers%2FShared%20Documents%2FCenter%20of%20Excellence%2FProject%20Operations%2FData%20Engineering%2FData%20Source%2FSocial%20Media%20Sentiment&viewid=d59398ae%2Dbc86%2D4d4e%2D9ff3%2Dc9cd2bfaea22)