Real-Time Market Sentiment Analysis and Retrieval System for Technology Companies

Applying Big Data and NLP to extract real-time sentiment insights from the technology industry news

Presented by: Group 2 (Yejin Moon, Shawn Sun, Rishabh Sethiya, Nikki Tai, Raunak Vikas Singh)

Content *overview*

1. Background & Business Use Case

The Challenge of Understanding Market Sentiment in Tech Media

The technology sector generates massive volumes of financial news daily. Investors, journalists, and researchers struggle to track changing public sentiment toward specific companies. Unstructured text data makes it difficult to extract trends or insights efficiently.

1. Background & Business Use Case

- **Problem Statement:**
  - Many investors find it difficult to stay informed about the constantly changing stock market due to overwhelming amount of scattered financial data.
  - Needs for a centralized, user-friendly platform that analyzes stock information to help users make informed decisions.
- **Solution:**
  - Build a real-time sentiment intelligence platform that analyzes large volumes of live unstructured text from news, social media, and reports.
  - By leveraging big data ETL and data warehousing to provide insights.
- **Business Impact:**
  - We hope this platform empowers users ranging from casual investors to researchers to make informed decisions, track market sentiment, and stay up-to-date with stock trends without deep technical expertise.

2. Data Source & Collection

1 Github(HF)

2 Python ELT

3 Snowflake

Primary Data Source:

- **FNSPID – Historical news dataset from multiple sources (23GB and condensed 5GB csv) (15.7 million news articles for 4,775 companies, from 1999 to 2023)**

Data Format and Access:

- **JSON format (headline, metadata, timestamp, source).**
- **Collected via Python `requests`**

Data Storage:

- **Cloud-native MPP warehouse for analytics, BI, and SQL workloads.**

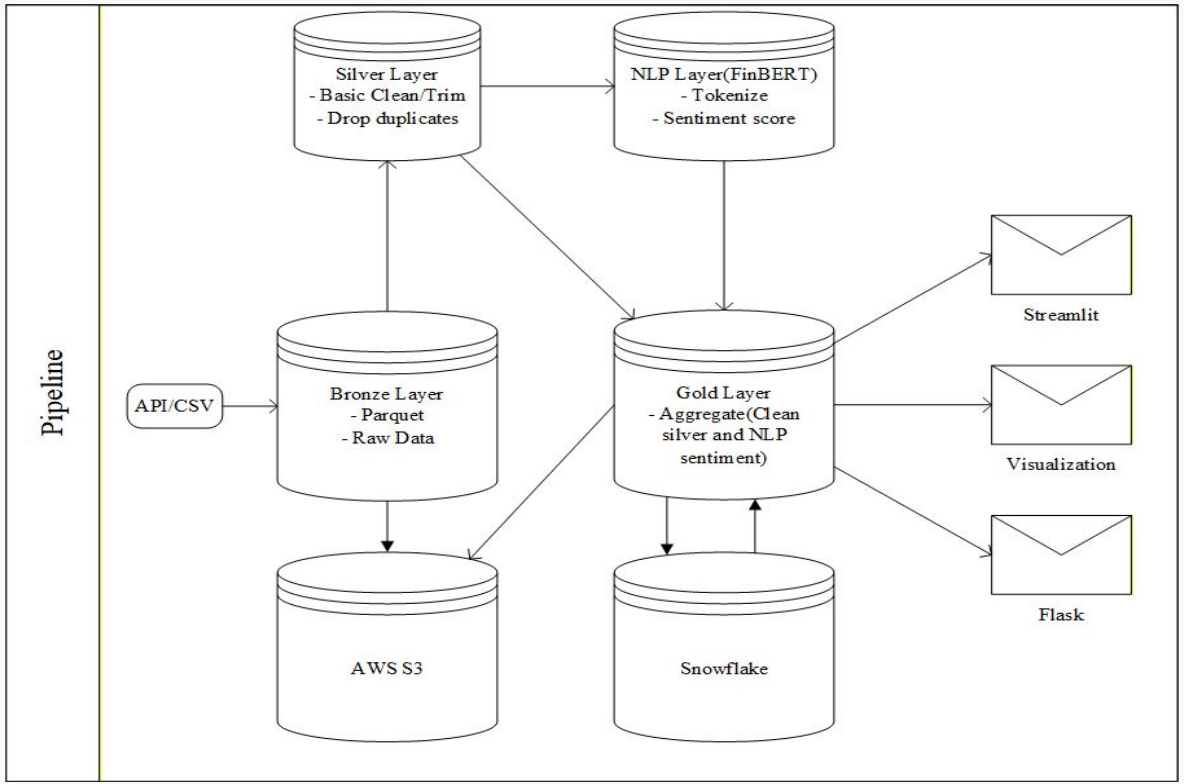
Dong, Z., Fan, X., & Peng, Z. (2024). FNSPID: A comprehensive financial news dataset in time series. arXiv. https://arxiv.org/abs/2402.06698
Github: https://github.com/Zdong104/FNSPID_Financial_News_Dataset?tab=readme-ov-file

