Improve Sequence Prediction By Learning Network Graphical Dependencies

Hung V Ngo

Time Series Dependencies

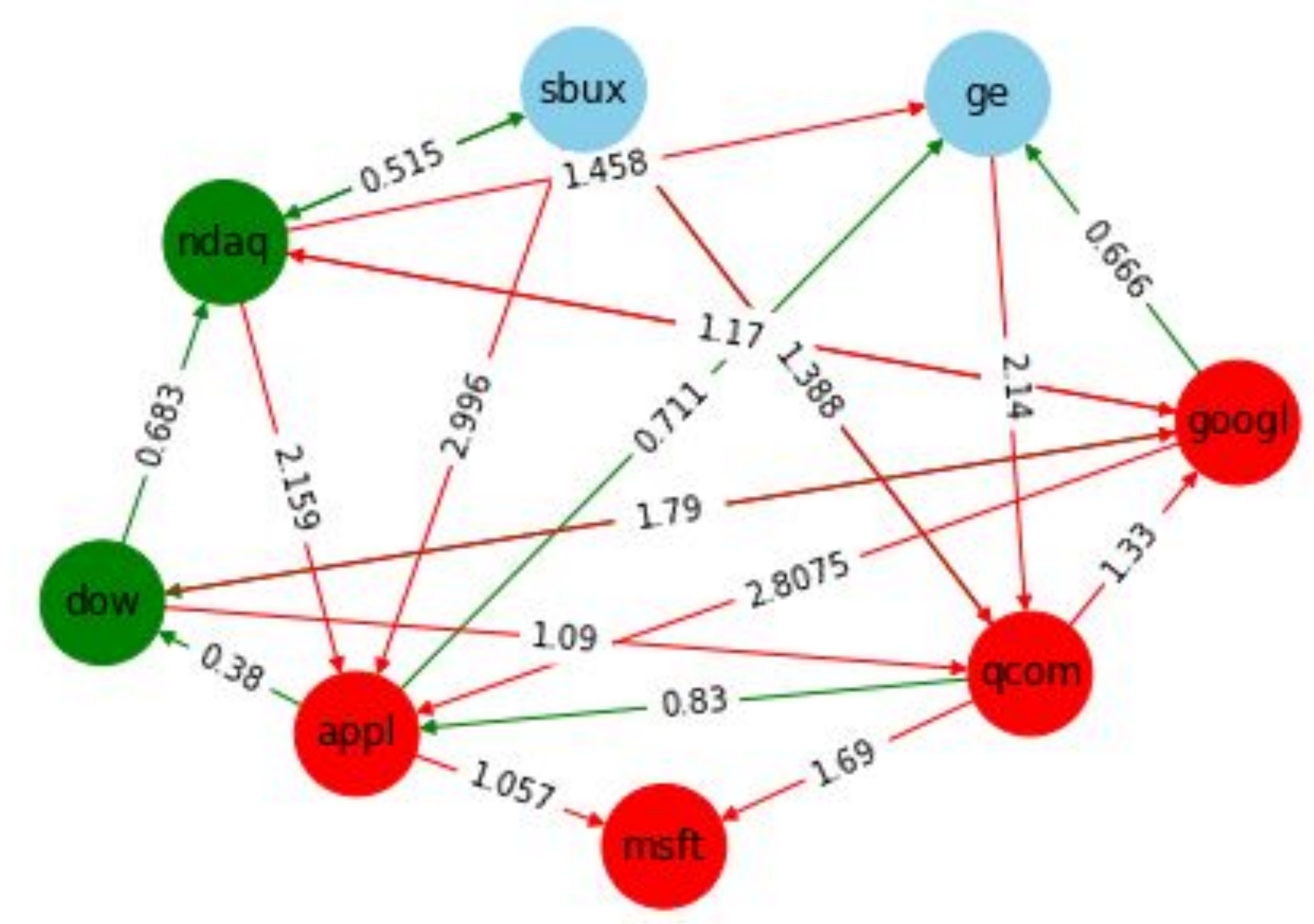

Figure 1: Company sentiment dependencies with PCA labels

