DUKE
FUQUA
SCHOOL OF BUSINESS

DUMAC Duke MQM: Capstone Project – Final Presentation

Team 11
Names: Ke-Hsuan Chu, Julia Tsai, Holly He, Zehua Yuan, Yan Guan

Agenda

1 Project Goal

2 Data Structure/ Data Preparation

3 Exploratory Data Analysis

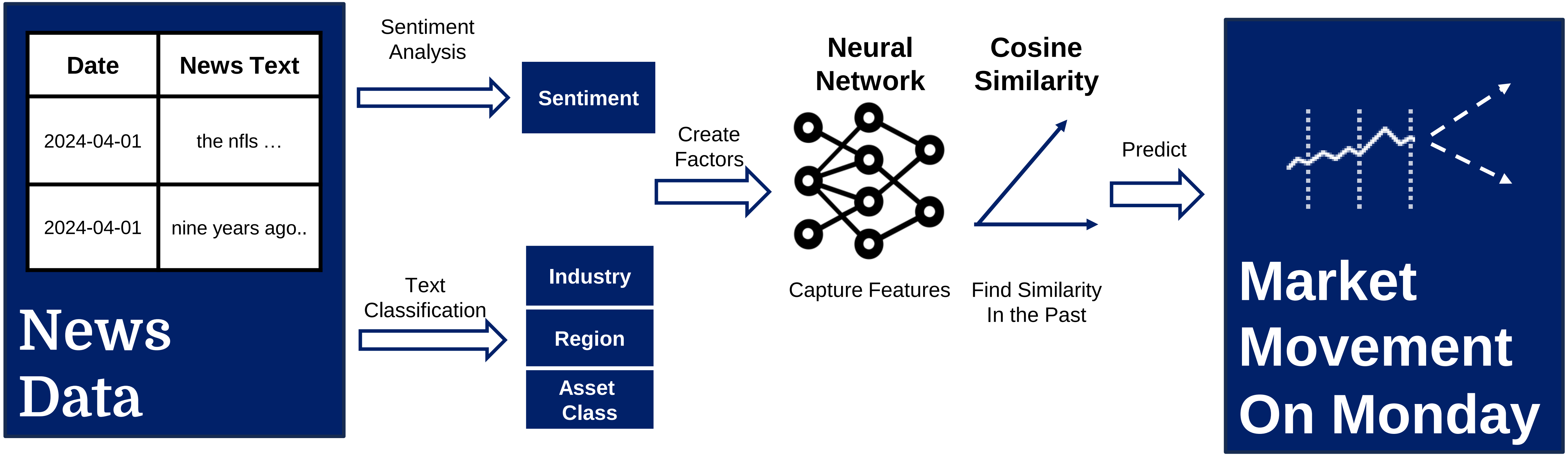
4 Predictive Model Design

5 Further Implementations

6 Q&A

Project Goal

Project Goal



Date	News Text
2024-04-01	the nfls …
2024-04-01	nine years ago..

News Data

Sentiment Analysis

Sentiment

Text Classification

Industry

Region

Asset Class

Create Factors

Neural Network

Capture Features

Cosine Similarity

Find Similarity In the Past

Predict

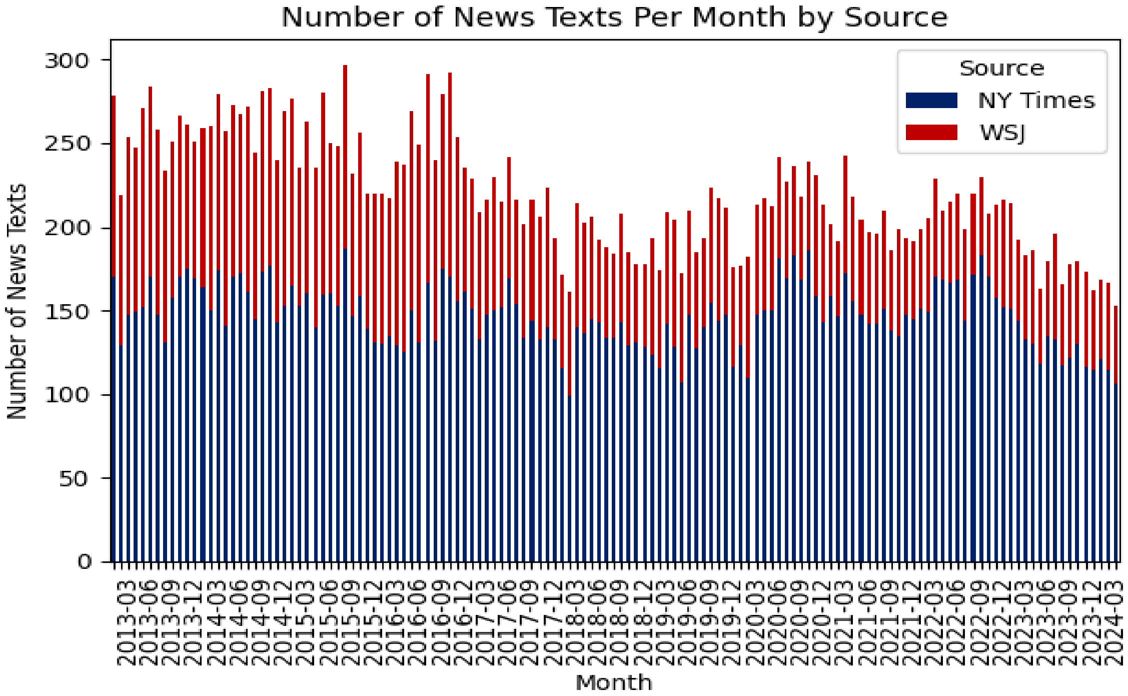
Market Movement On Monday

Data Structure/ Data Preparation

News Data from TDM Studio

29780 rows × 3 columns

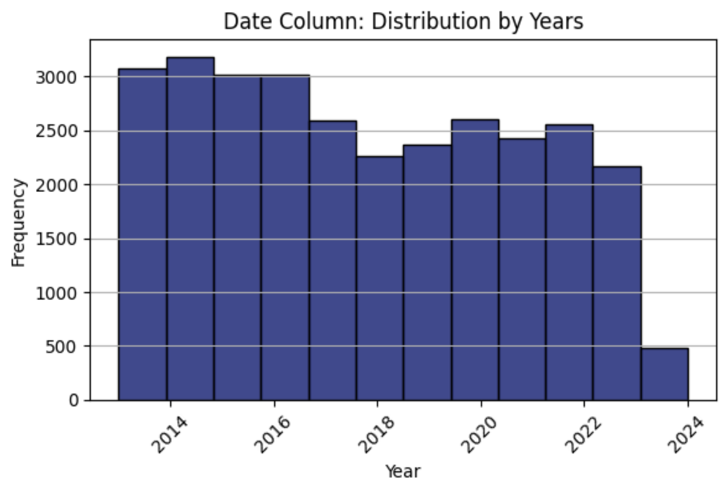
Date	News Text	Source
2013-01-01	the nfls …	WSJ
2013-01-01	nine years ago..	NY Times



