**Social Media Sentiment Analysis Report (NLP)- Reddit**

# Project Report

Group Name: Group 2 Data Analysis Internship Duration: 3 months

Submitted By: Nandni, Garima Mishra, Subhashree Samantac, Taufik Shaikh Company Name: Infotact Solution

# Introduction

Sentiment analysis is a widely used Natural Language Processing (NLP) technique that helps in extracting subjective information from textual data.

With the rise of social platforms like Reddit, users actively share their opinions on diverse topics such as machine learning.

This project focuses on analyzing sentiments in Reddit posts related to Machine Learning using two well-known NLP tools: VADER and TextBlob.

The objective is to understand the sentiment distribution, visualize key insights, and prepare the dataset for further NLP tasks such as topic modeling and sentiment classification.

# Project Overview

The goal of this project is to perform sentiment analysis on Reddit posts related to Machine Learning. We started by collecting Reddit posts data containing metadata such as titles, body text, scores, number of comments, and author details.

Sentiment features were generated using VADER and TextBlob, which provided compound sentiment scores and polarity values. The data was cleaned, visualized, and analyzed to uncover insights into how the Reddit community perceives discussions around Machine Learning.

# Dataset Description

The dataset consists of 829 posts with 12 columns after cleaning. - `title`: Post title describing the main subject. - `selftext`: Full body text of the post. - `score`: Reddit score calculated as upvotes minus downvotes. - `num\_comments`:

Number of comments associated with the post. - `created`: Date and time when the post was created (converted from UNIX timestamp). - `author`: Reddit username of the poster. - `vader\_compound`, `vader\_positive`, `vader\_negative`, `vader\_neutral`: Sentiment scores generated using VADER. - `textblob\_polarity`, `textblob\_subjectivity`: Sentiment metrics generated using TextBlob.

The dataset was carefully cleaned by handling missing values, removing duplicates, and formatting dates.

# Tools and Technologies Used:

The project was implemented using the following tools and technologies: - \*\*Python\*\*: Primary programming language for data analysis and NLP tasks. - \*\*Pandas\*\*: Used for data cleaning and manipulation. - \*\*Matplotlib\*\*: Employed for visualization of sentiment distributions and trends.

\*\*WordCloud\*\*: Used to generate word clouds from Reddit post titles and texts. - \*\*VADER (Valence Aware Dictionary for Sentiment Reasoning)\*\*: Lexicon-based sentiment analysis tool optimized for social media. - \*\*TextBlob\*\*: Python library that provides sentiment polarity and subjectivity analysis. - \*\*Jupyter Notebook\*\*: Environment for performing experiments and creating reproducible workflows.

# Insights:

From the analysis of the dataset, several insights were drawn: - The majority of Reddit posts had neutral to slightly positive sentiments. - Posts with higher scores (upvotes) tended to lean more positively in sentiment. - TextBlob polarity and VADER compound sentiment scores showed a positive correlation, though not perfect, suggesting that combining multiple tools adds robustness to sentiment analysis.

Weekly sentiment trends revealed spikes in discussions around major machine learning announcements, conferences, or research breakthroughs. - Word clouds highlighted frequent terms such as 'model', 'data', 'training', 'AI', and 'research'.

# Visualization:

The project leveraged multiple visualization techniques to interpret sentiment patterns within Reddit posts. These visualizations not only provided statistical insights but also offered a more intuitive understanding of how sentiment evolves over time and across different authors. Key visualizations included:

**1**.**Distribution of Post Scores**

Highlighted that most Reddit posts achieved moderate engagement while a few outliers reached exceptionally high upvotes, reflecting community-driven popularity of certain discussions.

**2**. **Distribution of VADER Compound Sentiment**

Displayed clustering around neutral sentiment, with a slight skew toward positive sentiment, suggesting a generally constructive tone in machine learning conversations.

**3**. **Scatterplot of VADER Compound vs TextBlob Polarity**

Showed a positive correlation between the two sentiment analysis tools but also highlighted differences in edge cases, proving that multi-tool sentiment analysis is more robust than relying on a single model.

**4**. **Time Series Sentiment Trend**

Weekly and monthly sentiment trends reflected spikes during periods when major machine learning research, conferences, or technology announcements were made. This indicates that community sentiment is event-driven and highly reactive to external developments.

**5. Word Cloud of Titles and Selftext**

Provided an at-a-glance overview of the most frequently used terms. Prominent words like “model”, “training”, “data”, “research”, and “AI” reaffirmed the dataset’s relevance to machine learning discussions.