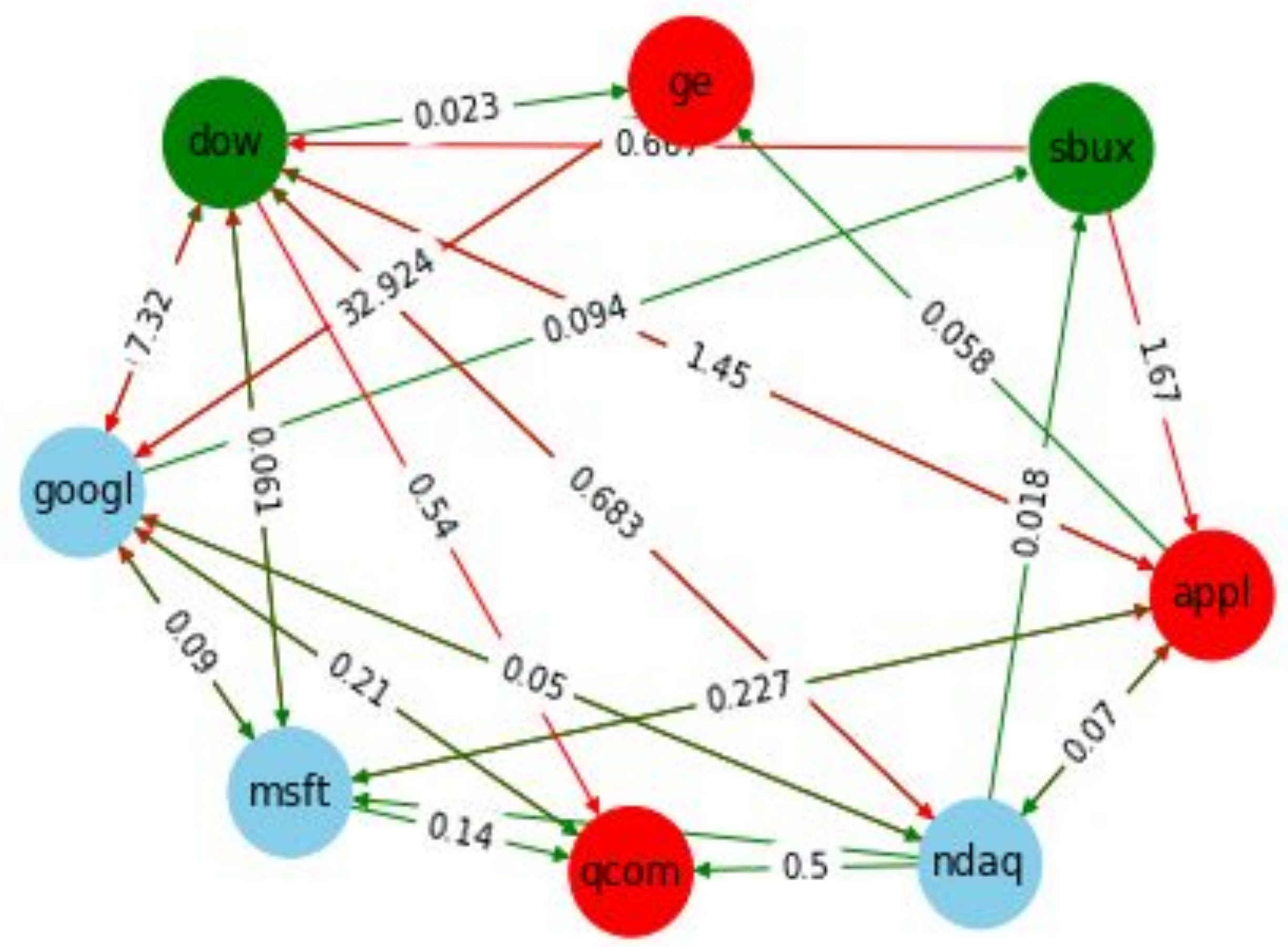

Figure 2: Company stock price dependencies with tSNE labels

Interesting Facts

> Tech software companies are highly influenced by each other with respect to sentiment, but not with stock price

