Archita Dhande

FinSight-AI Project: Detailed Report

1. Introduction

The FinSight-AI project is a machine learning-driven text classification system deployed using Google Cloud Vertex AI. The goal is to classify financial news headlines and make predictions based on the sentiment, i.e., whether the headlines reflect a positive or negative sentiment towards the stock market. Using a DistilBERT model, fine-tuned for financial news sentiment classification, the system processes text data to predict market movements, offering key insights for investors and business professionals.

2. Dataset Overview: Financial News Headlines Data

The Financial News Headlines Data used in this project consists of headlines scraped from prominent news sources like CNBC, The Guardian, and Reuters, focused on U.S. businesses and the overall economy. This dataset spans the period from March 2018 to July 2020 and provides a snapshot of the U.S. stock market and economic conditions.

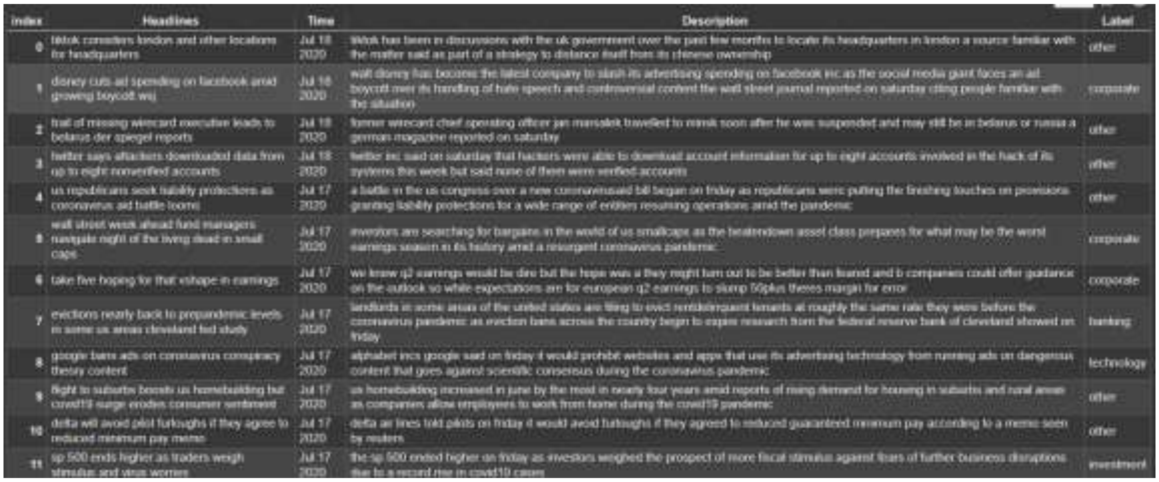
- *Source & Timeframe:*
    - *Reuters: Headlines, last updated date, and preview text of articles from March 2018 to July 2020.*
    - *Content: Includes a range of business-related topics such as market trends, corporate activities, economic policies, and more, covering stock performance, regulations, company earnings, GDP, and inflation.*
- *Inspiration:*
    - *The project investigates how sentiment in financial news impacts stock market performance. By applying Natural Language Processing (NLP) techniques, the primary goal is to explore whether positive or negative sentiment in headlines correlates with daily market movements.*

3. Modeling Process: DistilBERT for Financial News Sentiment Classification

The project uses DistilBERT, an optimized, smaller version of the BERT model, to classify sentiment (positive or negative) in financial news headlines.

- *Data Preprocessing:*
    - *Headlines are labeled with sentiment (positive = 0, negative = 1).*
    - *The text was tokenized using DistilBertTokenizer, converting the raw text into input features suitable for model processing.*
    - *Tokenization padded and truncated text to a maximum length of 32 tokens.*

- *Model Setup:*
    - *The DistilBertForSequenceClassification model was used for binary classification (Positive or Negative).*
    - *The model was deployed using TensorFlow and trained on the dataset with the Adam optimizer.*
- *Training & Evaluation:*
    - *Mixed Precision Training was used to scale gradients during backpropagation.*
    - *The model achieved 100% accuracy on the validation dataset, indicating excellent classification performance (though further testing on larger datasets is recommended).*



