Project Proposal

Name:- Joshi Jay Vijaybhai
Student id:- 200485155
Project Name:- AzureSent: A Sentiment Analysis Using Azure and Bing News API

Table of Contents

Topic	Page No.
Introduction ● The Main Idea of the Project ● Research Problem ● Objective	2
Background ● Previous Related Projects ● Relevant Research Papers ● Existing Techniques Relevant to the Project ● Learning Resources ● Advantages and Disadvantages of Techniques	3
Approach ● DATA ● METHOD ● EXAMPLE	6
Schedule	9
References	10

Introduction

The Main Idea of the Project:

In this project, we aim to create a news analytics application which does sentiment analysis on real-time news articles. Using Microsoft's Bing News Search API, Azure services and the Microsoft Fabric to measure news sentiment every day. The results will be visualized on a live Power BI dashboard, exposing patterns in news and public sentiment. The tech aims to aid organizations, companies and the media with tracking public opinion to use it for strategic decision-making, crisis management or planning.

Research Problem:

It is a project to solve real-time the problem of sentiment analysis across news sources. It solves a problem of fetching news data by an API, and analyzing the data to get insights on how the public is feeling about any particular thing. The primary research challenge is to design a pipeline which should be capable of: News Data Ingestion;Data Transformation and Sentiment Classification on Data in Real-Time

Objective:

In essence, the goal of this project is to construct a data pipeline that can accurately predict sentiment over news data retrieved via Bing News Search API. This includes use of data engineering techniques and machine learning models (such as Synapse ML model) to gauge sentiment out of news articles.

- **Input:** The input will be the daily news articles from Bing News Search API stored in Microsoft Data Lake. These articles are from Politics, Business, Technology along with the rest.
- **Output:** Results will be presented as a PowerBi dynamic dashboard highlighting daily changes in sentiment, summary tables for key metrics, and visual analytics (eg. graphs). We will set up an alert system so we can be notified via Microsoft Teams when sentiment goes above or below a threshold.
- **Usefulness:** Organizations will be able to monitor the pulse of public opinion on issues of the day in real-time for decision-making, crisis response, and strategic planning.

Background

Previous Related Projects:

During the CS714 (big data) course, I worked on a project that requires analyzing you-tube data collected using AWS. We were required to build an ETL pipeline and design a visualization in AWS QuickSight for the data gathered. This project itself was missing any alerting system or machine learning models for data analysis unlike my current one. In this project, we are planning on building a functional Sentiment Analysis model and real-time notifications in Microsoft Teams using machine learning. On top of that, I have barely touched the Azure ecosystem before so this will help me to broaden my cloud computing expertise into worlds outside of AWS.

Relevant Research Papers:

These are five important papers as of 2018–2024 that are related to sentiment analysis in real time especially for news articles by both machine learning and data engineering.

- **Sentiment Analysis Using Deep Learning Techniques: A Comprehensive Review[1]**
  This Paper shows some of the difficulties with the use of deep learning in sentiment analysis are large well-labeled datasets and models that generalize to different domains. For instance, it goes into details about the kinds of models used in sentiment analysis and focuses on ways such as regularization and data augmentation to avoid overfitting while being critical if we are to analyze news articles across different subjects in a real-time scenario.
- **Real-Time Sentiment Analysis of Natural Language Using Multimedia Input[2]**
  This study is based on the proposal of a model regarding real-time sentiment analysis of textual and multimedia data. Considering the integration of text, audio, and video data for the sentiments of the authors in this research will inspire new avenues towards news-given that a sentiment analysis system would work on systems that operate on multimedia data inside the news, be it video or podcast format.
- **Challenges and Future in Deep Learning for Sentiment Analysis: A Comprehensive Review and a Proposed Novel Hybrid Approach[3]**
  This paper will specifically dwell on multimodal sentiment analysis by using the integration of vision, audio, and textual features that are derived from user-generated videos. Using various sources of data such as social media, video news, and textual articles in building a real-time news analytic tool will support this with more robust and accurate sentiment analysis.
- **An Analysis of Sentiment: Methods, Applications, and Challenges[4]**
  This paper discusses the various approaches to sentiment analysis with an emphasis on machine learning and their hybridized techniques. It addresses pertinent issues like feature extraction, classification strategies, and domain adaptability, which are convoluted but crucial for sentiment analysis of heterogeneous news topics.

- **Enhanced Video Analytics for Sentiment Analysis Based on Fusing Textual, Auditory, and Visual Information[5]**
  This research clarifies the advantages of combining textual, auditory, and visual sources for the task of emotion detection. The form of emotion detection studied for this research was video analytics. Still, the general knowledge about fusion that can inform and expand emotion detection functionality beyond text, as applied here to include guidelines for using visual reporting sources.