> Pairs with high correlation in sentiment usually have low correlation in stock price

> Correlation matrix shows price and sentiment data is similar because companies behaves as a uniform market

> First two principal directions of stock price and sentiment data captures 95% dynamics of the system but PCA fail at clustering stock price data

> Same group in tSNE labels are companies that share similar inward edge pattern

> Same group in PCA labels are companies that share similar node ranks (total number of edges)

Data

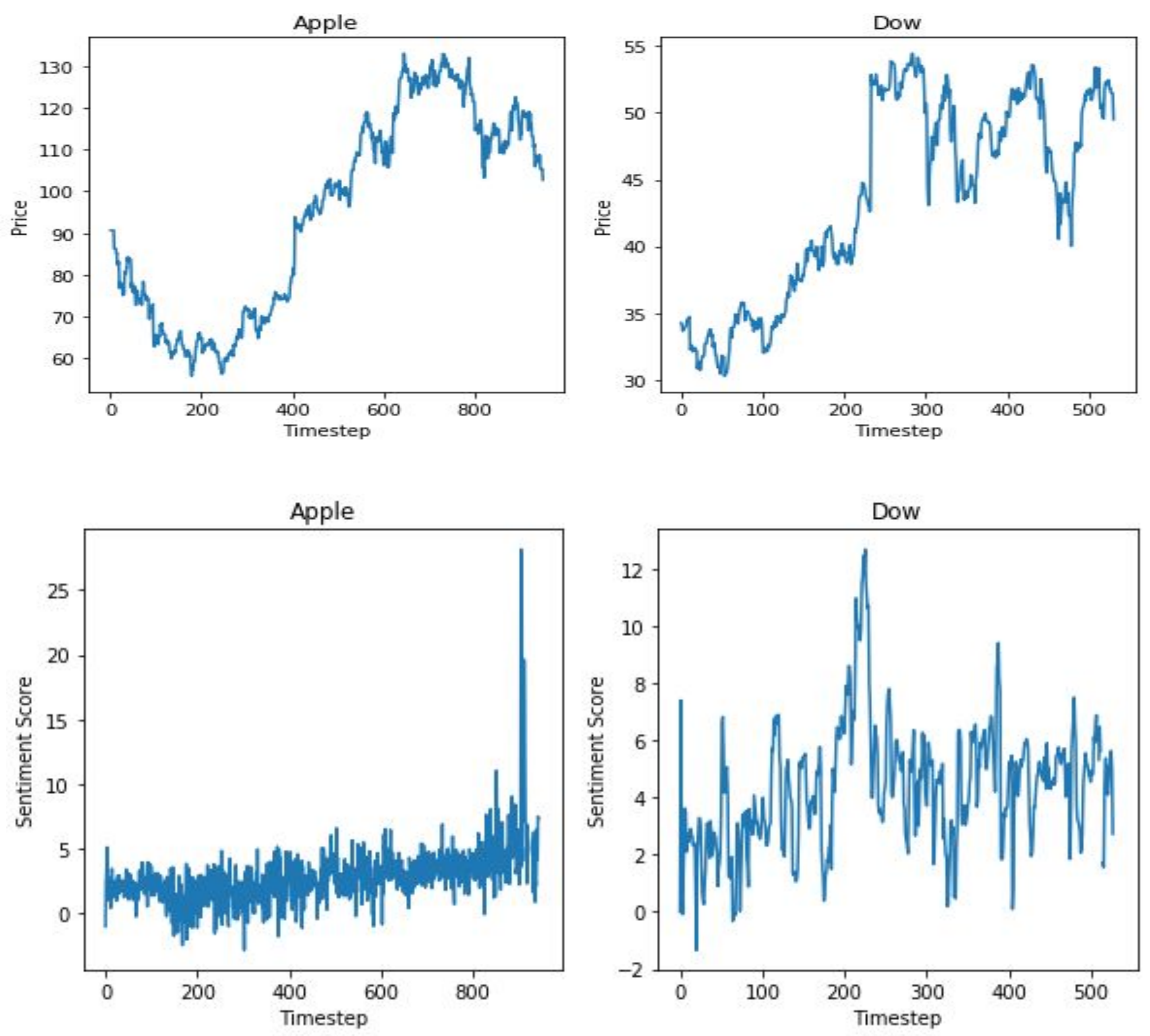

Figure 3: Sentiment and stock data
Number of companies: 8
Timesteps: 738
Preprocessing: Merge conflict companies that do not have sentiment data at some times