Date

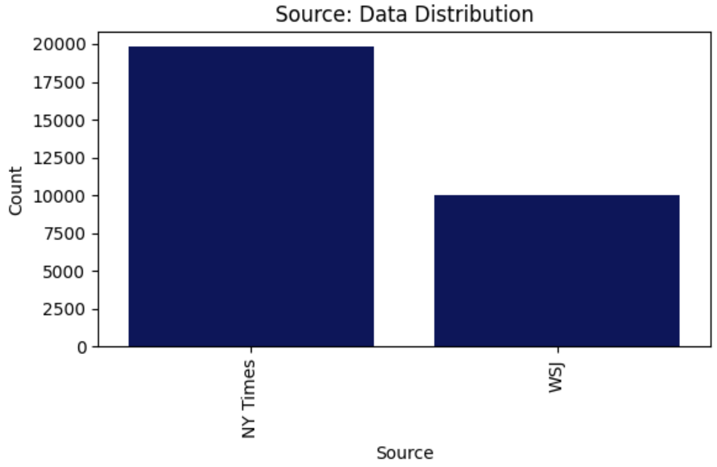
2013/01/01

2024-03-31



Source

WSJ: 9,980
NY Times: 19,800



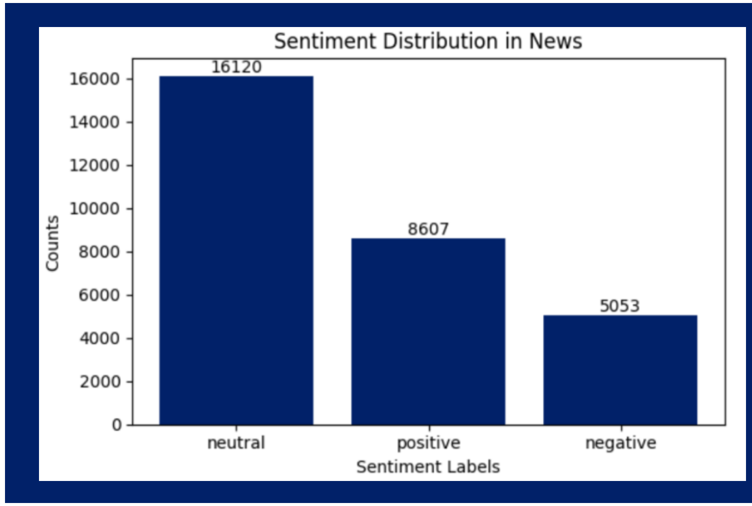
Sentiment Label from News Texts

Date	News Text	Source
2013-01-01	the nfls …	wallstreet
2013-01-01	nine years ago..	nyt

Sentiment Analysis

Pretrained Model

Sentiment
1
0



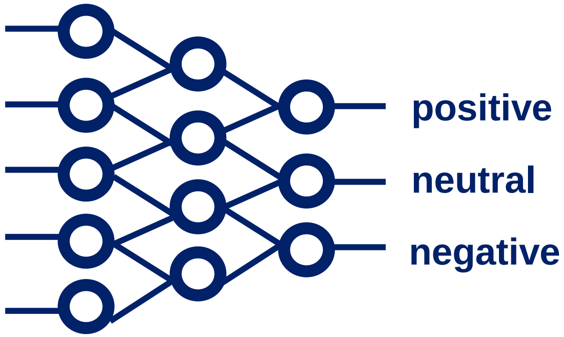
Sentiment

1 (Positive):	***8,607***	***/29%***
0 (Neutral):	***16,120***	***/54%***
-1 (Negative):	***5,053***	***/17%***

Sentiment Pre-train Model

DistilRoberta-financial-sentiment

Hugging Face



DistilRoberta

a smaller and faster version of RoBERTa

Accuracy

0.9823

Training Data

4.85k rows

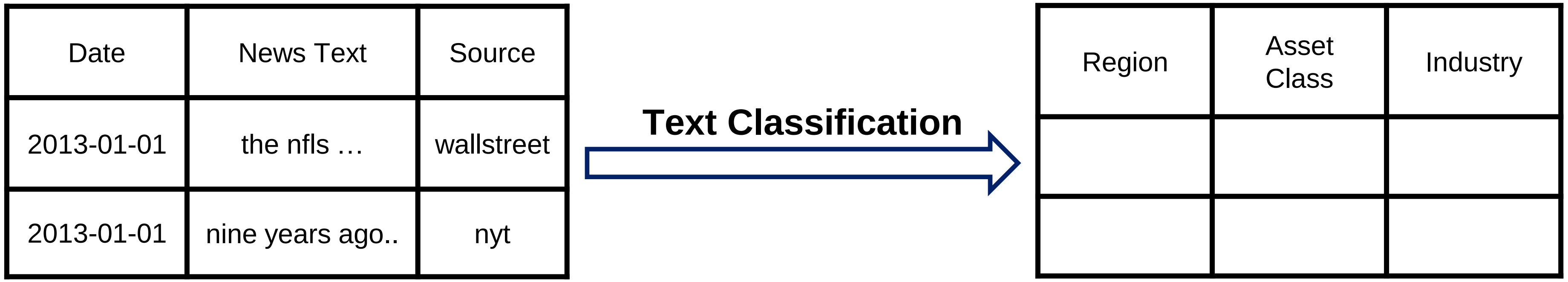
Sentence	Label
With the new production plant the company would increase its capacity to meet the expected increase in demand and would improve the use of raw materials and therefore increase the production profitability .	**positive**
According to Gran , the company has no plans to move all production to Russia, although that is where the company is growing .	**neutral**
The international electronic industry company Elcoteq has laid off tens of employees from its Tallinn facility; contrary to earlier layoffs the company contracted the ranks of its office workers, the daily Postimees reported.	**negative**

Source Link

DUKE FUQUA

Classification Labels from News Texts

29779 rows × 7 columns



Date	News Text	Source
2013-01-01	the nfls …	wallstreet
2013-01-01	nine years ago..	nyt