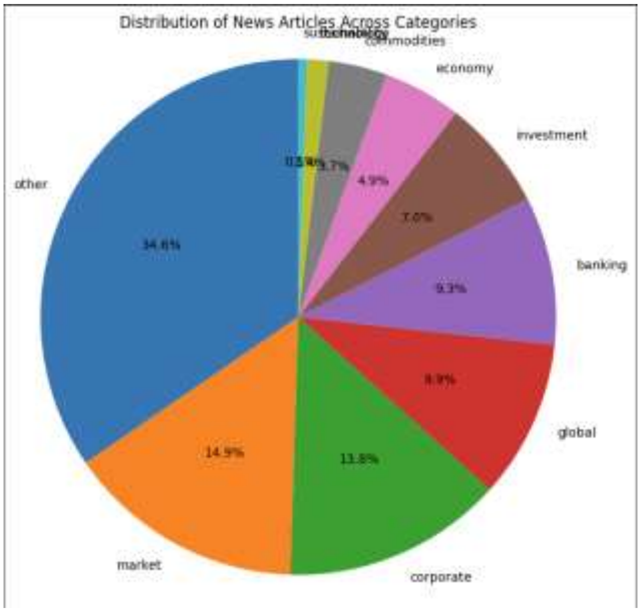
index	Headlines	Time	Description	Label
0	tiktok considers london and other locations for headquarters	Jul 18 2020	tiktok has been in discussions with the uk government over the past few months to locate its headquarters in london a source familiar with the matter said as part of a strategy to distance itself from its chinese ownership	other
1	disney cuts ad spending on facebook amid growing boycott wsj	Jul 18 2020	walt disney has become the latest company to slash its advertising spending on facebook inc as the social media giant faces an ad boycott over its handling of hate speech and controversial content the wall street journal reported on saturday citing people familiar with the situation	corporate
2	trail of missing wirecard executive leads to belarus der spiegel reports	Jul 18 2020	former wirecard chief operating officer jan marsalek travelled to minsk soon after he was suspended and may still be in belarus or russia a german magazine reported on saturday	other
3	twitter says attackers downloaded data from up to eight nonverified accounts	Jul 18 2020	twitter inc said on saturday that hackers were able to download account information for up to eight accounts involved in the hack of its systems this week but said none of them were verified accounts	other
4	us republicans seek liability protections as coronavirus aid battle looms	Jul 17 2020	a battle in the us congress over a new coronavirusaid bill began on friday as republicans were putting the finishing touches on provisions granting liability protections for a wide range of entities resuming operations amid the pandemic	other
5	wall street week ahead fund managers navigate night of the living dead in small caps	Jul 17 2020	investors are searching for bargains in the world of us smallcaps as the beatendown asset class prepares for what may be the worst earnings season in its history amid a resurgent coronavirus pandemic	corporate
6	take five hoping for that vshape in earnings	Jul 17 2020	we know q2 earnings would be dire but the hope was a they might turn out to be better than feared and b companies could offer guidance on the outlook so while expectations are for european q2 earnings to slump 50plus theres margin for error	corporate
7	evictions nearly back to prepandemic levels in some us areas cleveland fed study	Jul 17 2020	landlords in some areas of the united states are filing to evict rentdelinquent tenants at roughly the same rate they were before the coronavirus pandemic as eviction bans across the country begin to expire research from the federal reserve bank of cleveland showed on friday	banking
8	google bans ads on coronavirus conspiracy theory content	Jul 17 2020	alphabet incs google said on friday it would prohibit websites and apps that use its advertising technology from running ads on dangerous content that goes against scientific consensus during the coronavirus pandemic	technology
9	flight to suburbs boosts us homebuilding but covid19 surge erodes consumer sentiment	Jul 17 2020	us homebuilding increased in june by the most in nearly four years amid reports of rising demand for housing in suburbs and rural areas as companies allow employees to work from home during the covid19 pandemic	other
10	delta will avoid pilot furloughs if they agree to reduced minimum pay memo	Jul 17 2020	delta air lines told pilots on friday it would avoid furloughs if they agreed to reduced guaranteed minimum pay according to a memo seen by reuters	other
11	sp 500 ends higher as traders weigh stimulus and virus worries	Jul 17 2020	the sp 500 ended higher on friday as investors weighed the prospect of more fiscal stimulus against fears of further business disruptions due to a record rise in covid19 cases	investment

---

4. Financial News Sentiment Analysis Insights

- *Sentiment Impact on Stock Market:*
    - *Positive headlines generally indicate investor confidence and bullish trends, while negative headlines may reflect bearish market sentiments.*
    - *Analyzing sentiment over time can provide insights into how the stock market reacts to specific news, such as corporate earnings reports, policy changes, and global events like trade wars or economic recessions.*
- *Categories for Classification: To provide a more granular analysis, the headlines were classified into various business-related categories:*
    - *Market: Keywords like "stock", "market", "trading".*
    - *Corporate: Keywords like "merger", "acquisition", "earnings".*
    - *Economy: Keywords like "GDP", "inflation", "recession".*

- *Banking: Keywords like "finance", "loans", "interest".*
- *Investment: Keywords like "fund", "ETF", "portfolio".*
- *Technology: Keywords like "fintech", "blockchain", "crypto".*
- *Commodities: Keywords like "oil", "gold", "commodity".*
- *Global: Keywords like "trade", "tariff", "agreement".*
- *Sustainability: Keywords like "ESG", "sustainable", "climate".*



5. Project Architecture

The FinSight-AI project leverages Google Cloud Platform (GCP) to deploy the sentiment analysis model using Vertex AI, enabling real-time predictions from financial news headlines.