Previous Work

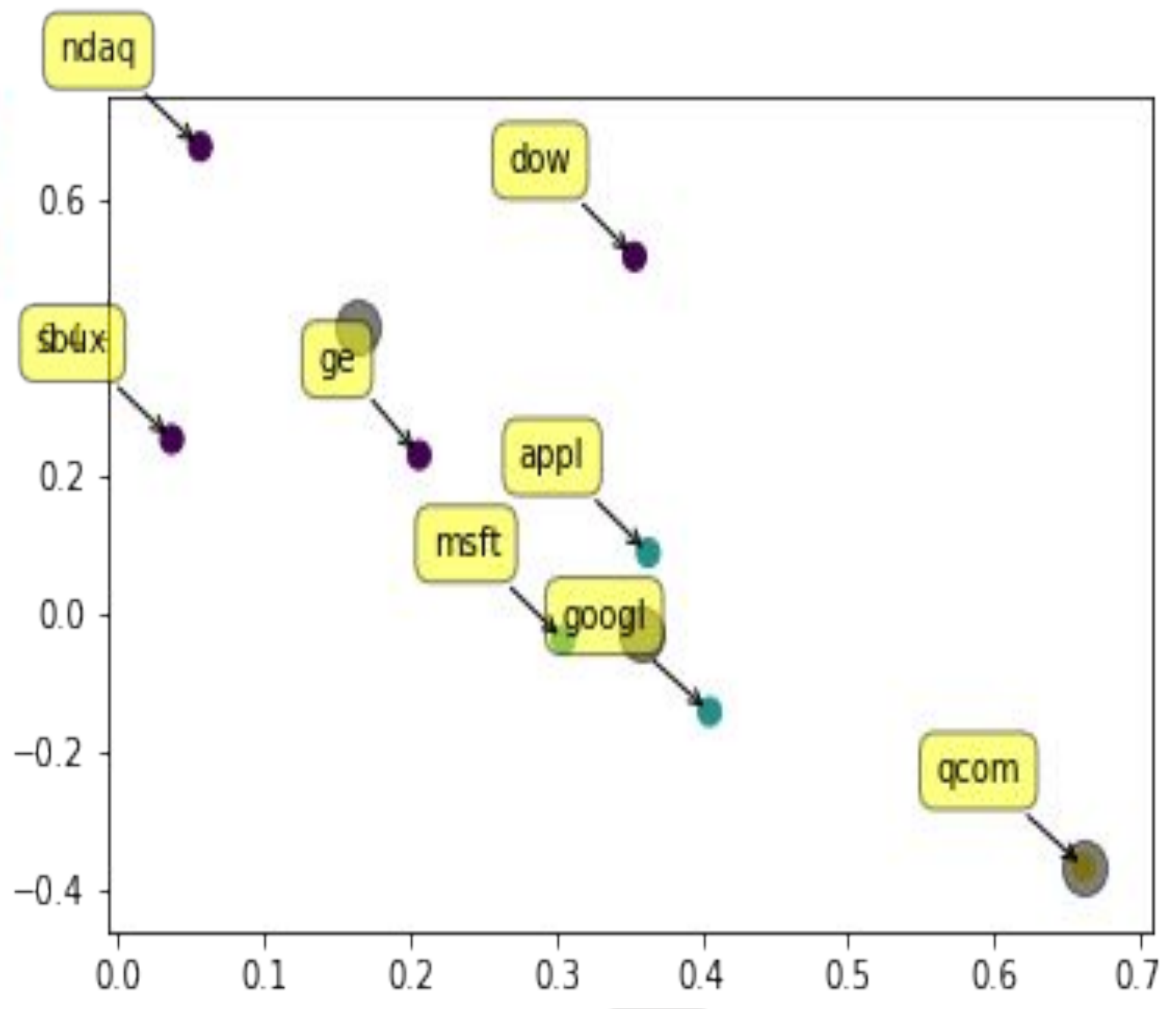

Figure 4: Principal component analysis on sentiment data

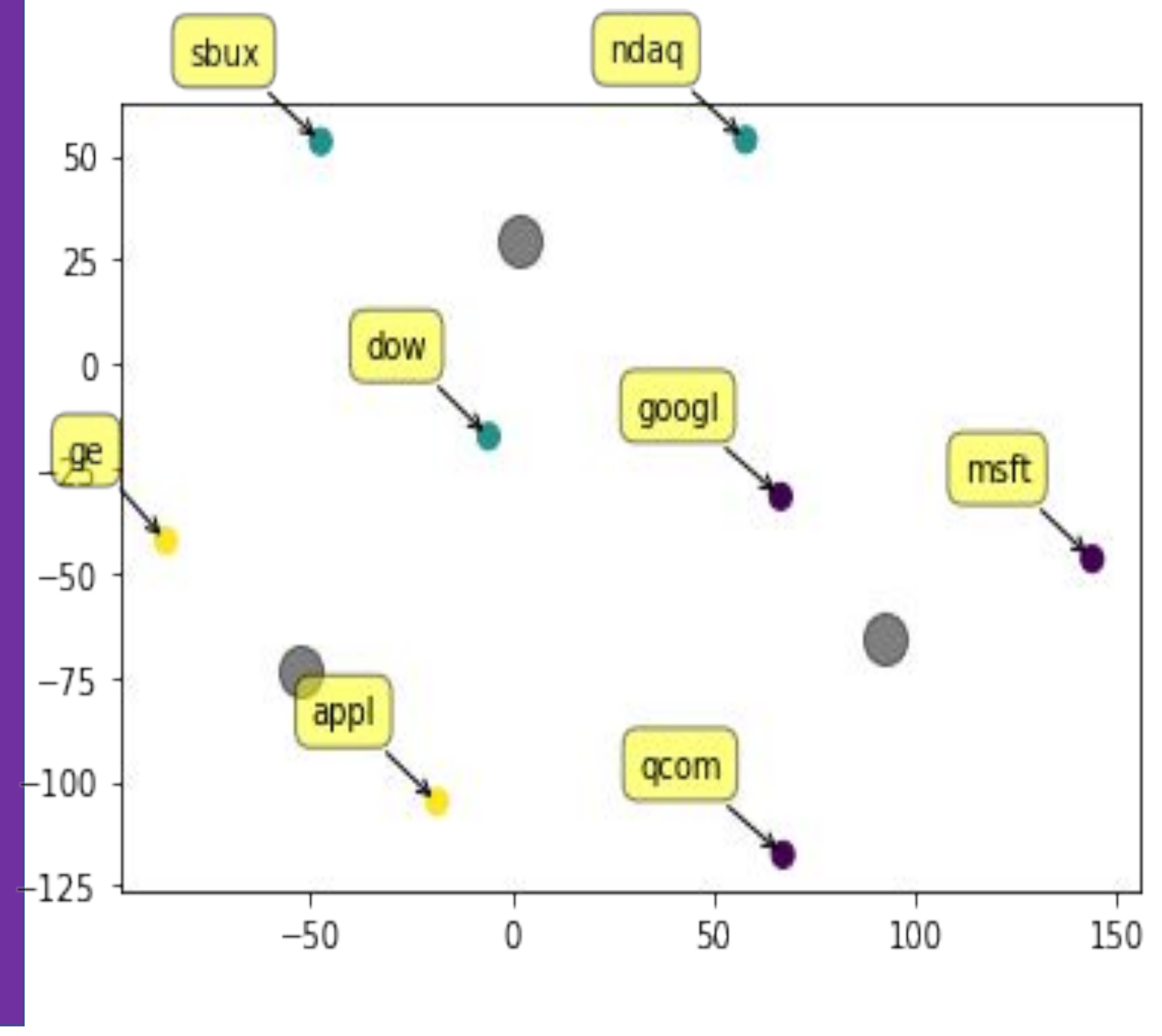

Figure 5: tSNE on company stock price

Mathematical Theory

This problem framework begins with time-series of the dynamical system with each vector y representing system at one time step with n variables. Next, we can compute the derivative y' through various difference schemes from scientific computing and this serve as the right hand side b for our regression framework to solve A x = B.

Intuitively, this approach works by building a library of functions from scientists' knowledge about the system that they are considering to be good candidates for the system and use sparse regression to evaluate which equations are more suitable to the given data.

$$\Theta = \{x,\ y,\ x^2,\ xy,\ y^2,\ x^3,\ x^2\ y^2,\ x^4, \dots\}$$

Where library functions are represented as a matrix with each column is the corresponding library function at each timestep ytand this matrix serve as matrix A in our regression framework. By using sparse regression algorithm such as Lasso, the algorithm can extract important coefficient that correspond to the library function that matches with the derivative y' of the system.