3. System Design & Technologies

Layer	Technology	Rationale
1. Extraction	Python + Requests + Airflow	**Python** and **Airflow** are used for automated API data extraction and easy integration with cloud data sources. **Airflow** is used to schedule and monitor ELT jobs, ensuring the pipeline automatically runs daily.
2. Data Lake (Raw Storage Load)	Cloud Object Storage (AWS S3)	Works as a data lake and integrates with Spark.
3. Transformation	Apache Spark (PySpark)	Data cleaning and transformation are performed using distributed computing for scalability.
4. NLP Model	PyTorch + HuggingFace Transformers + FinBERT	Finance-domain language model outperforms generic lexicons.
5. Data Warehouse (Gold layer Storage)	Snowflake	Cloud-native MPP warehouse for analytics, BI, and SQL workloads.
6. User Interface	Streamlit to Flask	**Streamlit** provides web dashboard for real-time visualization of market sentiment trends, class distributions (positive/neutral/negative), and keyword analytics. **Flask** can provide huge amount live users.

3. System Design & Technologies

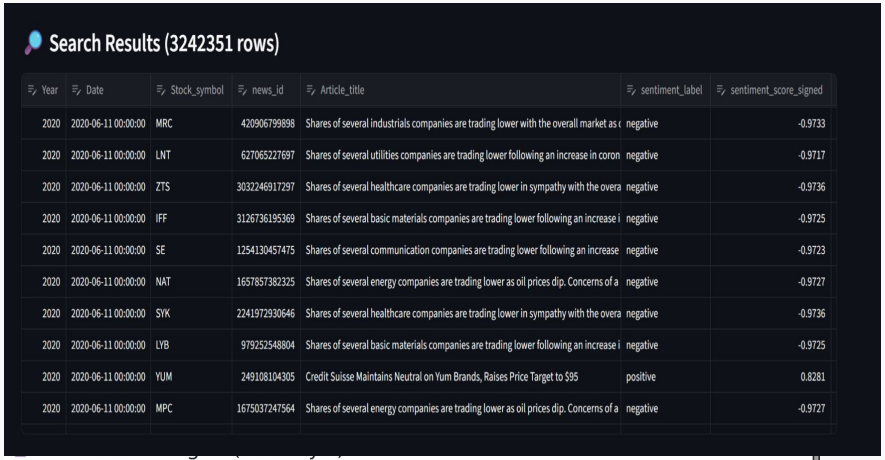
Workflow Design:

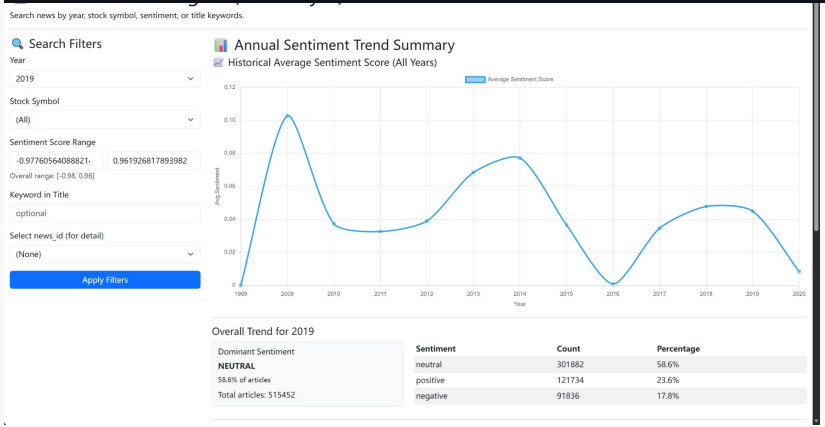


Data Quality:

- **Uniqueness**: used PySpark to drop duplicates.
- **Completeness**: filtered out rows with null headlines or timestamps.
- **Timeliness**: the system processes data daily.
- **Validity**: ensured sentiment scores are strictly between -1 and 1. (neutral: 0)

4. Product Demonstration(Steamlit & Flask)





License used: CC-by nc

5. Scalability & Costs:

Live users:

Streamlit: 500-1000
Flask: 12,000-20,000

Purpose	Key Drivers	Monthly Estimate	Annual Estimate
API	Paid News API	~$1,749	~$21,000
Data Processing (ETL)	EMR Large Cluster (x2idn.16xlarge) (Spark)	$6,000	$72,000
AI Compute	NLP GPU Instance (p4d.24xlarge)	$15,800	~$190,000
Analytics(Data warehouse)	Snowflake(Medium Multi-cluster 4 credit/hours)	~$5,000	~$60,000
Hosting(Streamlit,Flask)	AWS Fargate	$2,900	$34,000
Storage Infrastructure	S3 Data Lake(5-10TB)	~$236	~$2,832
GRAND TOTAL		**~$31.65k**	**~ $380K**

5. Scalability & Costs:

- Scalability

  For further public use, we could implement cloud computing, data lake for the parquet storage and cloud data warehouse to create dashboard, visualization, and web (UI).

- License

  For a commercial version, we would replace this dataset with a paid API like Bloomberg Terminal or Dow Jones Newswires.

- Future Work

  1. Integrate a trading bot to auto-trade based on sentiment.
  2. Develop website using flask for wider users.
  3. Add LLM-Based Company Extraction From Headlines and the content in news.
  4. Organized the stock symbol into specific categories for filter

Any questions?

Thank You