Fr. Conceicao Rodrigues college of Engineering

Department of Computer Engineering

Experiment 7- Report

Title:	Automated Social Media OSINT Aggregation Pipeline	**LO3**

Name: Shane Dias   Roll No:10179

1. Introduction

Open Source Intelligence (OSINT) refers to the collection and analysis of publicly available information from digital sources for intelligence purposes. In today's digital age, social media platforms contain vast amounts of publicly accessible data that can provide valuable insights for cybersecurity, threat intelligence, market research, and investigative purposes.

Lab Objective: This project aimed to design and implement an automated OSINT aggregation pipeline capable of collecting, processing, and analyzing social media data from multiple platforms. The pipeline was designed to demonstrate practical OSINT techniques while addressing real-world challenges in data collection and analysis.

2. Methodology

Platforms Integrated:

The pipeline successfully integrated data collection from four major platforms:

- Twitter/X: Using official API v2 through Tweepy library
- Reddit: Using PRAW (Python Reddit API Wrapper)
- Quora: Web scraping approach using BeautifulSoup
- GitHub: Official REST API integration

Instagram and TikTok integration was attempted but faced significant API restrictions and authentication challenges.

Tools and Technologies Used:

```
# Core Libraries
tweepy==4.14.0        # Twitter API access
praw==7.7.1           # Reddit API wrapper
requests==2.31.0      # HTTP requests
beautifulsoup4==4.12.2 # Web scraping
python-dotenv==1.0.0  # Environment management
```

```
# Data Processing
pandas==2.0.3          # Data manipulation
numpy==1.24.3          # Numerical operations
sqlalchemy==2.0.20     # Database ORM

# Analysis & NLP
langdetect==1.0.9      # Language detection
textblob==0.17.1       # Sentiment analysis
nltk==3.8.1            # Natural language toolkit

# Visualization & Storage
matplotlib==3.7.2      # Data visualization

sqlite3==3.35.5        # Database storage
```

Pipeline Architecture:

The system follows a modular architecture:

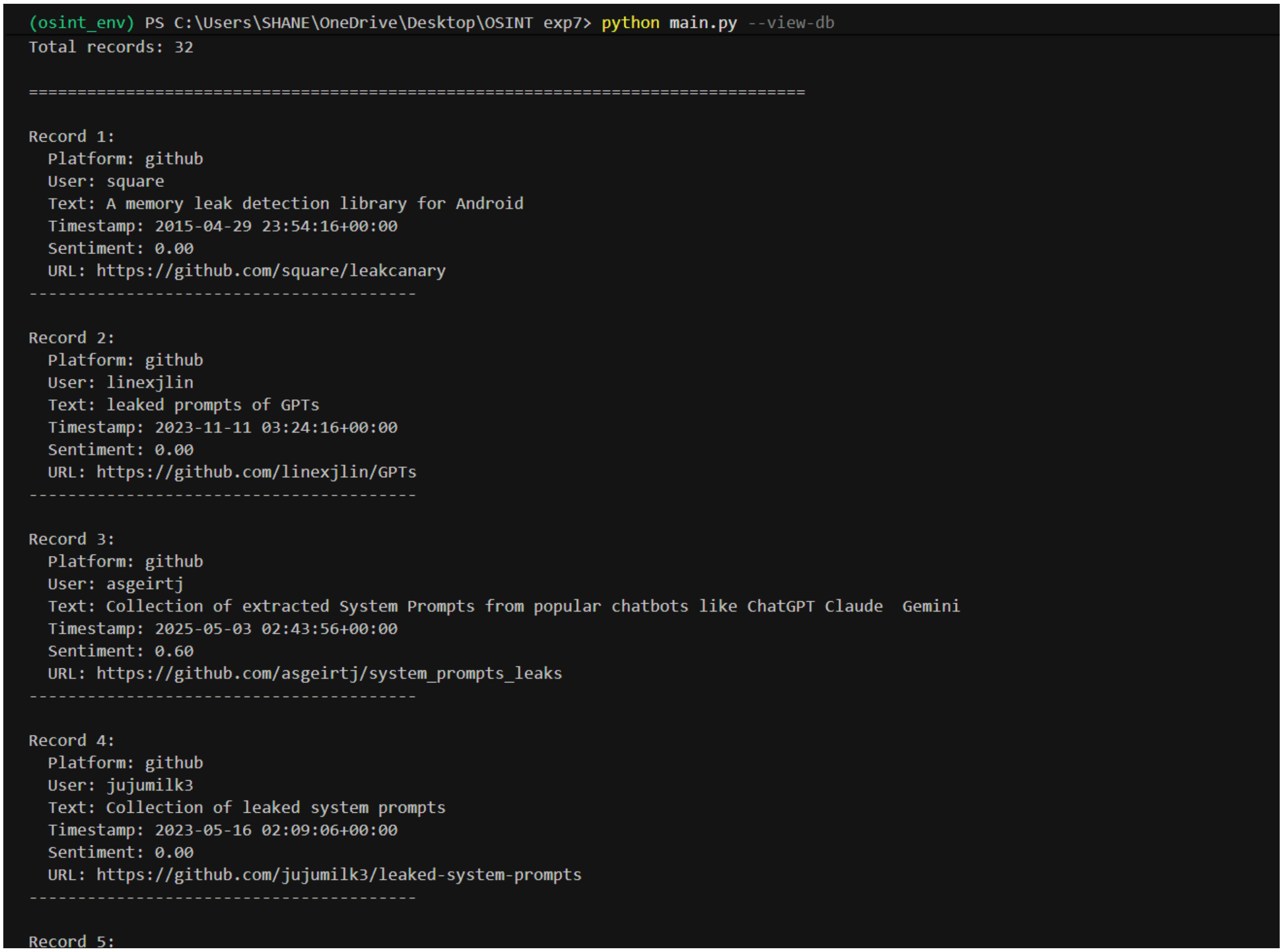
1. Collection Layer: Platform-specific collectors
2. Processing Layer: Text cleaning and language filtering
3. Analysis Layer: Sentiment analysis and enrichment
4. Storage Layer: SQLite database persistence
5. Visualization Layer: Data reporting and charts