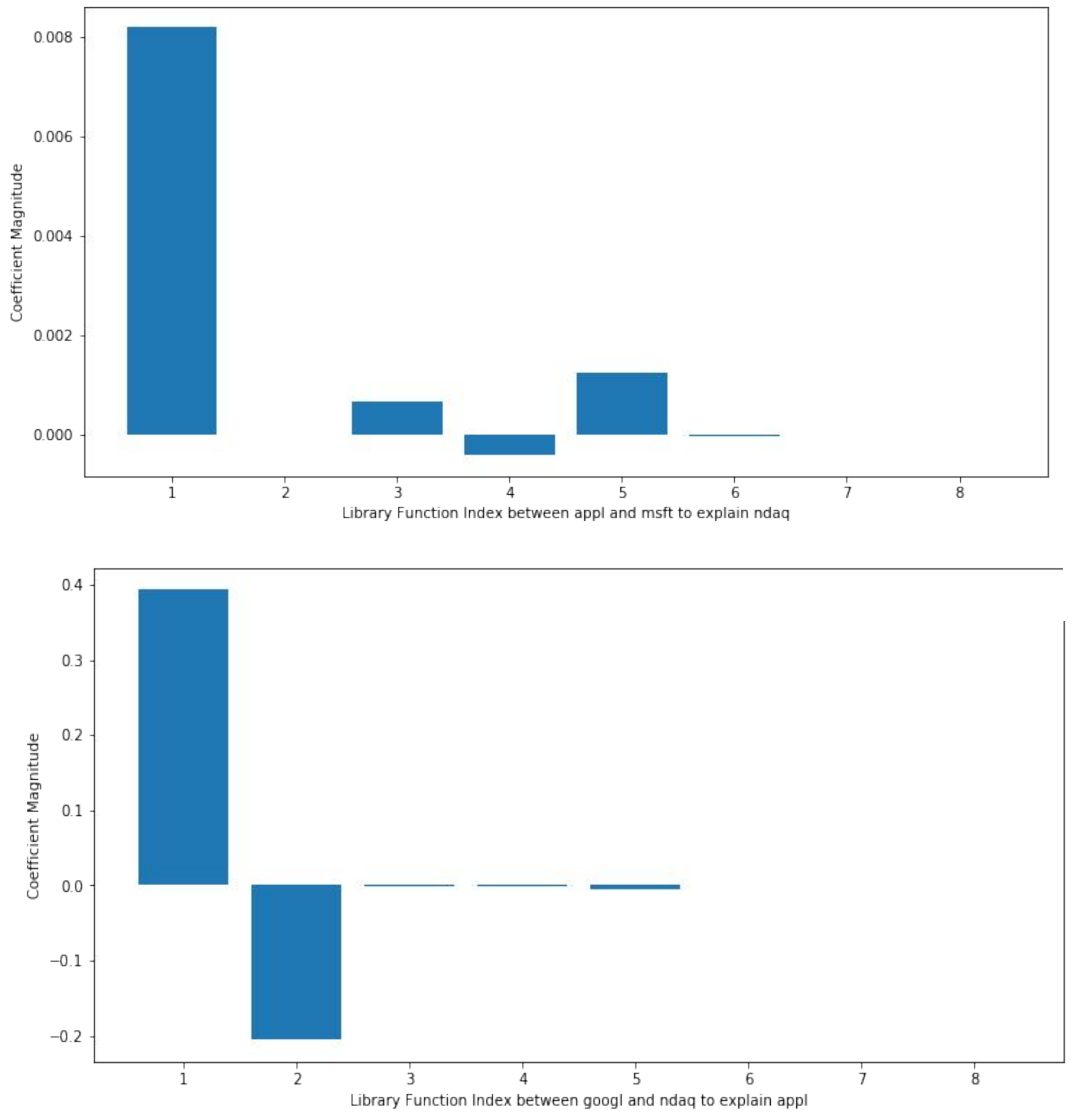

Figure 6: Sparse regression results with respect to library functions

Preprocessing Tool for Prediction Framework

This learned dependency can be used as a variable selection criterion to improve sequence learning accuracy in neural networks such as LSTM neural network and Transformers. This technique proposes a crucial hyper-parameter tuning step in time series data analysis in a data-driven way to discover and interpret the dependencies among variables to select qualitative components to improve prediction accuracy.

This new perspective enables us to visualize the dependencies among input variables to predictive models, which can help us distinguish important variables to the output rather than picking uncorrelated input that can decrease the accuracy.

The typical usage of neural network involves collecting a huge amount of data, but to collect valuable and important data to the output is even more important.

This machine learning approach to discover models provides a unique approach to visualize how variables in the dynamical system are interacting and we can extract mathematical equations from such scientific systems by experimenting expert knowledge to evaluate variable relationship through library functions.

Results

Company	Correlation Value
Dow	-0.355
Msft	-0.642
Googl	-0.35
Ndaq	-0.14

Table 1: Correlation value between edge weights with respect to the output versus RMSE prediction loss with LSTM network