Text Classification

Region	Asset Class	Industry

Region

USA	12,358	41.5%
Emerging Market	0	0%
Developed Market	17,422	58.5%

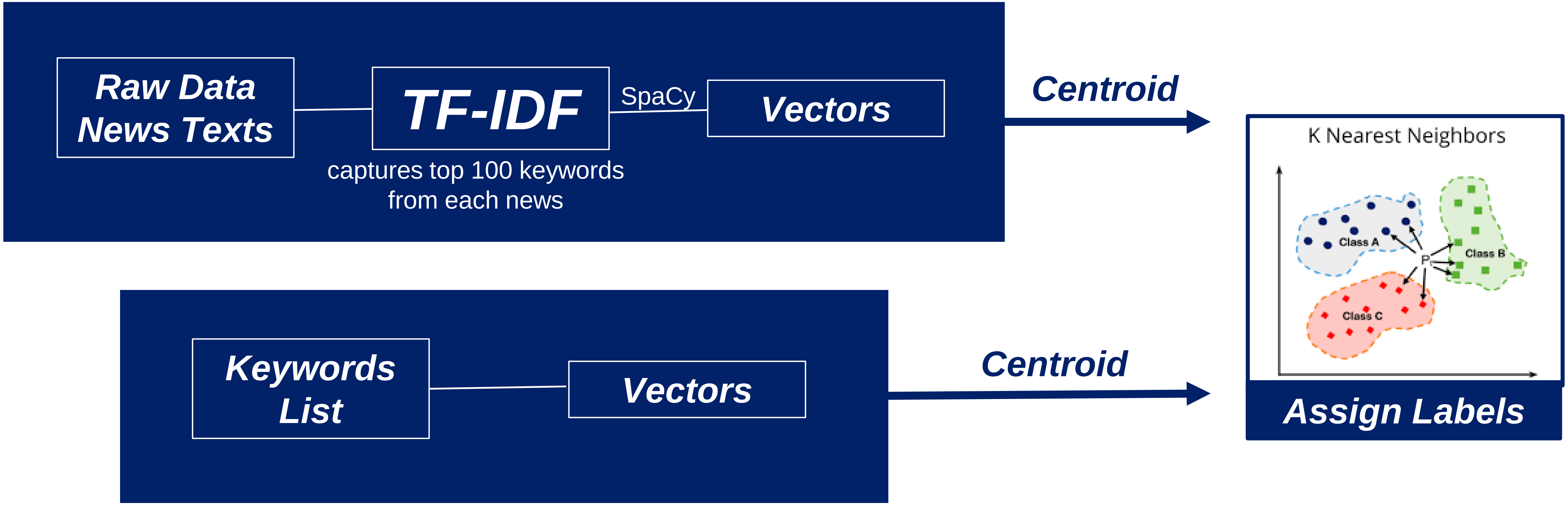
Asset Class

Equity	26,697	89.6%
Fixed Income	3,083	10.4%
Cash	0	0%

Industry

Energy	0	0%
Materials	7	0%
Industrials	11,688	39%
Consumer Discretionary	9472	32%
Consumer Staples	564	2%
Health Care	349	1.2%
Financials	82	0.3%
Information Technology	570	2%
Communication Services	6874	23%
Utilities	1	0%
Real Estate	173	0.6%

Text Classification Model



Raw Data News Texts

TF-IDF

SpaCy

Vectors

captures top 100 keywords from each news

Centroid

Keywords List

Vectors

Centroid

K Nearest Neighbors

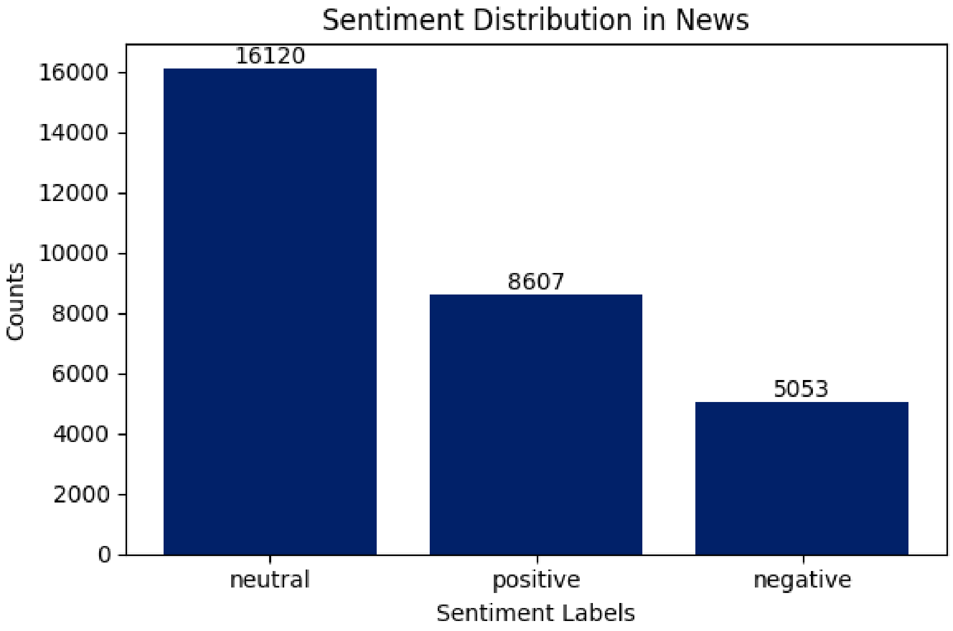
Assign Labels

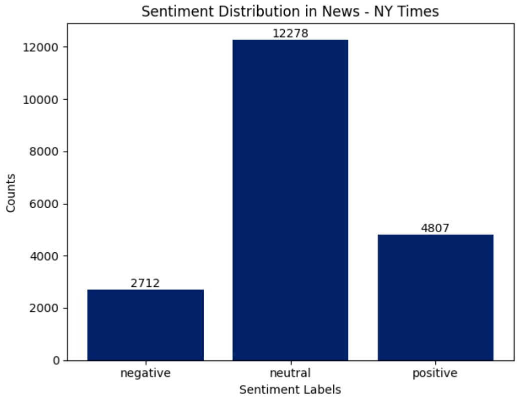
Text Classification Model

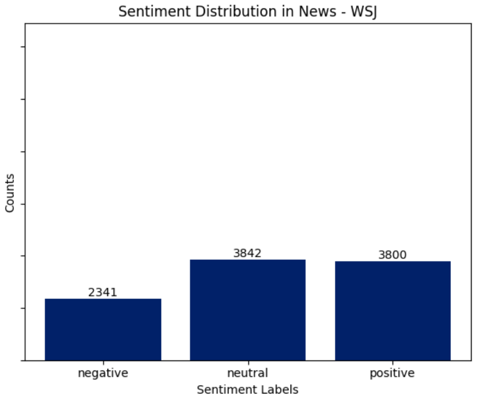
Industry	Energy	['energy','oil','gas','fuels','coal','drilling','drill','drills','drilled','petroleum','crude','renewable','oil price', 'pipeline’, 'opec','exxonmobil','chevron','bp','shell','total','conocophillips','schlumberger','halliburton']
	Materials	['chemicals','commodities','materials','iron','ore','metal','glass','glasses','plastic','container','gold','steel','silver','mining', 'forest','minerals','metallurgy','fertilizers','agricultural','vale','bhp','glencore','alcoa','newmont','anglo']
	Utilities	['utilities','electricity','gas','water','power','utility','renewable','energy','grid','infrastructure','energy','distribution`,'wastes`,'nextera','dominion','energy','exelon','edison','sempra']
Region	USA	['united','states','america','american','us','us dollar','usd','trump','biden','sustainability','silicon','valley','hollywood','federal','fed']
	Emerging Market	['developing nations','emerging markets','brics','asean','frontier markets','asia','africa','latin','china','india','brazil','russia','mexico','indonesia','turkey','philippines','thailand','vietnam','egypt','nigeria','argentina','pakistan','iran','colombia','bangladesh','malaysia','poverty,'infrastructure development','rural','aid']
	Developed Market	['advanced','oecd','g7','g8','g20','highincome','aging','sustainability','europe','eu','euro','canada','united','kingdom','germany','france','japan','australia','switzerland','netherlands','sweden','denmark','norway','finland','singapore','korea','taiwan']

