Do Stocks Create News, or does the News Drive Stocks?

Group number: G08

Wes Bailey, Trupti Abnave, and Abubeker Abdullahi

CPSC 6030, Fall 2020

Project Report

Introduction and Problem Identification

The stock market is the marketplace for buying and selling equities of publicly traded companies. Stock sales allow the public to invest in ownership of companies, and a means to raise capital. The stock market data is enormously large and complex. There are thousands of companies, and prices are constantly fluctuating. Stock performance is influenced by various factors including the news. We designed our stock market visualizations through the lens of news and opinion data.

Our team proposes to create a visual analysis of the effects of news and opinion stories on trading volumes.

1) Do stock prices drive news cycles or do news stories drive the trading floor?

2) Is it possible to evaluate a dataset of historical stock performance with respect to a dataset of historical financial news to gain insight and answers?

Dataset

We were curious about the possible correlation between news articles and stock sales. We located two datasets from the Kaggle website[9] that provided the needed information.

The first dataset is, Historical financial news archive: This news archive covers the last 12 years of US equities publicly traded on NYSE/NASDAQ which still has a price higher than $10 per share. This data includes full-text stories, dates, sources, categorization of news or opinion, and even links to the original online content.

The second dataset is, AMEX, NYSE, and NASDAQ stocks histories: This dataset contains almost all the stocks listed on these exchanges as of the date shown in the file name. The most recent date in the set is 6/12/2020. Due to the very large quantity of data, we ultimately reduced the scope to a five year window from 2015 through 2019.

Challenges:

Since we found a very extensive resource on AMEX, NYSE, and NASDAQ stocks histories, we quickly discovered that the sheer magnitude of this data made it unmanageable. The set was explored using Tableau in order to find the best way to focus on and reduce the quantity. We decided to narrow the scope to a 5 year period from 2015 through 2019.

The data was sorted and we selected the best 5 stocks that had a high volume of articles in the media source as well as fairly high trading volume. The decision to use volume as a comparative metric was chosen since the prices varied significantly, and trading volume seemed to be a reasonable way of normalizing the comparison.

Design Solutions:

Visual representation is one of the most efficient ways to assist investors to have a clear overview of movements of the stock market, as well as providing a deeper understanding of each individual stock. There is a lot of information in the raw datasets, so we created the interactive graphs in order to facilitate further exploration of the content and the possible relationship between the stories and the stocks.

Having narrowed the data to the years 2015 through 2019, we considered five well-known stocks that we represented using blue, aqua, green, gray, and red for Apple (AAPL), Amazon (AMZN), Bank of America (BAC), Microsoft (MSFT), and Tesla (TSLA) respectively. The data was sorted to determine the best 5 stocks that had a high volume of articles in the media sources as well as having fairly high trading volume. We used volume as a comparative metric to make a decision, the results showed us the trading volume seemed to normalize the comparison whereas the prices varied significantly.

The story data was categorized as either news or opinion. We correlated the stories by the referenced ticker symbol and their publication date. All of the negative and positive changes in the market are reflected in the media stories in some measure. For example in 2015, analysts at Credit Suisse, Morgan Stanley, and Stifel all reduced their estimates for Apple's iPhone sales that led to a 10% decrease in the Apple stocks in 2015.

Our final visualization is divided into two areas, one focused on the stock data and another focused on the story data.

1. Multiline Diagram:

When our team approached the data for initial evaluation we instinctively used a line graph since it so naturally presents time-series data in an understandable format. In the final visualization, we stayed with this paradigm, and it is difficult to imagine a better format for presenting a set of regularly-spaced data depicting values over a time interval. In a 2015 paper, *Task-Based Effectiveness of Basic Visualizations*, the authors noted that line charts perform very well at depicting trends as well as correlations in data, particularly where there is a pairwise comparison at a data point.[4]

Because our data was grouped as a time series, it is natural to present as a multi-line chart, however, this does lead to visual clutter when the graph is viewed at a wide time scale. In order to alleviate and to add meaningful interactivity, we added a 'selection brush' feature that permits the viewer to control the bounds of the time scale in order to reduce the data to a more digestible scope. We also added checkboxes to allow the viewer to isolate the graph to show only the desired stocks. We used these affordances to enhance usability on a desktop interface, but it is worth noting that they are useful tools to create manageable visualizations for mobile devices.[3] As digital consumption on handhelds becomes increasingly ubiquitous, we must be mindful and always design for compatibility with these smaller form-factors.

Data-Driven Documents (D3) is used to create a Stock Volume multi-line graph showing the daily trading volume for 5 stocks over a 5 year period. The chart is filtered in two ways: one, to show data for one or more of the stocks by selecting a checkbox and updating the chart. Secondly filtered to show a specific time frame for the chosen data by using a selection bar at the bottom of the chart. The user can select either end of the gray segment of the bar and move to a date, and can also drag the gray segment to a different date range.

We used the following marks and channels in our multi-line chart:

Marks:

Line (1D) - to represent the stock volumes over time

Channels:

Position (both, horizontal and vertical) - X and Y axes

Color - used to represent the different stock ticker symbols

2. Sankey Diagram:

We used a sankey diagram to describe the links between story types and stock symbols. Sankey diagrams are used to visualize flows in which the width of the lines is proportional to the flow rate, illustrating the flow from one set of values to the other. The attributes being connected are known as nodes, whereas the connection is called the link. Sankeys are mainly used to show a many-to-many mapping relationships between two domains or multiple paths through a set of stages.

