# Trading

**A**utomatic **M**edia **P**rocessing - **A**lgorithmic **M**arket **P**redictor (2-AMP)

Nic Dawson, Brody Curry, George Evans, Ryan Perry, Tanishq Somani

# The Vision


The United States stock market is a massive source of data, and arguably one of the most coveted in our modern day. Predictions, assumptions and calculated guesses are all used in order grow ones stock price and make a profit. Any advantage or ability to accurately predict these shifting markets can be very fruitful.

Our project seeks to make a tool to analyze this data and give a calculated decision on the stock's future, based on recent news and information, as modern media is often a huge player in shifting stock worth.

# The Logic

In order to do this, we aimed to train a classification model focusing on how stock prices fluctuate in response to news events. We planned to train the model using two datasets: historical stock prices and news articles related to the stocks in question.

Using these two sets of data, our model would be trained to recognize the relationship between the stock worth and the news about said stock.


# The Data

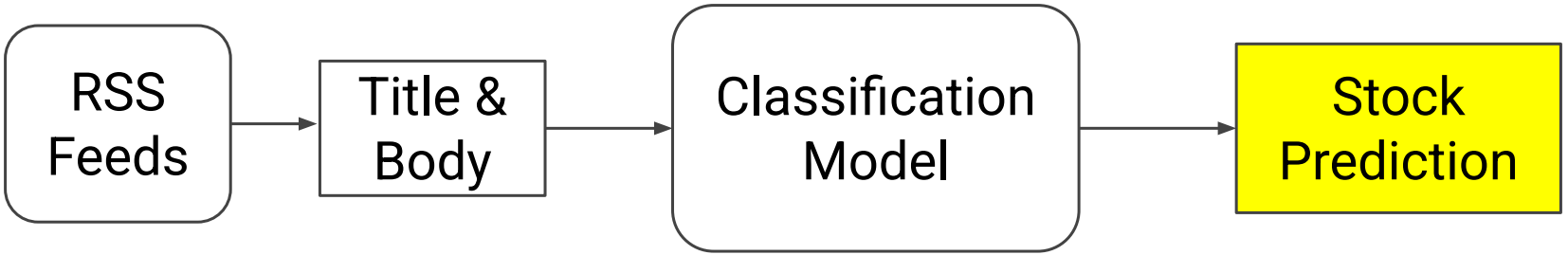
## Historical Nvidia Stock Prices (2011-2025)

- Yahoo Finance - NVIDIA Corporation (NVDA)
  - Daily stock prices

## Historical Nvidia News Articles (2011-2025)

- SEC EDGAR database
- GDELT Project
- Forbes

# Program Flow



# RSS Feeds

**Purpose:** Continuously monitor relevant news sources for new articles

**How it Works**

- Feeds are checked periodically for new articles
- Articles containing relevant keywords are forwarded to the classification model

# Initial Classification Model: TF-IDF + Logistic Regression

## Labeling the Data

- Each article assigned UP/DOWN/NEUTRAL based on Nvidia's 3-day return after publication
- Thresholds defined using k × volatility to account for market fluctuations.

## Training

- Input features: TF-IDF vectors of article *title + body*
- Output label: UP/DOWN/NEUTRAL based on future price

## Evaluation

- Evaluated on a Sentiment Analysis for Financial News dataset from Kaggle
- Low accuracy (~34%) motivated pivot to an ensemble approach using pre-trained financial sentiment classifiers.

