SentimentSift: AI-Powered Review Analysis Platform

7245 Big-Data

Anuj Rajendraprasad Nene Sicheng Bao Yung-Rou Ko

1. Introduction

Background

In today's digital marketplace, online reviews significantly influence consumer decisions. However, these reviews often contain emotional biases, exaggerations, or subjective content that can skew the overall perception of businesses. While Google reviews provide valuable consumer feedback, they don't differentiate between objective assessments and emotionally charged content, creating potential blind spots for both businesses and consumers seeking reliable information.

Objective

SentimentSift aims to develop an intelligent review analysis platform that uses advanced natural language processing and sentiment analysis to filter emotional content from Google reviews, providing users with more objective, factual, and balanced insights about businesses, products, and services.

2. Project Overview

Scope

- **Data Sources**: Google Reviews API for structured data, web scraping for additional unstructured review content from various platforms
- **Technologies**: Apache Airflow for ETL, FastAPI for backend services, Streamlit for frontend, LangChain/LangGraph for agent orchestration
- **Deliverables**: A fully functional web application that allows users to search for businesses, view filtered reviews, and gain objective insights through AI-powered analysis

Stakeholders

- **Primary Users**: Consumers making purchasing decisions
- **Secondary Users**: Business owners seeking accurate customer feedback
- **Tertiary Users**: Market researchers and analysts studying consumer behavior

3. Problem Statement

Current Challenges

1. **Emotional Bias**: Reviews often contain emotional language that can distort perception of actual quality
2. **Manipulation**: Some reviews may be inauthentic or manipulated to inflate or deflate ratings
3. **Inconsistency**: Review quality and helpfulness vary widely, making it difficult to extract reliable information
4. **Information Overload**: Users must read through numerous reviews to form their own assessment

Opportunities

1. **Objectivity Enhancement**: Provide more balanced, fact-based assessment of businesses
2. **Time Efficiency**: Save users time by automatically filtering and summarizing review content
3. **Business Insights**: Help businesses understand genuine customer concerns separated from emotional reactions
4. **Trust Building**: Create a more trustworthy information ecosystem for consumer decision-making

4. Methodology

Data Sources

1. **Structured Data**:
    - Google Places API: Primary source for business information and associated reviews
    - Public Review Datasets in CSV/JSON format: For initial model training and validation

- User feedback data stored in structured database tables

2. **Unstructured Data**:
    - Web Scraping Pipeline: Reviews from various platforms (Yelp, TripAdvisor)
    - PDF documents: Business reports, academic papers on sentiment analysis
    - Social media feeds: Unstructured text data from Twitter/X, Instagram comments

Technologies and Tools

1. **Data Collection & Processing**:
    - Apache Airflow for ETL pipeline orchestration and data preprocessing
    - Python (BeautifulSoup, Scrapy) for web scraping
    - MongoDB for storing unstructured review data (from web, PDFs)
    - PostgreSQL for structured business data (API responses)
2. **Backend Development**:
    - FastAPI for multiple RESTful API endpoints
    - Agent orchestration using Langraph, CrewAI, and Autogen frameworks
    - Integration with multiple LLMs (GPT-4, Claude, Llama) for sentiment analysis
    - MCP server integration for distributed processing
    - Docker for containerization
3. **Frontend Development**:
    - Streamlit for user interface and dashboard
    - Plotly and Altair for interactive data visualizations
    - Streamlit components for enhanced UI capabilities

Data Pipeline Design

1. **Data Ingestion**:
    - Scheduled API calls to Google Places API
    - Parallel web scraping jobs for supplementary review sources
    - Data validation and cleaning procedures
2. **Storage Layer**:
    - Distributed database architecture for scalability
    - Caching mechanisms for frequently accessed data
3. **Processing Layer**:
    - Review text preprocessing (tokenization, lemmatization)
    - Sentiment analysis and emotional content detection
    - Fact extraction and verification

Data Processing and Transformation

1. **Sentiment Analysis**: Identify and quantify emotional content in reviews
2. **Fact Extraction**: Use NLP techniques to extract objective statements from reviews
3. **Credibility Scoring**: Develop an algorithm to score review credibility based on various factors
4. **Summary Generation**: Create concise, objective summaries of multiple reviews

A proof of concept for basic sentiment analysis and emotional content detection will be implemented using Python scripts with sample data, demonstrating the feasibility of our approach.

