Siu King Sum

📂 **Complete Source Code and Model Output** — Available on GitHub
🔗 **View the Code and Results on GitHub**

Data-Driven Tech Stock Analysis: U.S. vs. China

Every day, the stock market is flooded with news and social media content, making it easy for investors to be swayed by emotions. The goal of this project is to integrate two key analytical methods—valuation and sentiment analysis—using real data, helping us break free from intuition and subjectivity. By relying on data, I aim to make more rational and stable investment decisions.

I selected 41 technology-related companies from the U.S. and China—24 from the U.S. and 17 from China. Using `yfinance`, I automatically retrieved key valuation, including PE, PB, PS, and EV/EBITDA.

In addition to valuation data, I used a web scraping program to collect the latest news headlines for each company from the website FinViz. These headlines were then processed using the VADER sentiment analysis tool, which assigns a sentiment score ranging from -1 (very negative) to +1 (very positive). We calculated the average sentiment score for each company to obtain a "News Sentiment Score."

Since indicators like PE and PB have different units and ranges, they cannot be directly compared. Therefore, I used the z-score method to convert each indicator into standardized units. A z-score represents how much a number is above or below the industry average. Because I want higher scores to represent cheaper valuations, I multiplied the z-score by -1, so that companies with lower valuations receive higher scores.

After the transformation, I averaged the standardized scores of the four valuation indicators for each company to obtain a Composite_Z score. This score represents how undervalued a company is overall. Weibo (WB) has a Composite_Z of 0.772, indicating that its valuation is cheaper than the industry average. Tesla (TSLA) has a Composite_Z of around -2.148, suggesting that its current stock price is expensive (Figure 1).

Figure 1

Summary of Valuation Metrics and Standardized Scores

Summary of Valuation Metrics and Standardized Scores:

Ticker	Market	Sector	PE	PB	PS	EV_EBITDA	z_PE	z_PB	z_PS	z_EV_EBITDA	Composite_Z
WB	China	Communication Services	7.905173	0.626580	1.313997	3.235	1.681267	0.542198	0.253670	0.612099	0.772308
BIDU	China	Communication Services	9.900772	0.117674	0.233022	6.072	1.313894	0.726199	0.634896	0.410265	0.771314
BILI	China	Communication Services	NaN	0.564823	0.295737	-3.630	NaN	0.564527	0.612779	1.100499	0.759268
INTC	US	Technology	NaN	0.978365	1.841924	17.476	NaN	0.816779	1.165292	0.264757	0.748942
QCOM	US	Technology	15.039913	5.733437	3.789034	12.267	1.121335	0.435861	0.766367	0.411365	0.683732
ZM	US	Technology	22.819315	2.506501	4.792350	15.610	0.763763	0.694363	0.560808	0.317276	0.584052
DOYU	China	Communication Services	NaN	0.050209	0.052000	8.259	NaN	0.750592	0.698737	0.254674	0.568001
XPEV	China	Consumer Cyclical	NaN	0.638626	0.487407	-5.724	NaN	0.480933	0.381009	0.762490	0.541477
JD	China	Consumer Cyclical	10.813008	0.241971	0.052082	0.962	0.584276	0.573948	0.523860	0.465213	0.536824
CSCO	US	Technology	25.135965	5.006115	4.208428	16.641	0.657281	0.494125	0.680442	0.288258	0.530026
DADA	China	Consumer Cyclical	NaN	0.138498	0.051308	1.650	NaN	0.598212	0.524114	0.434622	0.518983
PDD	China	Consumer Cyclical	10.867112	0.503999	0.400830	2.845	0.583226	0.512503	0.409419	0.381489	0.471659
NIO	China	Consumer Cyclical	NaN	1.296361	0.117464	-0.843	NaN	0.326695	0.502405	0.545468	0.458189
BABA	China	Consumer Cyclical	18.880293	0.299013	0.314764	12.073	0.427600	0.560572	0.437662	-0.028812	0.349255
IBM	US	Technology	37.867810	8.259499	3.597908	25.826	0.072075	0.233504	0.805525	0.029744	0.285212
IQ	China	Communication Services	19.000000	0.150544	0.068865	13.077	-0.361196	0.714315	0.692789	-0.088095	0.239453
ADBE	US	Technology	24.272968	12.199780	7.102689	17.963	0.696948	-0.082143	0.087465	0.251050	0.238330
YQ	China	Consumer Defensive	NaN	0.036328	0.067765	-1.856	NaN	0.707107	0.707107	-0.707107	0.235702
CRM	US	Technology	40.130500	4.013744	6.472517	21.836	-0.031928	0.573621	0.216575	0.142044	0.225078
NTES	China	Communication Services	16.604431	2.401758	0.647555	6.611	0.079808	-0.099639	0.488703	0.371919	0.210198
TCEHY	China	Communication Services	22.868328	0.605267	0.886795	2.951	-1.073322	0.549904	0.404331	0.632304	0.128304
SNOW	US	Technology	NaN	15.570395	12.888991	-34.117	NaN	-0.352156	-1.098034	1.716852	0.088887
MSFT	US	Technology	30.041061	9.164619	10.594618	19.748	0.431823	0.160996	-0.627962	0.200811	0.041417
...											
GOOG	US	Communication Services	18.983831	5.733228	5.279069	13.846	-0.358220	-1.304173	-1.144689	-0.142804	-0.737471
NVDA	US	Technology	34.625850	31.410060	19.034307	29.418	0.221088	-1.621036	-2.357087	-0.071353	-0.957097
META	US	Communication Services	22.271470	7.375928	8.188062	15.754	-0.963446	-1.898111	-2.170601	-0.278546	-1.327676
TSLA	US	Consumer Cyclical	131.665020	11.788990	8.800404	64.292	-1.762810	-2.133809	-2.346882	-2.350606	-2.148527