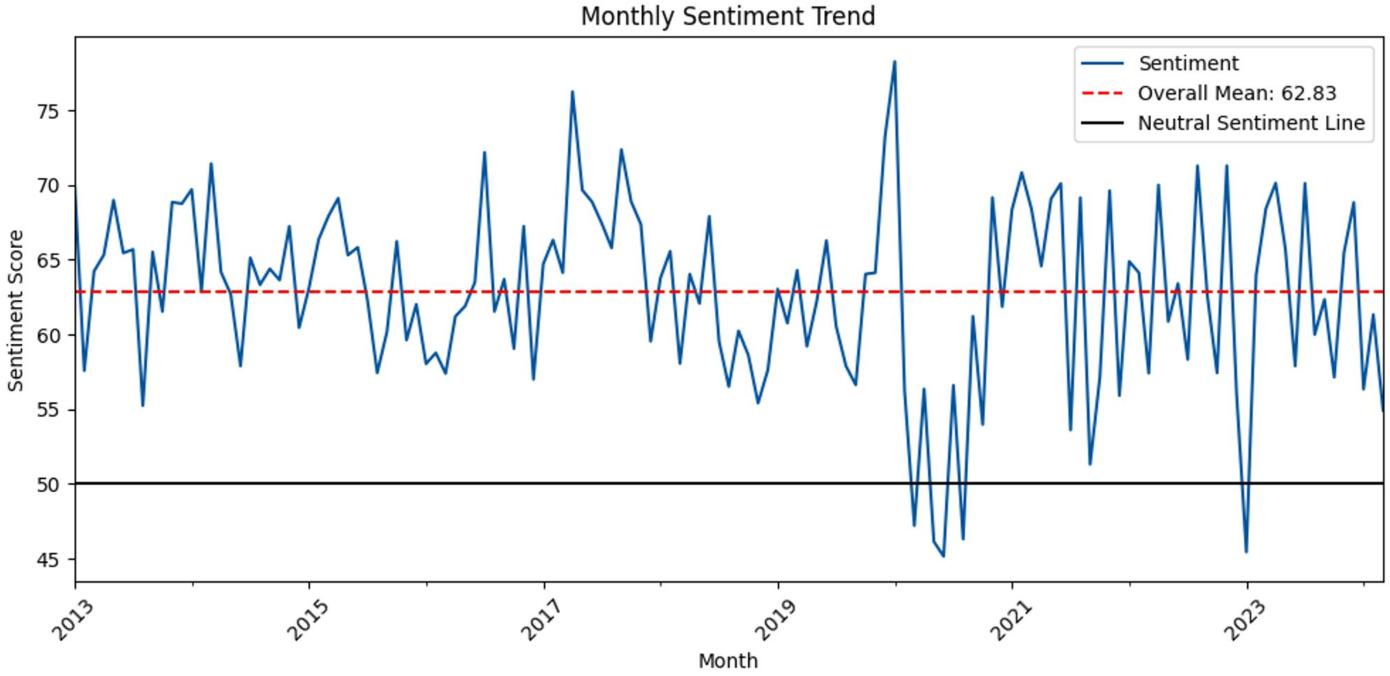
Exploratory Data Analysis (EDA)

Overall Sentiment Distribution and Trend



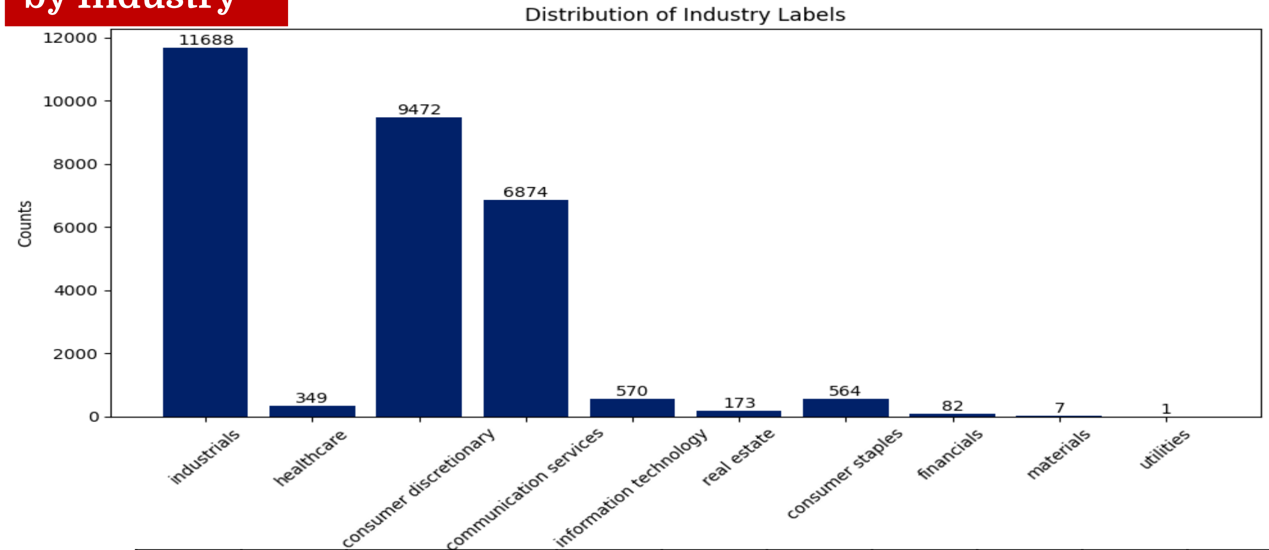


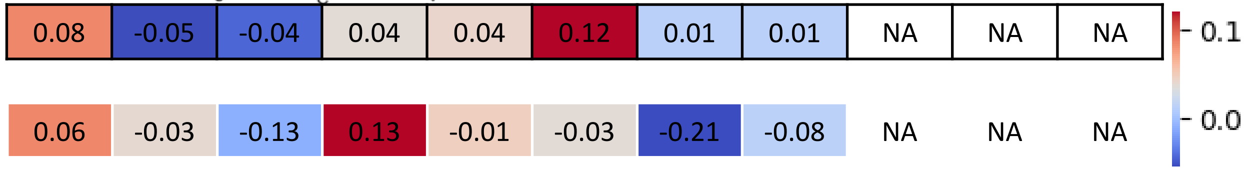




Correlation between Sentiment and Return of Index

by Industry





Weekly	0.08	-0.05	-0.04	0.04	0.04	0.12	0.01	0.01	NA	NA	NA
Monthly	0.06	-0.03	-0.13	0.13	-0.01	-0.03	-0.21	-0.08	NA	NA	NA

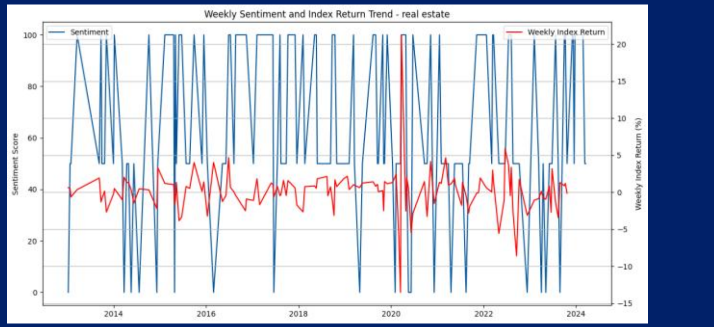
Correlation between Sentiment Score and Return of the Sector Index

- Industries with more data remain the same direction
- Monthly correlation shows more negative correlations

Assume positive correlations, we think the market reactions to sentiment might be shorter.