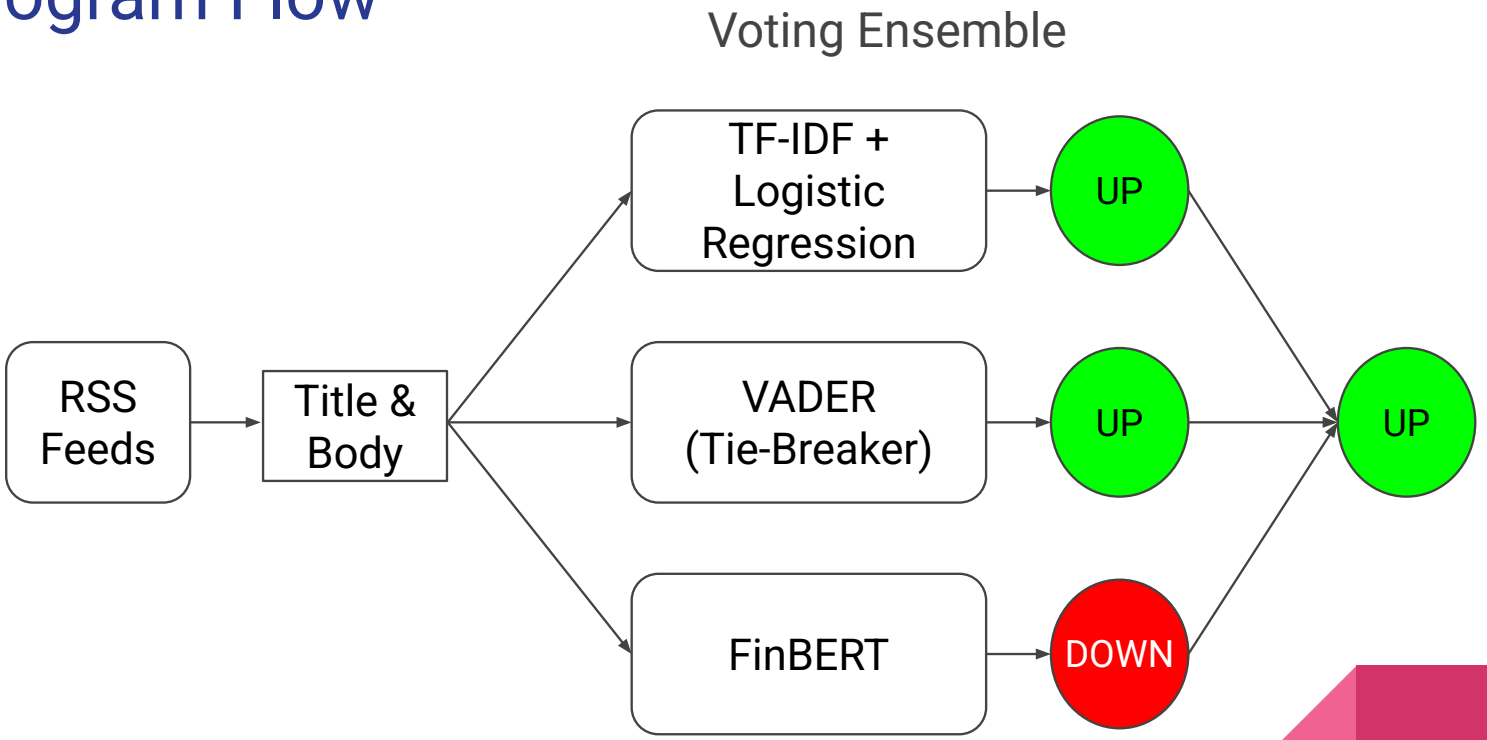
# Timeline of Work

October 14th, Sprint 1: Our scraper for historical stock data and news was created, making a concise set of csv files for three common tech companies stock values at the time of different article releases. The main one to be used will be the Nvidia csv. Furthermore, the Forbes article finder and and results were designed and implemented as well, providing most needed data for continued development.

November 6th, sprint 2: Work began on testing and understanding our chosen dataset and how to apply it to our chosen LLM, as well as a refinement to our article finders.

November 18, Sprint 3: With changes and realizations being made, our sentiment analysis system was converted into our LLM  work to ensure an easier and more realistic end goal while work on LLM prompts and set up continued. We achieved a functioning prototype that achieved our goals after much difficulty.

# Program Flow



# Voting Ensemble

**Objective:** Improve prediction accuracy beyond TF-IDF + logistic regression classifier.

**Voting**

- Each model predicts UP / DOWN / NEUTRAL
- Majority vote determines final prediction
- Tiebreaker: VADER prediction is used if votes are equal because it demonstrated the highest accuracy alone

**Evaluation**

- Accuracy (~54%) improved significantly over the TF-IDF + logistic regression alone

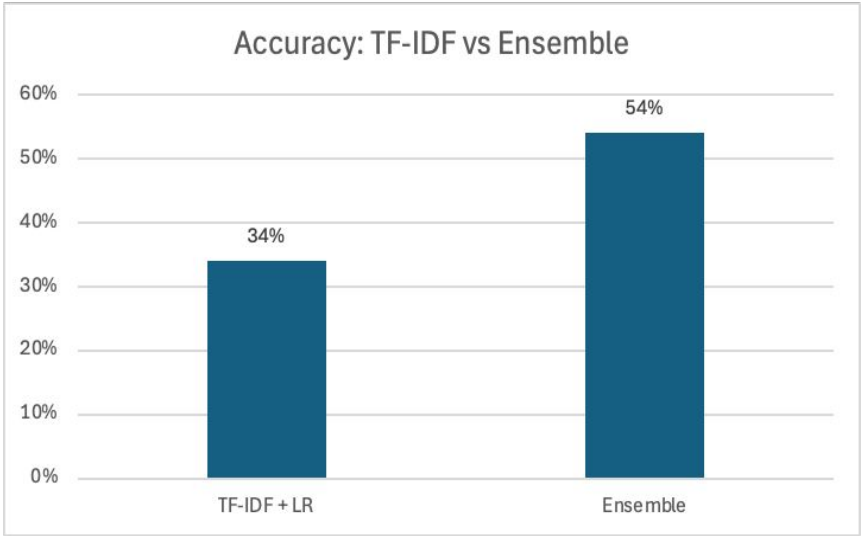
# Why an Ensemble? (Accuracy Comparison)