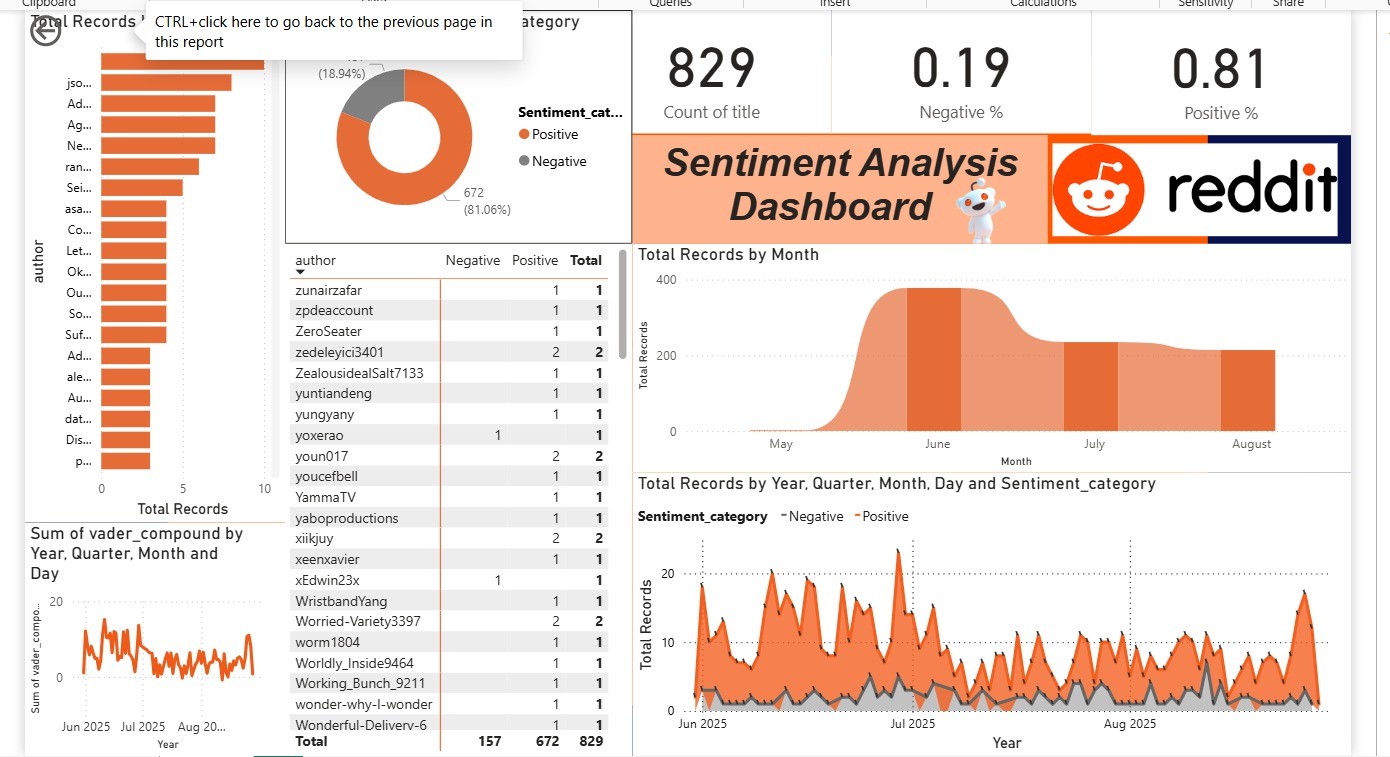
**6.Power BI Dashboard**

To provide a comprehensive summary, a Power BI interactive dashboard was created, consolidating key metrics and trends: - The dashboard revealed that \*\*81% of posts were positive\*\* while \*\*19% were negative\*\*. –

A timeline visualization showed sentiment fluctuations across June to August, capturing both daily and monthly variations. - Author-wise breakdown highlighted frequent contributors, allowing us to see which users drive sentiment-heavy discussions.

Line and area charts emphasized the dynamic nature of sentiment, with visible peaks corresponding to community events and trending topics. –

Overall, the dashboard acted as a central visualization hub, making insights more interactive and digestible for stakeholders. Below is the Power BI dashboard screenshot embedded for reference:



# Future Scope

The current project provides a foundation for sentiment analysis on Reddit posts, but there are several directions for future work: - Extend the dataset by including more subreddits beyond machine learning, such as data science, deep learning, or AI ethics.

Incorporate advanced NLP models like BERT or GPT for more accurate sentiment classification. - Perform topic modeling to identify the most frequently discussed themes and connect them with sentiment trends. - Develop a real-time dashboard that monitors Reddit discussions and updates sentiment insights continuously.

Explore causal relationships between external events (e.g., publication of a major research paper) and shifts in community sentiment.

# Conclusion

This project successfully demonstrated the application of sentiment analysis to Reddit posts related to Machine Learning. By using VADER and TextBlob, we were able to evaluate the emotional tone of discussions and visualize patterns across posts, comments, and time.

Insights such as the predominance of neutral-positive sentiment, spikes in engagement during major events, and correlation between sentiment and post popularity provide a deeper understanding of the community’s perception of Machine Learning.

The work sets the stage for more advanced NLP applications such as topic modeling and deep learning-based sentiment classification.