I also created a bar chart showing the composite valuation scores of all companies, arranged from left to right (Figure 2). This highlights which companies are relatively cheap and which are more expensive. Next, I used a boxplot to compare the valuation distribution between U.S. and Chinese companies (Figure 3). The results showed that Chinese companies have lower valuations and smaller volatility compared to their U.S. counterparts.

Figure 2

The composite valuation scores

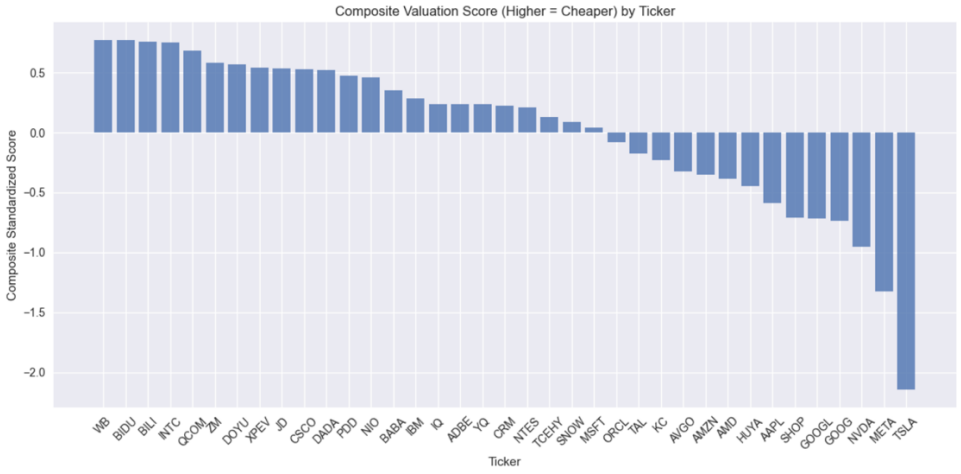
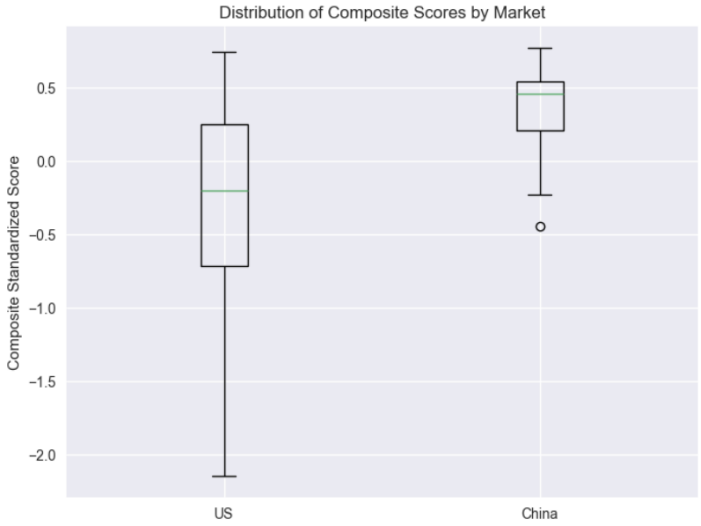


Figure 3

The valuation distribution between U.S. and Chinese companies



I also plotted a scatter plot of z-score PE versus z-score PB (Figure 4). The chart shows that U.S. companies are widely dispersed—some are cheap, others are expensive—whereas Chinese companies are clustered in the top right, indicating that they are generally more attractively valued.