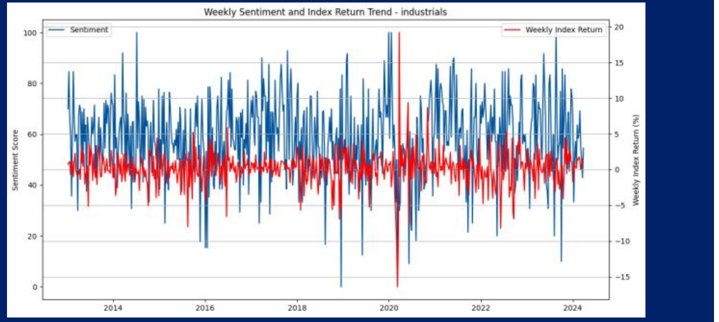
Real Estate

0.12



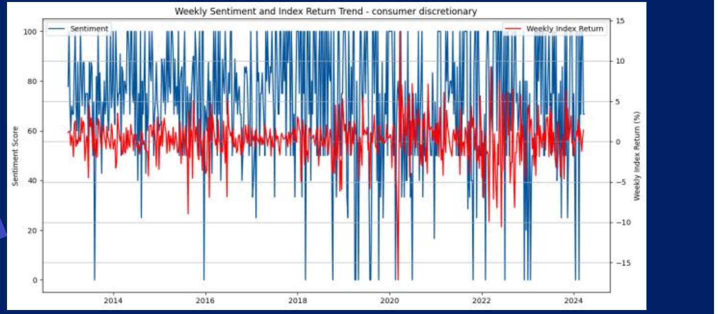
Industrials

0.08



Consumer Discretionary

-0.04



Predictive Model Design

Features and Target Variable of the Neural Network

Target variable – Weekly Index Movement Signal

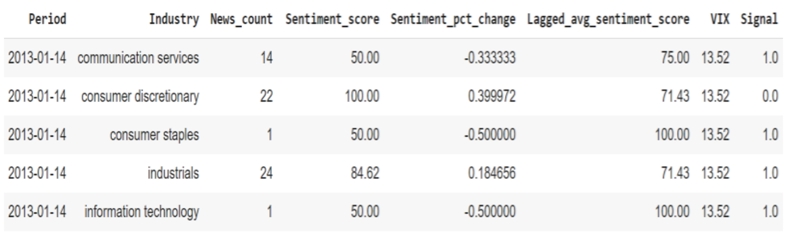
- MSCI World Sector Indices, Close price from Jan 1, 2013 to Mar 29, 2024
- Label 1 if the close price of next Monday > the close price of this Friday; else label 0

7 Features

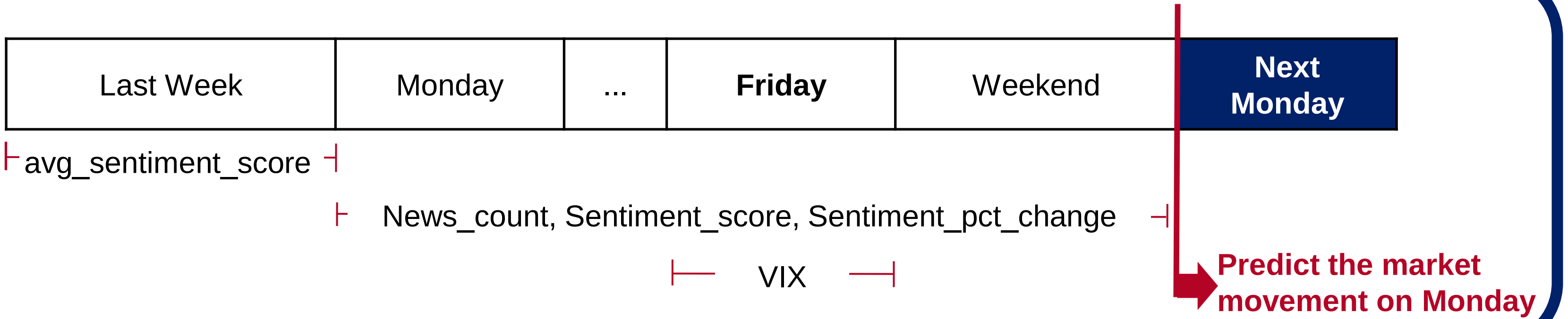
By industry, By week

- Period: The start date of a week
- Industry label
- News_count: Total number of news of an industry in a week
- Sentiment_score: Number of positive sentiment of an industry/ (Number of positive sentiment + Number of negative sentiment) * 100
- Sentiment_pct_change: Percentage change of the sentiment score
- Lagged_avg_sentiment_score: Average sentiment score of last week
- VIX: The close price of the VIX index on Friday

Example Data



Period	Industry	News_count	Sentiment_score	Sentiment_pct_change	Lagged_avg_sentiment_score	VIX	Signal
2013-01-14	communication services	14	50.00	-0.333333	75.00	13.52	1.0
2013-01-14	consumer discretionary	22	100.00	0.399972	71.43	13.52	0.0
2013-01-14	consumer staples	1	50.00	-0.500000	100.00	13.52	1.0
2013-01-14	industrials	24	84.62	0.184656	71.43	13.52	1.0
2013-01-14	information technology	1	50.00	-0.500000	100.00	13.52	1.0



Predict the market movement on Monday

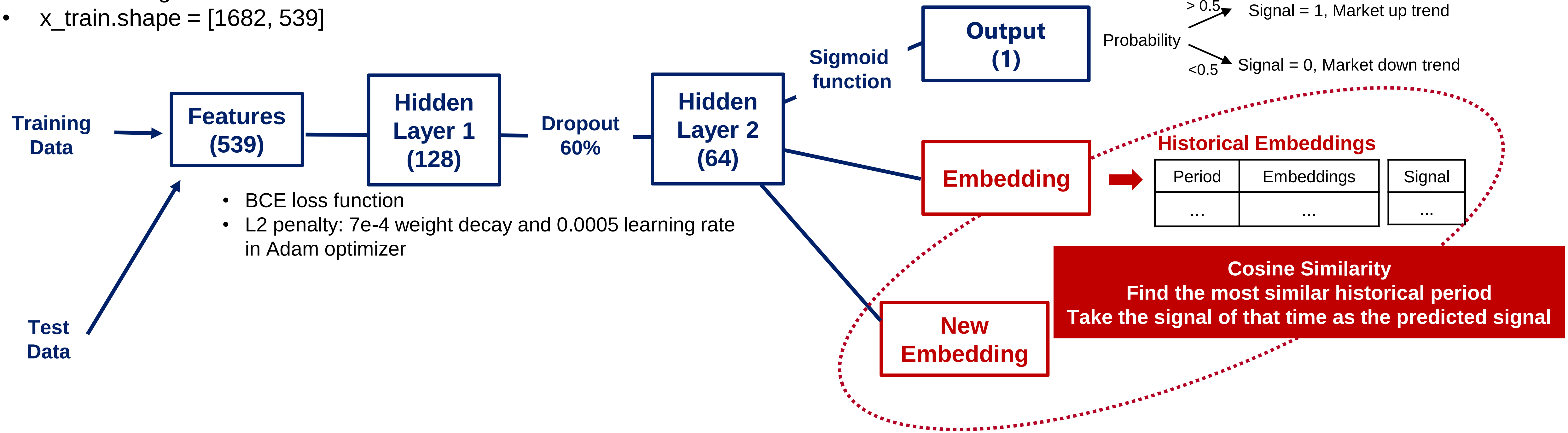
Algorithm of the Model: Neural Network + Cosine Similarity