- *High-Level Architecture:*
  - *Google Cloud Storage (GCS): Stores model files and training data.*
  - *Vertex AI: Hosts the BERT model and handles inference requests.*
  - *Cloud Functions (Optional): Handle events such as model updates or new data for inference.*
- *Deployment Flow:*

1. The model is trained locally and saved to Google Cloud Storage (gs://finsight-ai-bucket/bert_text_classifier).

2. The model is deployed to Vertex AI via a deployment script.

3. The model serves real-time inference requests in us-west4 (Region A).

4. *Increased usage leads to exceeding the Custom Model Serving CPUs quota in us-west4, impacting scaling.*

---

6. Key Technical Components

- *Vertex AI & Model Deployment:*
  - *Custom Model Serving: Used for deploying and serving predictions via cloud resources.*
  - *Model Container: Built with TensorFlow and optimized for inference tasks.*
  - *Autoscaling: Configured to scale based on the request volume but requires quota management to avoid resource exhaustion.*
- *Data Management:*
  - *Training Data: Stored in Google Cloud Storage, tokenized, and encoded before training.*
  - *Inference: Tokenized input text is processed through the model for real-time predictions.*

---

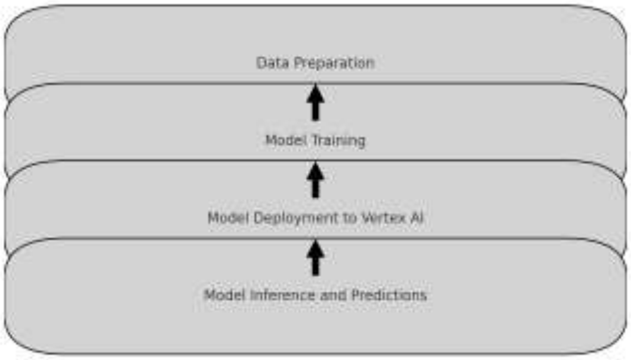
7. Challenges & Solutions

- *Quota Management Challenges:*
  - *The deployment in us-west4 region has exhausted its available CPU quota for Custom Model Serving, hindering further scaling.*
- *Suggested Solutions:*

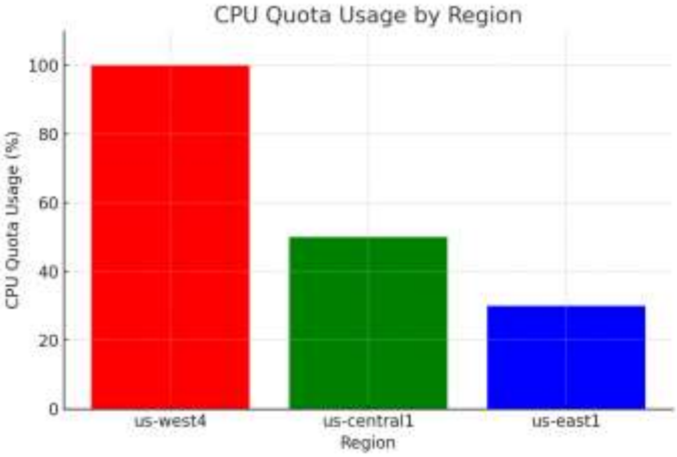
1. *Increase Quota: Request a quota increase for Custom Model Serving CPUs in us-west4.*
2. *Migrate to us-central1: This region offers more available CPU quota for serving models.*
3. *Optimize Resource Usage: Review and optimize the CPU allocation for the model. Implement autoscaling based on actual usage.*

---

8. Deployment Workflow


Data Preparation
Model Training
Model Deployment to Vertex AI
Model Inference and Predictions

CPU Quota Usage Visualization: A chart would show CPU usage across regions, highlighting the exceeded quota in us-west4.



9. Conclusion

The FinSight-AI project successfully deploys a BERT-based model using Vertex AI for real-time financial news sentiment classification. However, quota limitations in the us-west4 region have restricted scaling. Solutions such as increasing quotas, migrating to us-central1, and optimizing resource allocation are suggested to improve model deployment efficiency and scalability.