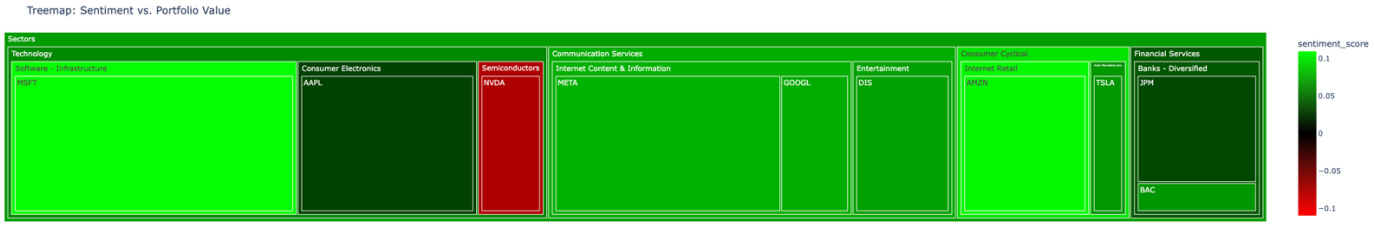
Figure 4

A scatter plot of z-score PE versus z-score PB

The sentiment scores derived from news headlines allow us to quantify the current market sentiment toward each stock. For example, Amazon (AMZN) recently had the highest average sentiment, close to +0.10, while NVIDIA (NVDA) had the lowest, around -0.08. I visualized this using a Treemap that combines market capitalization with sentiment scores, allowing us to see not only which stocks have the most positive or negative sentiment, but also which types of stocks make up the largest portions of an investment portfolio (Figure 5).

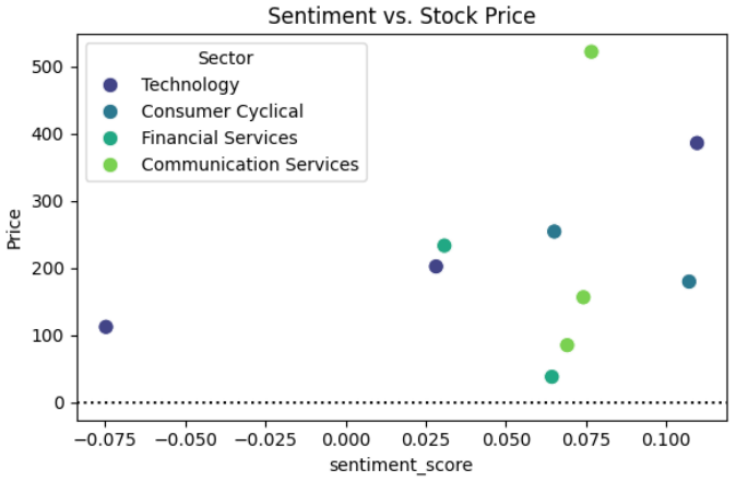
Figure 5

A Treemap that combines market capitalization with sentiment scores



I also created a scatter plot showing the relationship between each stock's sentiment score and its stock price. The overall trend suggests that stocks with higher sentiment scores tend to have higher prices, indicating a positive correlation between sentiment and price (Figure 6).

Figure 6

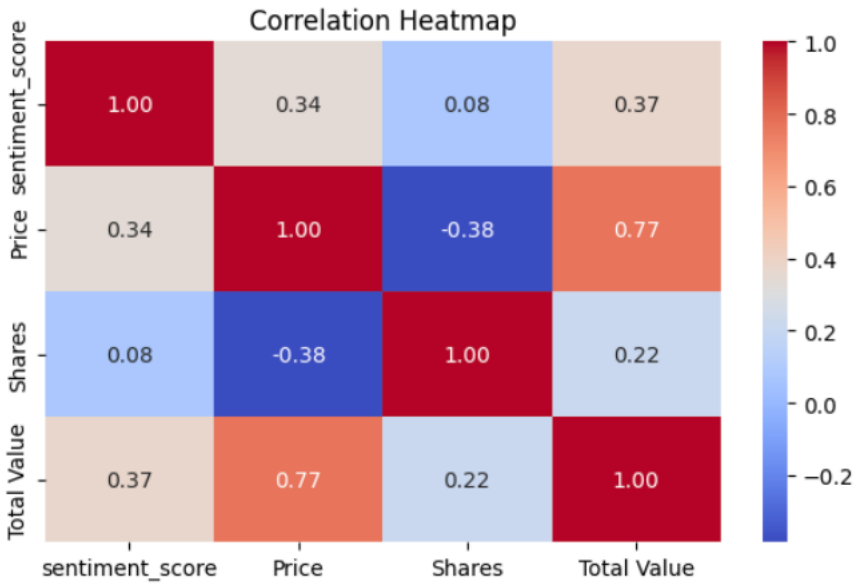


Finally, I used a heatmap to examine the correlations between different variables. I calculated the correlation coefficients among four variables: sentiment score, stock price, number

of shares held, and market capitalization. The results showed that the correlation between sentiment and price was around 0.34, indicating a moderately positive relationship. The highest correlation was between price and market capitalization, at 0.77, which makes sense because market cap is calculated by multiplying price by the number of shares. There was a negative correlation of -0.38 between price and number of shares held, suggesting that investors tend to buy more shares of cheaper stocks.

Figure 6

A heatmap to examine the correlations between different variables



Through this data analysis, I built an investment analysis system that evaluates both valuation and sentiment. This system not only helps us make rational decisions about whether a stock is undervalued or overvalued, but also helps us understand market sentiment with data, reducing the chances of blindly chasing highs or selling at lows.

In the future, this framework could be expanded to include more data sources—such as text analysis from social media, technical indicators, or financial growth metrics—to further enhance the model's comprehensiveness. It is a dynamic tool, capable of real-time updates as markets evolve, and serves as a highly practical data-driven assistant for investment decision-making.