This research is a timely review of important developments, challenges, and trends in sentiment analysis including the ongoing research that expands the scope of sentiment analysis to include different modalities like embedded social media content and images, both which are crucial in creating an overall, real-time application for news sentiment analysis.

Existing Techniques Relevant to the Project:

- **API Integration:** The Bing News Search API will be employed to ingest fresh news articles into the solution on a daily basis. The ingestion of data with existing expression, working with APIs as a data engineer working on data engineering projects.
- **Data Ingestion and Transformation:** The raw news articles will all be consumed data and stored in Azure Data Lake, and I've developed a pipeline to do incremental data load through Azure Pipelines, using PySpark for data transformation.
- **Sentiment Analysis:** Synapse ML models will be leveraged to perform sentiment analysis on the ingested news data to label articles as positive, negative, or neutral based on their emotional tone.
- **Reporting and Dashboard:** Power BI will be used as the visualization tool to get the dashboard of sentiment trends and insights in a manner that refreshes on an ongoing daily basis of ingestion of the new news data.
- **Alerting System:** There will be an alerting system using teams to flag when the sentiment of new articles surpasses thresholds for alerting (for example, when the articles have high negative or high positive sentiment using Data Activator).

Learning Resources:

In order to gain a solid understanding of the skills and knowledge base that arises from implementation of these techniques, I will refer:

- Official Azure documentation.
- Guides for using Microsoft Fabric.
- API documentation for guidance on integration.
- Previous coursework on cloud computing and machine learning.

Advantages and Disadvantages of Techniques:

API Integration:

• Advantage: Enables real-time data ingestion; simple access to a variety of news sources.

• Disadvantage: Dependent upon stable and reachable API; possible rate limits can postpone data ingestion.

Data Ingestion and Transformation:

• Advantage: Azure Data Lake provides performant storage and scalable data processing; PySpark offers opprobrious data transformation opportunities with its extensive libraries.

• Disadvantage: Steeper learning curve when assessing Azure services; complex pipeline can be a challenge to manage.

Sentiment Analysis:

• Advantage: Synapse ML models provide advanced machine learning capabilities for sentiment analysis; finds the patterns between news articles.

• Disadvantage: Models are biased; can be influenced by the quality of underlying data.

Reporting and Dashboarding:

• Advantage: Power BI offers easy to use visualization tools; offers dashboards active reports for real-time data insight.

• Disadvantage: Data sources need to be maintained on a continual basis to ensure accuracy; or when source complexity increases visualization the benefit will diminish in exchange for Institutional knowledge.

Alerting System:

• Advantage: Near real-time notification of a shift in sentiment increases responsiveness; integrates with Microsoft Teams in the program.

• Disadvantage: Alert fatigue can happen if a threshold is not determined; close to near real-time data flow or current sentiment enabled alerts.

Approach

DATA:

Data sources for this project come from the Bing News Search API. The Bing News Search API has granted us access to real news stories across categories falling within news-related themes and genres, such as politics, tech, entertainment, and so forth. For each news article, metadata include its title, description, date published, and URL; all these are ingested into the platform. It is a well-documented, trusted source which can be integrated without any hitch into Azure services and, therefore, is a sound source for the continuous ingestion of news data.

METHOD:

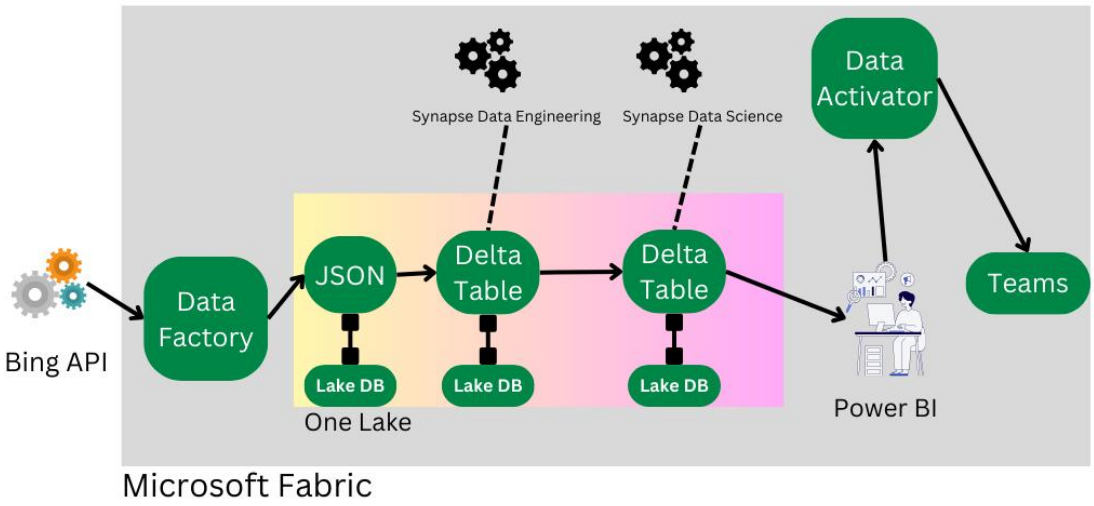


Fig:-Project Architecture

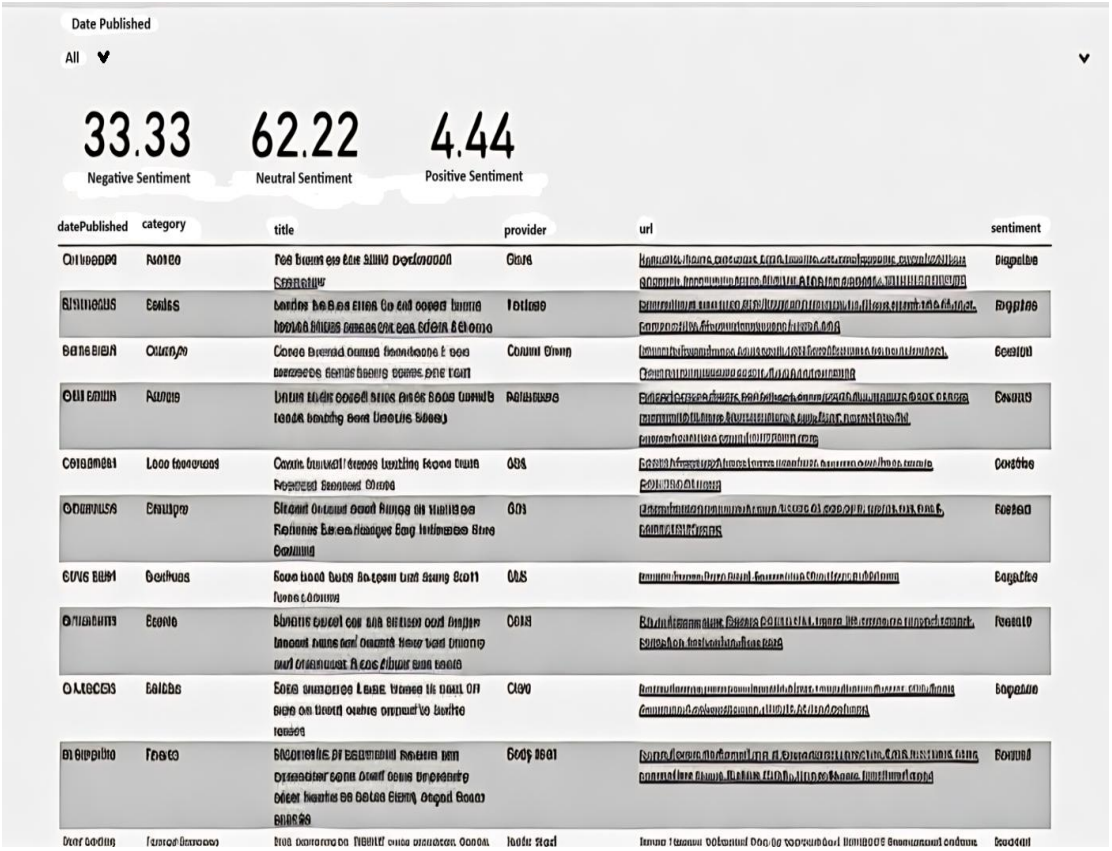
The project will be executed through several processing stages, leveraging Microsoft Fabric's capabilities and integrating various Azure tools to achieve the goal of real-time sentiment analysis for news articles. The method consists of the following stages:

- **API Setup and Configuration:** We will set up the Bing News Search API in Azure that can tap real-time data of news. It will deliver fresh news articles every day and will be one of the very important pieces of the data source for the project.

