Project Report

Reddit Post Classification Pipeline Summary

Introduction

This project details the creation of an end-to-end data analytics pipeline, transforming raw text from the Reddit API into actionable, visual insights. The primary objective was to collect data, perform advanced text preprocessing, classify post sentiment using machine learning techniques, and deploy the findings on an interactive Power BI dashboard. The resulting structured dataset, sentiment labels, and integrated dashboard explain posting behavior, sentiment trends, and keyword patterns across online communities.

Methodology & Implementation

Data Acquisition

- **Source:** Reddit's AskReddit community via PRAW API
- **Collection:** 500 recent posts using subreddit.new() and subreddit.hot()
- **Fields:** Title, body text, timestamp, subreddit, post ID
- **Output:** reddit_raw.csv for downstream processing

Data Preprocessing

Text Normalization Pipeline:

- Timestamp conversion from UNIX to datetime format
- Lowercasing and whitespace normalization
- URL and special character removal using regex
- Creation of cleaned columns: clean_title, clean_text

Sentiment Classification

- **Tool:** VADER (Valence Aware Dictionary for Sentiment Reasoning)
- **Scoring:** Compound score range [-1, 1]
- **Thresholds:**
    - Positive: ≥ 0.05
    - Negative: ≤ -0.05
    - Neutral: (-0.05, 0.05)
- **Output:** classified_reddit_data.csv with sentiment labels

Dashboard Development

Power BI Implementation:

1. **Visual 1: Sentiment Distribution** - Stacked bar chart showing post volume by sentiment
2. **Visual 2: Temporal Analysis** - Line chart tracking daily sentiment trends
3. **Visual 3: Keyword Analysis** - Word cloud of frequent discussion topics

Interactive Features:

- Sentiment filter (Positive/Negative/Neutral)
- Date range selection
- Cross-visual highlighting

Technical Architecture

Development Stack

Python Environment:

- pandas (data manipulation)

- PRAW (Reddit API integration)

- VADER-NLTK (sentiment analysis)

- regex (text cleaning)

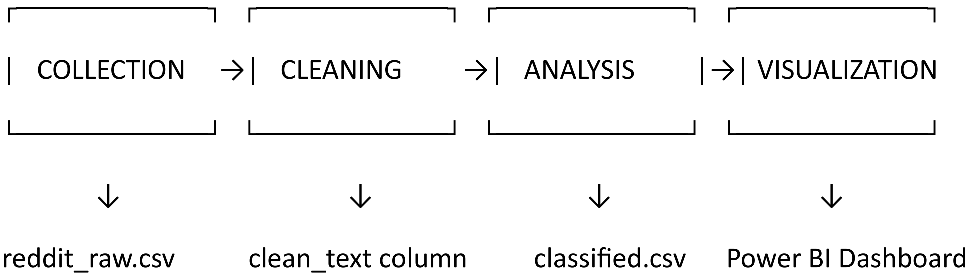
Visualization Platform:

- Microsoft Power BI Desktop

- Custom word cloud visual

- DAX measures for calculated fields

Data Pipeline



Challenges & Solutions

Challenge	**Solution Implemented**
API rate limiting	Controlled request pacing with delay intervals
Informal text patterns	Regex-based cleaning for URLs, emojis, slang
Sentiment ambiguity	VADER compound scoring with validated thresholds
Dashboard performance	Aggregated data views and optimized queries
Word cloud integration	Preprocessed word frequency table in Power Query

Results & Insights

Analytical Findings

- **Sentiment Distribution:** Quantitative breakdown of community emotional tone
- **Temporal Patterns:** Identification of posting frequency cycles and sentiment shifts
- **Content Themes:** Dominant discussion topics revealed through keyword frequency
- **Interactive Exploration:** User-driven analysis capability through filter controls

Business Value

- **Community Management:** Early detection of negative sentiment spikes
- **Content Strategy:** Data-driven understanding of engaging topics
- **Research Utility:** Foundation for social dynamics analysis
- **Technical Demonstration:** End-to-end data science pipeline implementation

Future Development Roadmap

Short-Term Enhancements

- Real-time data streaming integration
- Hybrid sentiment models (VADER + transformer ensembles)
- Mobile-responsive dashboard design

Medium-Term Objectives

- Multi-subreddit comparative analysis
- Topic modeling with BERTopic/LDA
- Automated report generation and alerting

Long-Term Vision

- Cross-platform sentiment comparison
- Predictive trend forecasting
- Production deployment with REST API

Conclusion

This project successfully demonstrates a complete analytics pipeline from social media data collection to interactive business intelligence. The implemented system provides practical sentiment insights while establishing an extensible foundation for advanced natural language processing applications. The modular architecture supports incremental enhancement, positioning the solution for both immediate utility and future development.