Data Preprocessing

- Categorical features (Period, Industry): OneHotEncoder
- Numeric features (News_count, Sentiment_score, Sentiment_pct_change, Lagged_avg_sentiment_score, VIX): StandardScaler

Structure of the Neural Network

- 70% training data + 30% test data
- x_train.shape = [1682, 539]



- BCE loss function
- L2 penalty: 7e-4 weight decay and 0.0005 learning rate in Adam optimizer

Historical Embeddings

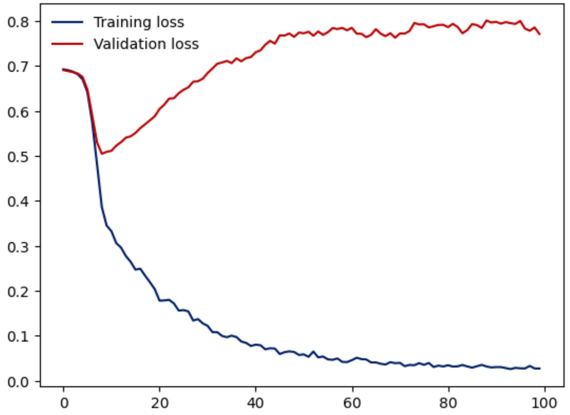
Period	Embeddings	Signal
…	…	…

Cosine Similarity
Find the most similar historical period
Take the signal of that time as the predicted signal

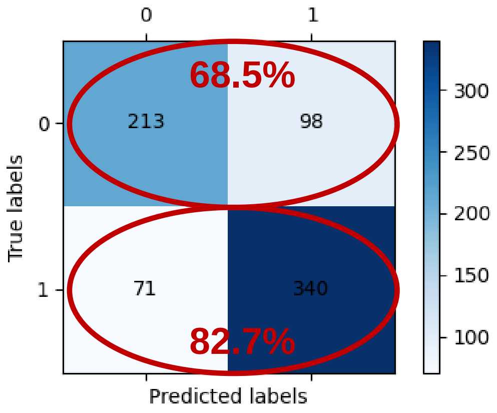
Evaluation of the Model

Model Performance

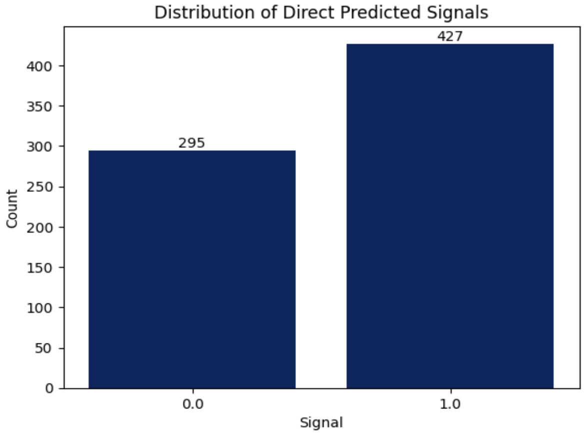
Losses of the Neural Network

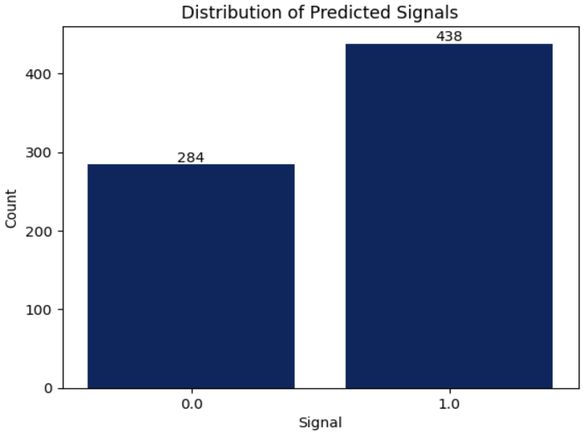


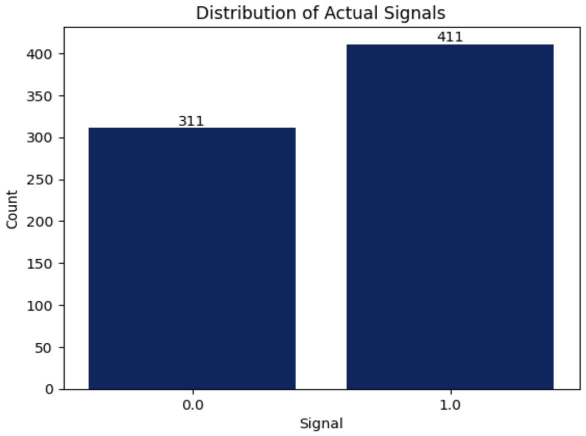
Confusion Matrix



Comparison







Accuracy of Direct NN

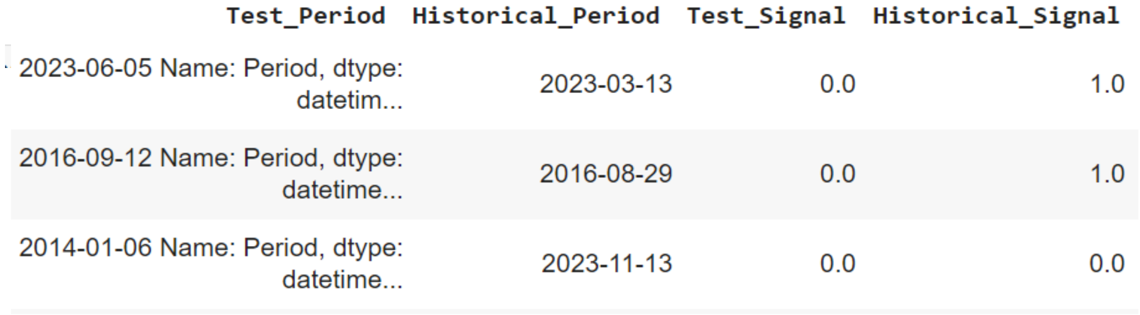
77.84%

Accuracy of NN + Similarity

76.59%

Precision	***Recall***	***F1 Score***
77.63%	***82.73%***	***80.09%***

Example Data

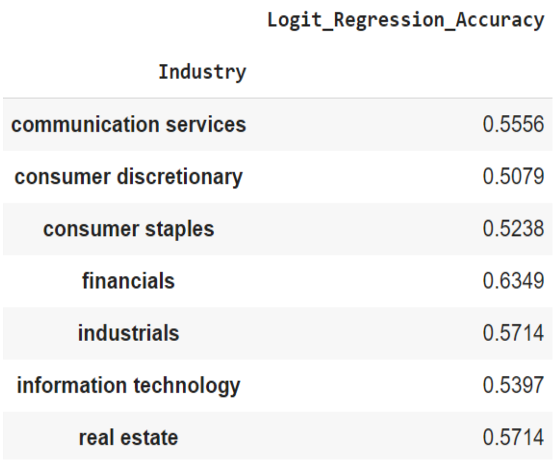


Test_Period	Historical_Period	Test_Signal	Historical_Signal
2023-06-05 Name: Period, dtype: datetim...	2023-03-13	0.0	1.0
2016-09-12 Name: Period, dtype: datetime...	2016-08-29	0.0	1.0
2014-01-06 Name: Period, dtype: datetime...	2023-11-13	0.0	0.0