5. Project Plan and Timeline

Milestones and Deliverables

1. **Project Setup and Initial Development (April 5-7)**

- Repository setup and project scaffolding
- Team roles assignment
- Data source API integration setup
- Initial Airflow pipeline configuration

2. **Data Collection and Processing (April 8-10)**
    - Implementation of web scraping components
    - Development of data processing pipelines
    - Initial sentiment analysis model integration
    - Data storage implementation
3. **Backend and Agent Implementation (April 11-13)**
    - FastAPI endpoint development
    - Agent orchestration with Langraph/CrewAI
    - LLM integration for sentiment analysis
    - MCP server setup
4. **Frontend and Integration (April 14-16)**
    - Streamlit application development
    - Data visualization implementation
    - API integration with frontend
    - Initial system testing
5. **Final Testing and Deployment (April 17-18)**
    - Comprehensive testing and debugging
    - Documentation completion

    - Cloud deployment
    - Final presentation preparation

Timeline

Due to the compressed timeframe (April 5-18), the team will employ an agile sprint methodology with daily stand-ups to ensure rapid progress and address any blockers immediately. A detailed task board will be maintained in GitHub Projects with daily updates and priority adjustments.

6. Resources and Team

Personnel

- **Team Member 1 (33.3%)**: Data Engineering Lead
    - Responsible for ETL pipeline development
    - API integration
    - Database management
- **Team Member 2 (33.3%)**: AI/ML Specialist
    - Sentiment analysis model development
    - LLM integration
    - Agent orchestration
- **Team Member 3 (33.3%)**: Full-Stack Developer
    - Frontend development
    - Backend API implementation
    - System integration

7. Risks and Mitigation Strategies

Identify Risks

1. **API Rate Limiting**: Google Places API has usage limits that could restrict data collection
2. **Model Accuracy**: Sentiment analysis and fact extraction models may have accuracy limitations
3. **Scalability Challenges**: Processing large volumes of reviews may create performance bottlenecks
4. **Legal/Ethical Concerns**: Web scraping and data usage may have legal implications

Mitigation Strategies

1. **Distributed Data Collection**: Implement rotating proxies and scheduled collection to avoid rate limits
2. **Model Ensemble Approach**: Use multiple models and validation techniques to improve accuracy
3. **Efficient Database Design**: Optimize storage and indexing for scalability
4. **Legal Compliance**: Ensure all data collection adheres to Terms of Service and applicable regulations

8. Expected Outcomes and Benefits

Measurable Goals and Testing Framework

1. **Accuracy Metrics**:
    - Achieve >85% accuracy in sentiment classification (measured using precision, recall, F1-score)
    - Implement A/B testing to compare filtered vs. unfiltered review insights
    - Use human evaluation panels to validate AI-filtered content quality
2. **Performance Testing**:
    - Process and analyze >10,000 reviews per hour

    - Stress testing of API endpoints with simulated traffic
    - Benchmark response times across different server configurations
3. **User Experience Evaluation**:
    - Achieve >80% user satisfaction in initial user testing
    - Implement usability testing protocols with defined task completion metrics
    - Track user engagement and retention metrics
4. **System Integration Testing**:
    - End-to-end testing of the entire pipeline from data collection to visualization
    - Component testing for each agent in the multi-agent system
    - Cross-compatibility testing across different devices and browsers

Expected Benefits

1. **Enhanced Decision Making**: Users can make more informed decisions based on objective information
2. **Time Savings**: Reduce the time needed to research businesses by 50%
3. **Business Improvement**: Provide businesses with clear, actionable feedback separated from emotional noise
4. **Market Analysis**: Enable better understanding of market trends through objective review analysis

9. Conclusion

The SentimentSift platform addresses a critical gap in online review systems by providing tools to filter emotional bias and extract objective information. By leveraging advanced AI techniques, our project aims to transform how consumers interact with and utilize online reviews, ultimately creating a more transparent and reliable information ecosystem for both consumers and businesses.

The project aligns perfectly with the course requirements, utilizing data and models as services, implementing a robust ETL pipeline, integrating multiple data sources, developing user-friendly interfaces, and orchestrating intelligent agents through modern frameworks. Our team is committed to delivering a high-quality solution that demonstrates practical applications of the concepts learned throughout the course.