- **Data Ingestion Using Microsoft Fabric:** We will connect Microsoft Fabric's data tool to the API so that we can pull data into it. The API is full of data in JSON file, which we'll copy and store to OneLake, a data storage service by Fabric. OneLake has so far enabled the creation of many databases, and one of them is Lake Database, which we'll use here to store raw JSON files.
- **Data Transformation:** Then, raw JSON will be transformed and structured to a proper format. The transformation in question will then be carried out using Synapse Data Engineering; that is by creating Spark Notebooks that will clean the JSON data into the tabular format. Clean data will then be stored within the Lake Database, in a pre-defined schema Delta Table. Incremental loading will also be included in the process; this will mean new data loading continuously and without extra overhead.
- **Sentiment Analysis:** The cleaned and formatted data will then be run for sentiment analysis through Synapse Data Science. It will thereby derive the emotional tone of the news articles or classify them as positive, negative, or neutral. This processed data with sentiment scores will again be warehoused in the form of a Delta Table in Lake Database.
- **Dashboard Creation Using Power BI:** From this analysis, I will build a comprehensive news sentiment dashboard using Power BI. This shall include depicting the key insights for the customer, such as general trends over time news sentiment, top stories, and even the distribution of sentiment to different categories.
- **Alerting System Using Data Activator:** The system will create alerting using the Data Activator, configured to be able to monitor the mood thresholds over time and to produce alerts in real-time using Microsoft Teams. It is illustrative to send out signals at times when overall sentiment of news articles exceeds certain predefined thresholds, highly negative or positive.

This approach relies on integration of data engineering, data science, and visualization tools in Microsoft Fabric with the result in a robust pipeline for real-time sentiment analysis and reporting over news data.

EXAMPLE:



Above is just an example of what I would show as a dashboard that I will provide at the completion of the project, with alert notifications through Teams; however, this remains a conceptual example, and actual product may vary.

Schedule

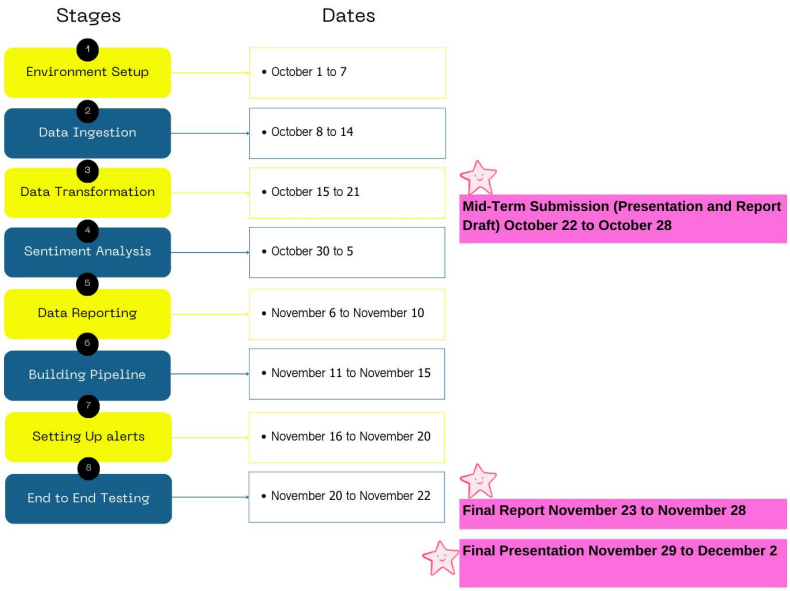


Fig:-Schedule

We have broken the project into eight manageable stages, as indicated by the graph above. There are also three special events, indicated by stars: mid-term submission, final report and presentation. These will be what we need to reserve dedicated time for in preparation. All through the project, it is our intention to stick as closely to this schedule as possible, to ensure that all these things get done on the expected time frames and within the planned framework.

References

- Sahoo C, Wankhade M, Singh BK. Sentiment analysis using deep learning techniques: a comprehensive review. International Journal of Multimedia Information Retrieval. 2023 Dec;12(2):41.[1]

- Jain R, Rai RS, Jain S, Ahluwalia R, Gupta J. Real time sentiment analysis of natural language using multimedia input. Multimedia Tools and Applications. 2023 Nov;82(26):41021-36.[2]

- Islam MS, Kabir MN, Ghani NA, Zamli KZ, Zulkifli NS, Rahman MM, Moni MA. Challenges and future in deep learning for sentiment analysis: a comprehensive review and a proposed novel hybrid approach. Artificial Intelligence Review. 2024 Mar 5;57(3):62.[3]

- Sharma HD, Goyal P. An Analysis of Sentiment: Methods, Applications, and Challenges. Engineering Proceedings. 2023 Dec 19;59(1):68.[4]

- Al-Azani S, El-Alfy ES. Enhanced video analytics for sentiment analysis based on fusing textual, auditory and visual information. IEEE Access. 2020 Jul 27;8:136843-57.[5]