Comparison of Different Models

Logistic Regression

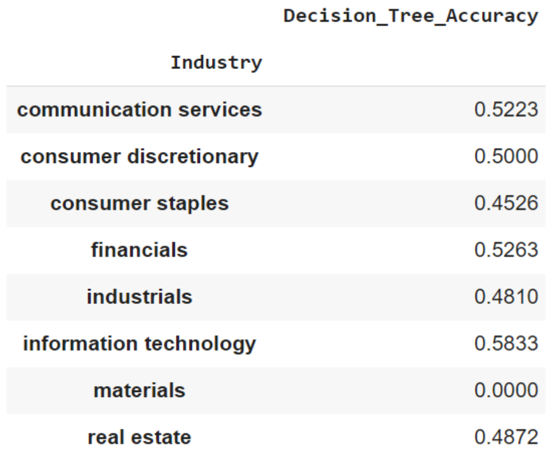
regression_model = smf.logit( )



	Logit_Regression_Accuracy
Industry	
communication services	0.5556
consumer discretionary	0.5079
consumer staples	0.5238
financials	0.6349
industrials	0.5714
information technology	0.5397
real estate	0.5714

Decision Tree

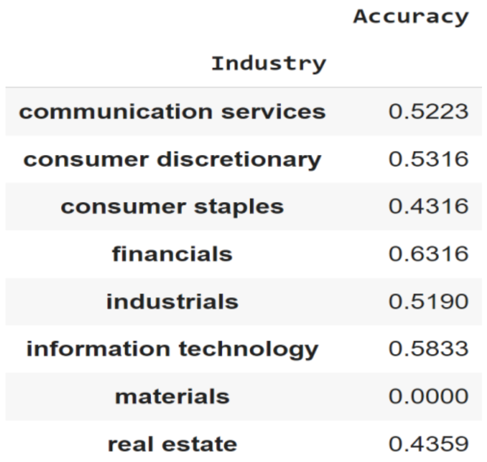
tree_model = DecisionTreeRegressor( )



	Decision_Tree_Accuracy
Industry	
communication services	0.5223
consumer discretionary	0.5000
consumer staples	0.4526
financials	0.5263
industrials	0.4810
information technology	0.5833
materials	0.0000
real estate	0.4872

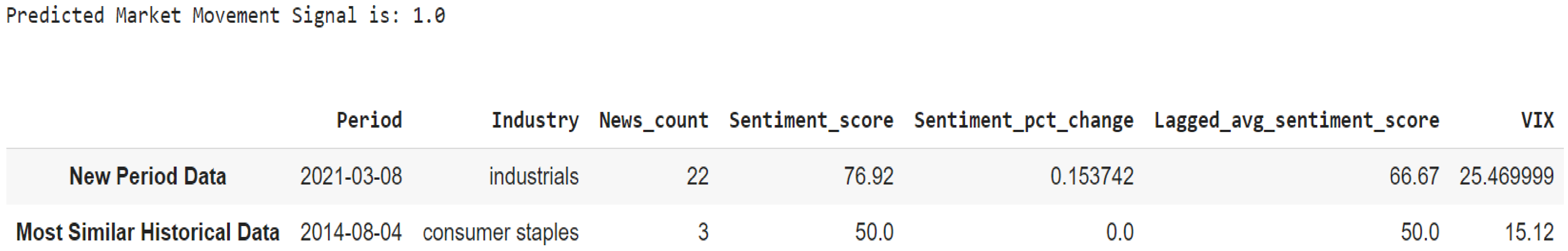
Feature Cosine Similarity

Overall Accuracy: 52%



	Accuracy
Industry	
communication services	0.5223
consumer discretionary	0.5316
consumer staples	0.4316
financials	0.6316
industrials	0.5190
information technology	0.5833
materials	0.0000
real estate	0.4359

Application of the Model



```
Predicted Market Movement Signal is: 1.0
```

	Period	Industry	News_count	Sentiment_score	Sentiment_pct_change	Lagged_avg_sentiment_score	VIX
New Period Data	2021-03-08	industrials	22	76.92	0.153742	66.67	25.469999
Most Similar Historical Data	2014-08-04	consumer staples	3	50.0	0.0	50.0	15.12

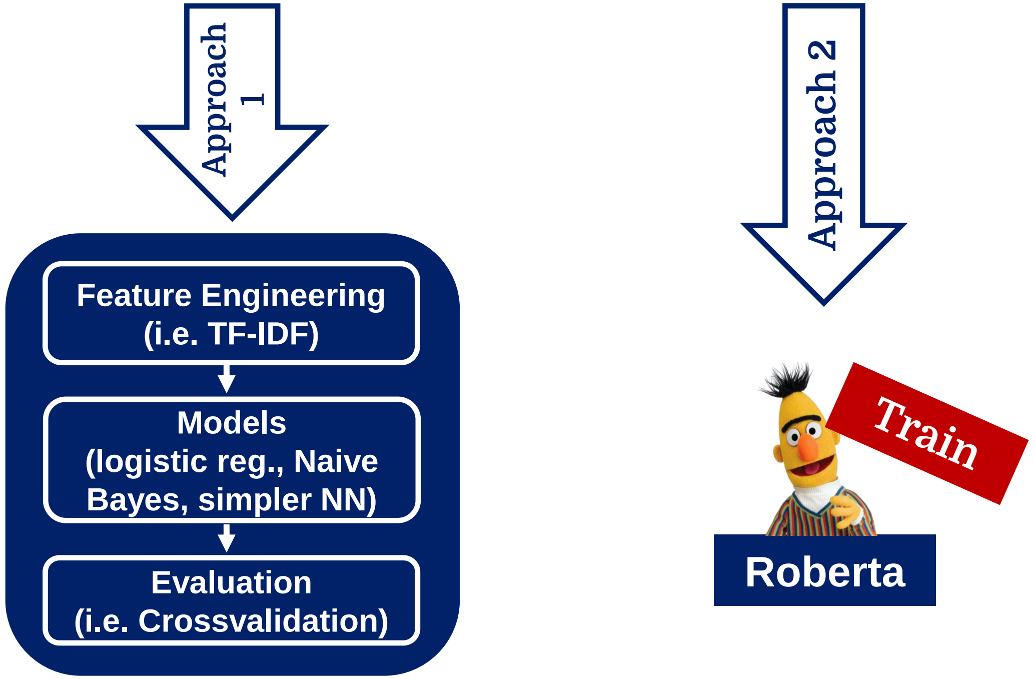
Further Implementations

Further Implementations

Text Classification

News Data

News Text	Region	Industry	Asset Class
the nfls …	usa	energy	Fixed income
nine years ago..	emerging	industrial	equity



Feature Engineering (i.e. TF-IDF)

Models (logistic reg., Naive Bayes, simpler NN)

Evaluation (i.e. Crossvalidation)

Roberta

Predicted Market Movement Signal

Features included:

Industry Label

+

Region Label

+

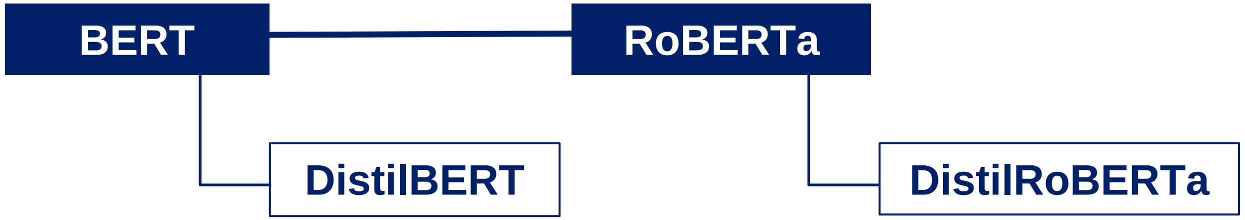
Asset Class Label

- We now have around 30k news over last 11 year period. Gather more data to support predicting market trends for all three labels.