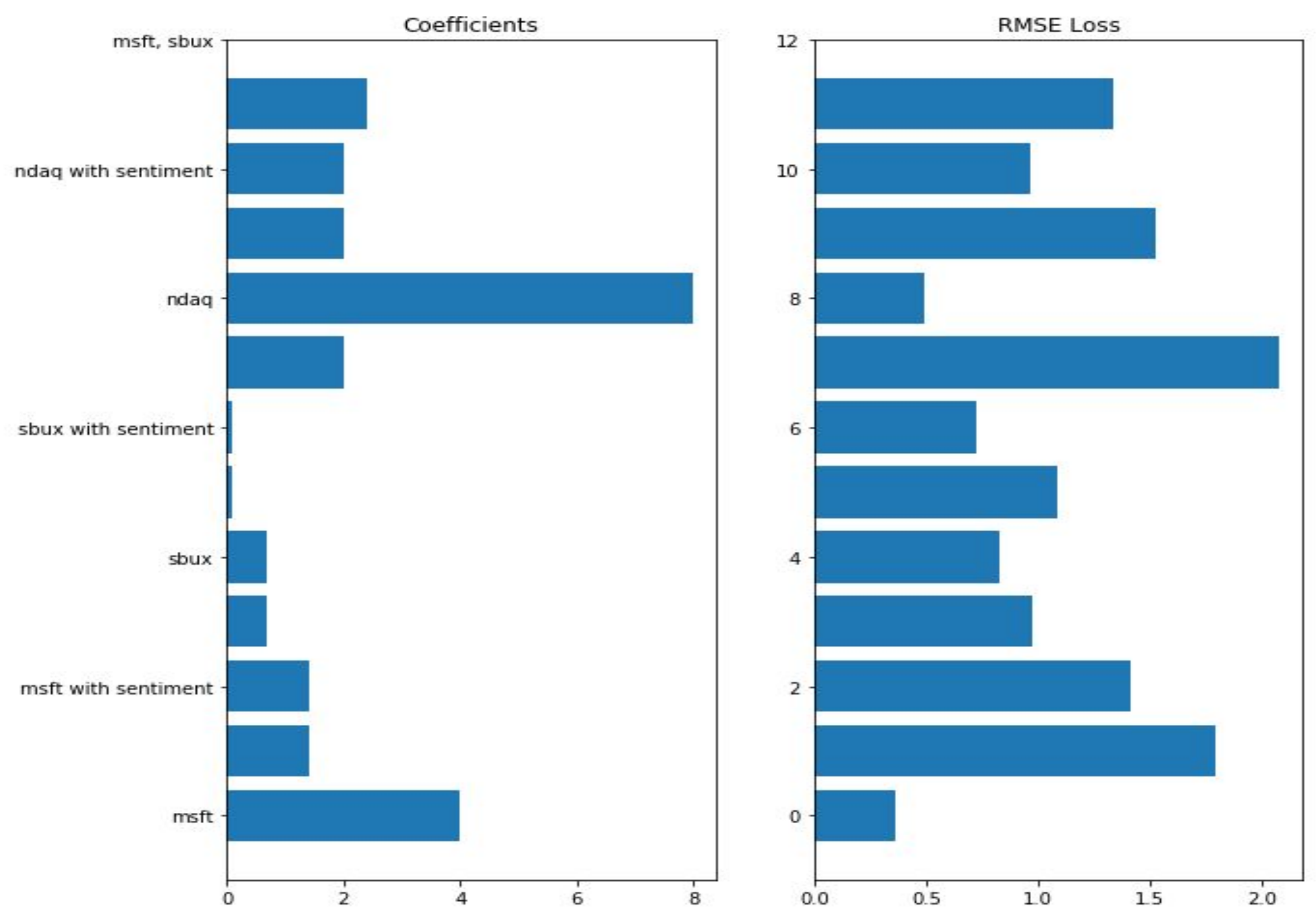

Figure 7: RMSE loss result with many variable combinations with LSTM network for Dow stock price

Input Variables	Outcome Variable	RMSE loss
🟥 Ndaq	Ndaq	0.661
🟥 Dow, Googl	Ndaq	0.395
🟩 Appl, Googl	Ndaq	1.031
🟥 Msft	Googl	0.984
🟩 Ndaq	Googl	25.2
🟥 Msft, Ndaq	Googl	15.4
🟩 Ndaq	Dow	0.723
🟥 Msft	Dow	1.792
🟥 Msft, Ndaq	Dow	0.963
🟥 Sbux	Dow	0.825
🟥 Msft, Sbux	Dow	1.338

Table 2: RMSE prediction results with various variable combinations.

- Selecting high coefficients and avoid low coefficients with respect to the outcome variable is not enough, we need to consider the interaction among the predictor variables as well.
- Furthermore, we need to weigh the contrasting relationship between stock price and sentiment data edge weights as well.

FUTURE APPLICATION

This paper apply the idea of sparse regression to perform model discovery on dynamical system to extract mathematical equations that describe the underlying mechanism of complex system. This idea plays a vital role in science applications since it is a data-driven approach to give scientists a hint in understanding the mathematics under high-dimensional data that people may not be able to see through. For instance, this approach can help scientists extract insights and understand more about physical phenomena such as turbulence, fluid flow or complex biological systems such as biological neural networks with a variety of free variables.

Further work

- Experimenting with larger amount of data and try different prediction models such as Transformer neural network.
- Establish new algorithms that choose predictor variables intelligently with respect to the outcome variable as well as considering the relationship among the predictor variables such as Maximum Spanning Trees (MST) and Dijkstra's algorithm.