Each model has different strengths:

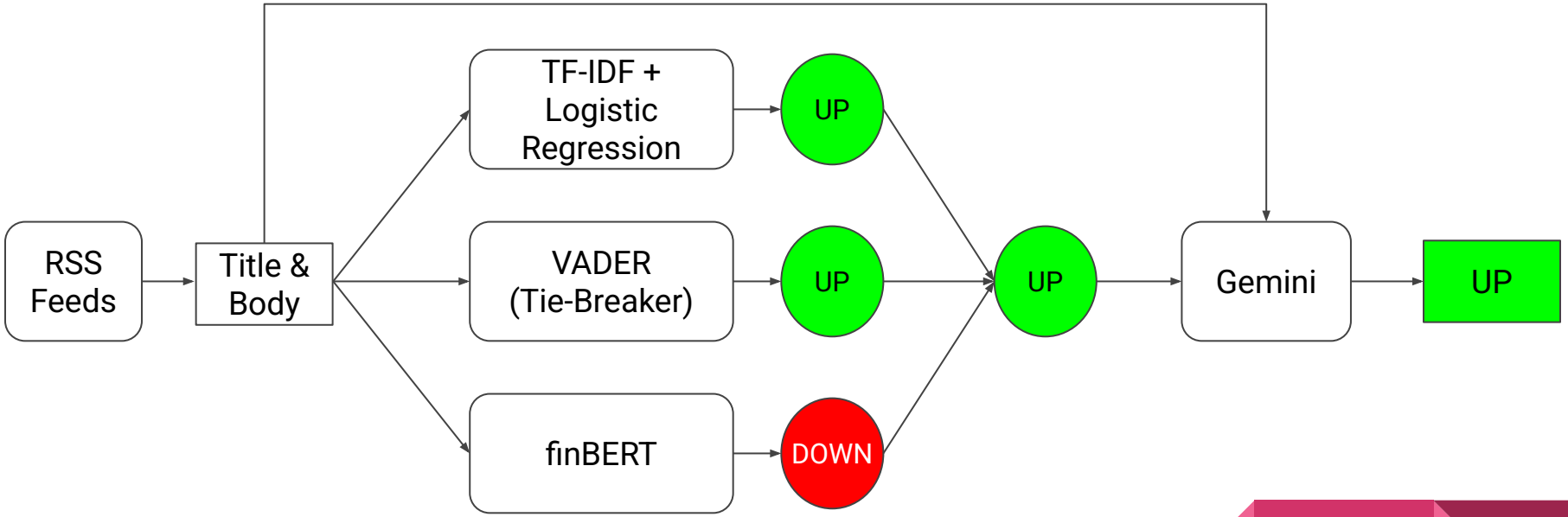
- **TF-IDF**: good linguistic features
- **VADER**: strong on sentiment
- **FinBERT**: excels on financial text

By combining the strengths of all three models the ensemble boosts overall accuracy from **34% to 54%**

Accuracy was still poor, so ensemble output was used as contextual input for an LLM



# Current Program Flow




Voting Ensemble

# Current State

**Model:** Gemini-2.5-Flash

- **Input:** Nvidia articles & ensemble output
- **Output:** Stock Prediction
- **Advantages**
  - Adaptive Reasoning
  - Improved Generalization

**Prediction History:** Predictions and articles are saved locally for future reference

```
[2025-11-24 Mon 13:09][Checking feeds for Nvidia articles]
[2025-01-27 Mon 04:11][DOWN][Biggest Market Loss In History: Nvidia Stock Sheds Nearly $600 Billion As DeepSeek Shakes AI Darling]
[2025-11-11 Tue 07:56][DOWN][Nvidia Loses 2 Billionaires Amid Softbank-Led Selloff]
[2025-11-18 Tue 09:51][UP][Nvidia's Next Trillion: The Story We've Told Since 2015]
[2025-11-24 Mon 12:15][UP][Revolut Secures Nvidia Investment As Valuation Soars To $75 Billion]
[2025-11-24 Mon 13:09][Checked all feeds]
```

# Future Work

- **Prompt Engineering**: Optimize Gemini-2.5-Flash for more accurate predictions
- **Expand Stock Coverage**: Include additional companies beyond Nvidia
- **Ensemble Improvements**: Experiment with additional pre-trained sentiment models