Q&A

Appendix

DistilRoBERTa for Sentiment Pretrained Model

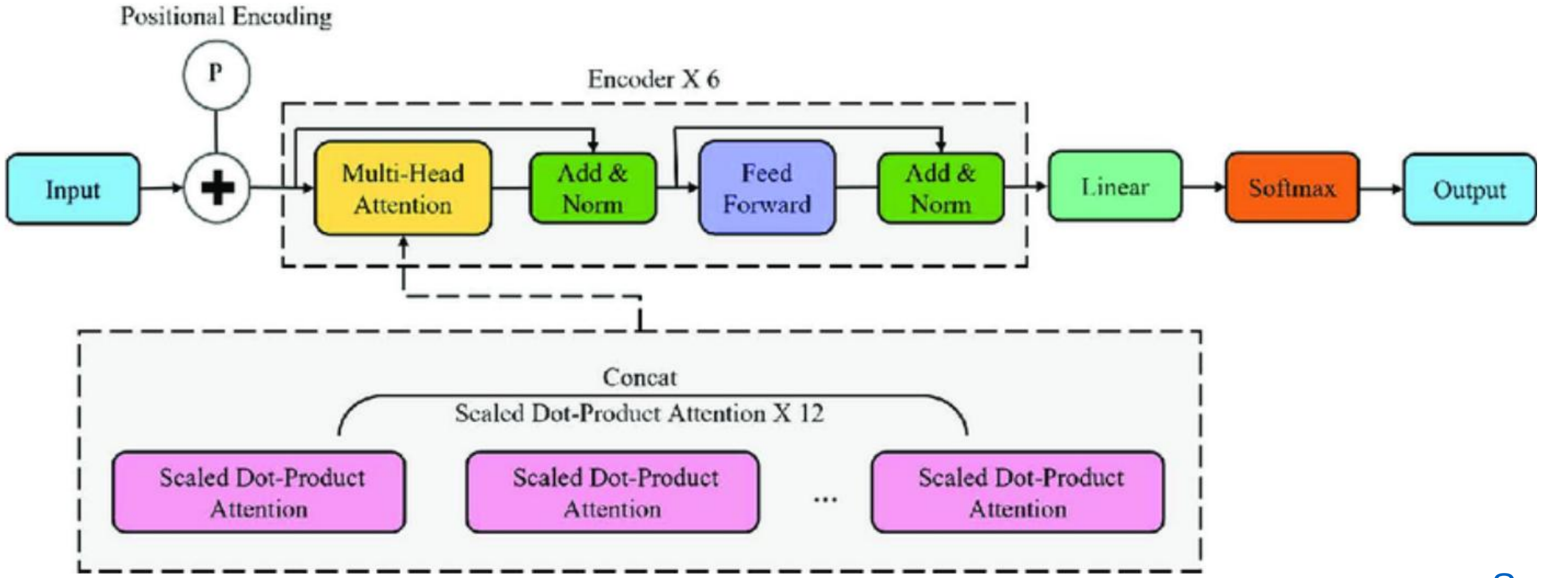


BERT

RoBERTa

DistilBERT

DistilRoBERTa


Positional Encoding
P
Encoder X 6
Input
Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Linear
Softmax
Output
Concat
Scaled Dot-Product Attention X 12
Scaled Dot-Product Attention
Scaled Dot-Product Attention
...
Scaled Dot-Product Attention

Source

TF-IDF and SpaCy for Text Classification

TF-IDF

Term Frequency - Inverse Document Frequency

how many times a word appears in a document X the inverse document frequency of the word across a set of documents

More Frequent in news TF ↑ **More common over news IDF ↓**

Source

SpaCy

Word Embeddings
pre-trained word embeddings

capture semantic similarity and relationships between words/documents

BERT/ RoBERTa

can capture deep contextual relationships in text and generate high-quality word embeddings that are sensitive to context.

Market Sentiment Definition

The prevailing view among investors regarding the current market conditions or specific securities is broadly recognized.

The collective mood among investors often leads to stock market volatility through bullish or bearish sentiments.
Investor sentiment is often seen as a self-fulfilling prophecy. For instance, if a business continues to grow but at a slower rate than before, a bearish sentiment might emerge. As this negative outlook becomes mainstream, investors may start selling their shares, which can lead to price drops and the onset of a bear market.

Thus, investor sentiment can drive market movements, even when not directly linked to fundamental economic indicators.
Ultimately, the market often reflects what investors collectively believe, turning perceptions into financial reality.

https://www.vectorvest.com/blog/stockmarket/what-is-market-sentiment/#:~:text=A%20good%20rule%20of%20thumb,and%20stocks%20could%20be%20undervalued.

Market Sentiment Application

Different investment strategies influence how investors interpret stock market sentiment and make investment decisions accordingly.

- **Swing Trading and Market Indicators** *Using Technical Analysis to Time Trades*

For instance, swing traders utilize technical indicators to detect shifts in volatility or market stability, aiding their decisions on when to enter or exit trades profitably.

- **Options Trading and Risk Management** *Leveraging Investor Sentiment for Portfolio Profitability*

Furthermore, options traders manage risk by gauging market sentiment. Early signs of bullish or bearish trends enable them to adjust their strategies, ensuring their portfolios remain profitable in line with current market dynamics and overall investor mood.

- **Contrarian Investing Strategy** *Capitalizing on Opposing Market Sentiments*

Contrarian investors, on the other hand, deliberately take positions against the predominant market sentiment. For example, they might choose to purchase stocks when general investor sentiment is bearish, betting against the majority.

Market Sentiment Indicators

VIX (CBOE Volatility Index) The VIX, also known as the "fear index," measures expected volatility over the next 30 days based on options prices. A rising VIX suggests higher risk and volatility, indicating the need for investor caution, while a lower VIX suggests lower volatility without indicating market direction.

Bullish Percent Index (BPI) The BPI tracks stocks displaying bullish patterns. Values above 80% suggest extreme market optimism and possible overvaluation, while values below 20% indicate pessimism and potential undervaluation.

High-Low Index This index compares the number of stocks hitting 52-week highs to those hitting lows, assessing market sentiment. A reading under 30 suggests bearish sentiment, while a reading above 70 indicates bullish sentiment.

Commitment of Traders Report (COT) The COT reports on futures holdings of select commodity traders, helping gauge market sentiment. Contrarian investors particularly use this to assess market sentiment.

Moving Averages Investors use moving averages like the 50-day and 200-day to gauge market sentiment. A 50-day average above the 200-day signals positive sentiment (bullish), while below it suggests negative sentiment (bearish).