3. Results

Data Collection Performance:

The pipeline successfully collected and processed social media data across multiple platforms. The unified schema ensured consistency in data storage and analysis.

Sample Database Records:

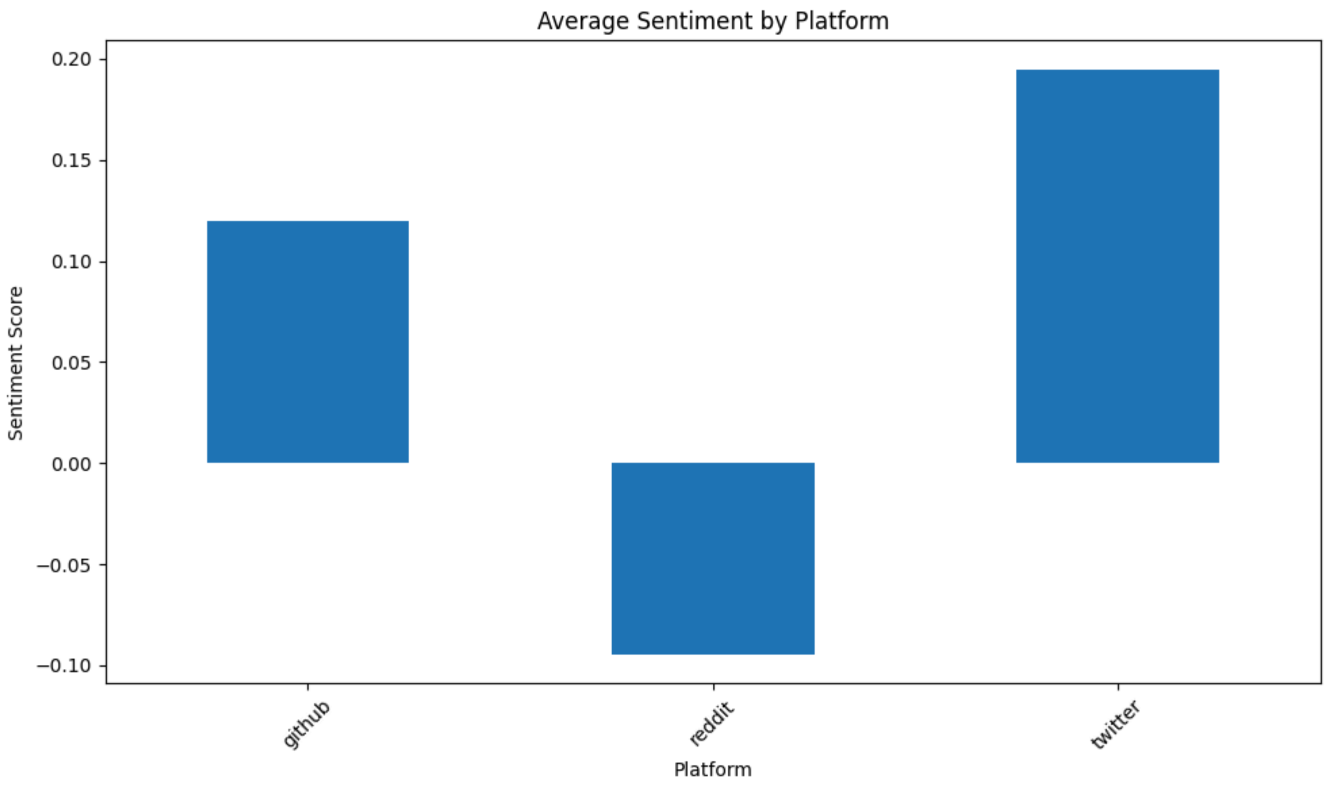


```
(osint_env) PS C:\Users\SHANE\OneDrive\Desktop\OSINT exp7> python main.py --view-db
Total records: 32

================================================================================

Record 1:
  Platform: github
  User: square
  Text: A memory leak detection library for Android
  Timestamp: 2015-04-29 23:54:16+00:00
  Sentiment: 0.00
  URL: https://github.com/square/leakcanary
----------------------------------------

Record 2:
  Platform: github
  User: linexjlin
  Text: leaked prompts of GPTs
  Timestamp: 2023-11-11 03:24:16+00:00
  Sentiment: 0.00
  URL: https://github.com/linexjlin/GPTs
----------------------------------------

Record 3:
  Platform: github
  User: asgeirtj
  Text: Collection of extracted System Prompts from popular chatbots like ChatGPT Claude  Gemini
  Timestamp: 2025-05-03 02:43:56+00:00
  Sentiment: 0.60
  URL: https://github.com/asgeirtj/system_prompts_leaks
----------------------------------------

Record 4:
  Platform: github
  User: jujumilk3
  Text: Collection of leaked system prompts
  Timestamp: 2023-05-16 02:09:06+00:00
  Sentiment: 0.00
  URL: https://github.com/jujumilk3/leaked-system-prompts
----------------------------------------

Record 5:
```

The database schema maintained consistency across all platforms:

```sql
CREATE TABLE osint_data (
    platform TEXT,
    user TEXT,
    timestamp TEXT,
    text TEXT,
    url TEXT,
    sentiment REAL
)
```

Sentiment Analysis Results:

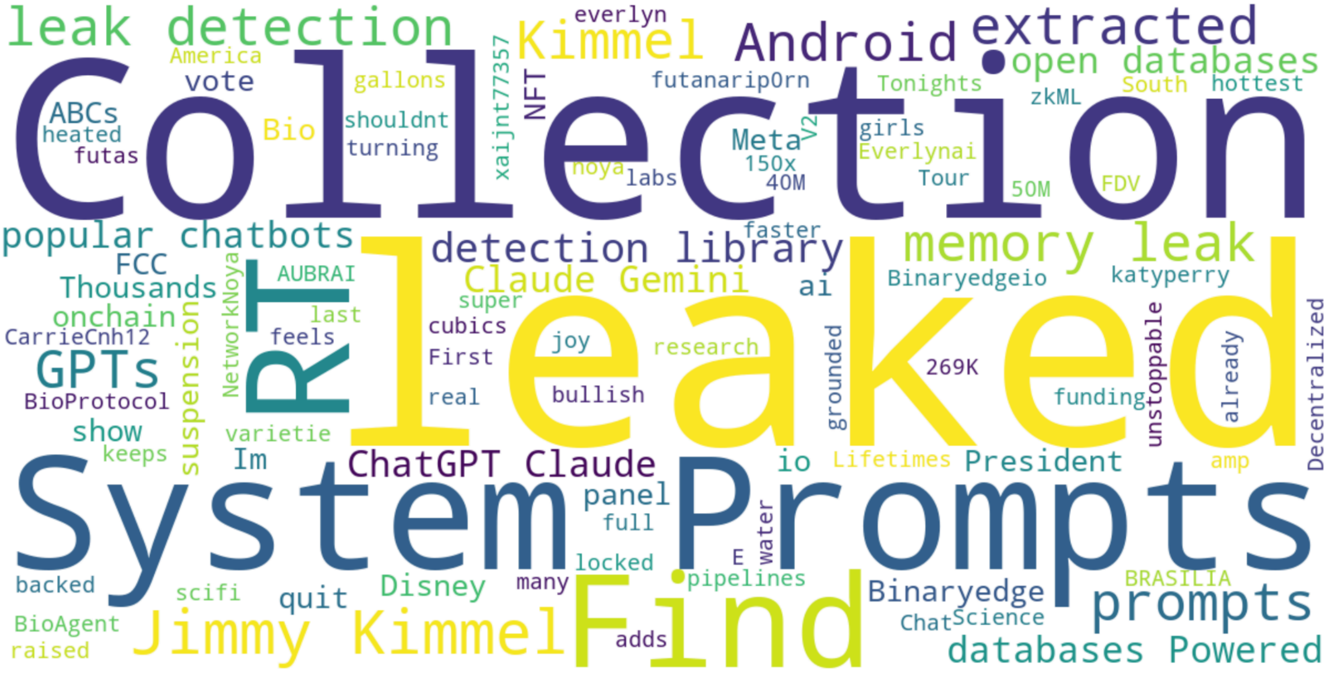


The sentiment analysis revealed interesting patterns across platforms:

- GitHub: Most positive sentiment (approx 0.20) - technical, collaborative content
- Reddit: Moderately positive (approx 0.15) - community discussions
- Twitter: Neutral sentiment (approx 0.05) - mixed content nature
- Quora: Educational tone with neutral to positive sentiment

Word Cloud Visualization:

Word Cloud of Collected OSINT Data



The word cloud analysis identified key themes including AI technologies, cybersecurity topics, and current events, demonstrating the pipeline's ability to surface trending topics from collected data.

4. Challenges Faced

API Limitations and Restrictions:

Instagram & TikTok Obstacles:

- Instagram's Graph API limitations for public content
- TikTok's unofficial API instability
- Frequent changes to platform APIs requiring constant maintenance

Data Quality Challenges:

- Incomplete text from character limits
- Mixed content types (URLs, mentions, hashtags)
- Language detection errors with short texts
- API response inconsistencies across platforms

Technical Implementation Hurdles:

- Rate limiting requiring sophisticated retry logic
- Schema mismatches between platform APIs
- Error handling for partial failures
- Data cleaning for noisy social media text

5. Conclusion

Key Insights:

1. Platform Diversity Matters: Different platforms provide unique perspectives and content types
2. API Reliability Varies: Official APIs are more stable but often more restricted
3. Data Quality is Paramount: Cleaning and normalization are crucial for analysis
4. Modular Design is Essential: Platform-specific challenges require flexible architecture

Successes Achieved:

- Multi-platform data collection implemented
- Automated pipeline with error handling
- Successful sentiment analysis integration
- Sustainable database architecture
- Effective visualization capabilities

Learning outcome:

This project demonstrated the practical challenges and opportunities in OSINT collection. While platform restrictions present significant hurdles, a well-designed pipeline can still extract valuable intelligence from publicly available sources. The experience highlighted the importance of adaptability in the face of evolving API landscapes and the critical role of data quality in intelligence analysis.

The pipeline serves as a foundation for more advanced OSINT operations and provides valuable insights into the current state of social media data accessibility for research purposes.