Our News Stories Sankey diagram shows the segmentation of news and opinion stories for each of the stocks. The chart is synchronized with the selected stocks and date range of the previous stock trading graph. When the user moves their cursor over the links of the sankey chart, a tooltip shows information including the type of story and the stock symbol, the source and date, and the headline. Additionally, the link is highlighted in blue, indicating that this is a link. When clicked, the link will open the original online story in a new tab.

We used the following marks and channels in our sankey chart:

Marks:

Area (2D) - the nodes representing the stocks and stories vary in area based on size of the data point

Channels:

Size - the overall width of the links between the nodes represents the number of news and opinion stories.

Color - to represent news vs. opinion and the different stock ticker symbols

We received the following comments to our initial submission of our visualization:

- Single lines in the Sankey diagram should be highlighted with a stronger color.
- A suggestion for the line chart. Start the webpage with a small timeframe and let the user expand it. This way, they will encounter problems with efficiency only if they really want to.

We made the following changes to address these suggestions:

- Changed the default time span (when the page first loads) from 5 years to approximately 1 month. We also added text below the chart to provide instruction for use.

- Revised the reload function for stock choices. Now, if a user selects or removes symbols while zoomed into a specific timeframe that timeframe stays constant when the page reloads.
- Modified the color of the sankey links to the darkest available shade of blue. The bigger issue for viewing these links had to do with the scale of the selected time span in the multi line graph. Now that the default selection is set for a shorter time, the links are visible as planned.

Literature review

Compared with traditional data mining, the combination of visualization methods can help people understand data faster. In the massive data, a large part contains time attributes, which belong to time series data.[6] The key problem in visualizing the stock market data is not only to allow readers with a graphical representation of the time-series stock market data but also to design an interactive visualization that is navigable to extract the desired information. The usual visualization of the stock market data is a 2-D figure with the x-axis representing the time and the y-axis the stock price. This method will be ideal if the visualization is used to represent one stock. However, when visualizing multiple stock symbols on the same plot, a multi-line graph would be ideal. The average number of traded stocks would be the best metrics to use for the y-axis by keeping the x-axis the same, as stock prices for different companies vary widely in price.

Multi-line graphs visualize trends among dense datasets. Representing time-dependent data plays an important role in information visualization since time is an abstract concept. Graph interactivity can be added to represent developments over time. A large number of research papers have been published to solve the problem of representing time-dependent datasets. According to the research paper [Simone Kriglstein, Margit Pohl, and Michael Smuc][7], time-dependent datasets can be categorized into two based on visual variables:

1) Space: this method discusses that time can be represented by space by which the presentation of the length of lines in space is used to represent the length of lines.

2) Animation: this method uses animation to represent the development and changes over time, where each frame in the graph represents a data point in time.

Our multi-line chart uses a static spatial rendering of the stock data as multiple lines, each representing a single equity. The axes are dynamically adjusted based on the domain of the selected equities.

Sankey diagrams are used to visualize flows of materials and energy in many applications, to aid understanding of losses and inefficiencies, to map out production processes, and to give a sense of scale across a system.[8] These diagrams are ideal to visualize the links and connections between two entities, where the width of the line connecting the nodes with the links is proportional to the flow rate and also the sum of the incoming weights for each node is equal to its outgoing weights. They can also be used to visualize multidimensional data in a clear and meaningful way. In recent years, Sankey diagrams have become a standard model used in science and engineering to represent multiple visual representations.

Although originally used to indicate flows of energy and other material resources, sankey diagrams have been adopted within the field of data visualization to depict a variety of multidimensional data.[1] The concept is discussed in-depth in a 2006 paper, *Parallel Sets: Interactive Exploration and Visual Analysis of Categorical Data*, in which the authors note that categorical data is often organized hierarchically by the viewer in an effort to aggregate and organize complex sets.[2] And although our dataset had a low number of dimensions and hierarchy, the pattern can be effectively applied to much higher-dimensional and deeper datasets.

We felt that the sankey depiction was a good choice, since the sankey links would be used as literal hyperlinks in our implementation. Additionally, this design offers a clear and unambiguous visual depiction of the data in a logically segmented layout that should be intuitive for viewer exploration.

References

[1] R.C. Lupton, J.M. Allwood. 2017. "Hybrid Sankey diagrams: Visual analysis of multidimensional data for understanding resource use", Resources, Conservation and Recycling, Volume 124, Pages 141-151. https://www.sciencedirect.com/science/article/pii/S0921344917301167

[2] R. Kosara, F. Bendix and H. Hauser, "Parallel Sets: interactive exploration and visual analysis of categorical data," in IEEE Transactions on Visualization and Computer Graphics, vol. 12, no. 4, pp. 558-568, July-Aug. 2006, doi: 10.1109/TVCG.2006.76.

[3] Trevor D'Souza, Padmalata V. Nistala, Swapna Bijayinee, Sonali Joshi, Prachi Sakhardande, and Kesav V. Nori. 2017. Patterns for Interactive Line Charts on Mobile Devices.Proceedings of the 22nd European Conference on Pattern Languages of Programs. Association for Computing Machinery, New York, NY, USA, Article 21, 1–13. DOI:https://doi-org.libproxy.clemson.edu/10.1145/3147704.3147727

[4] Bahador Saket, Alex Endert, and Çagatay Demiralp. 2015. "Task-Based Effectiveness of Basic Visualizations". Journal of LATEX Class Files, Vol 14, No 8. https://arxiv.org/pdf/1709.08546.pdf

[5] Krešimir Šimunić (2003). Visualization of Stock Market Charts. In *In Proceedings from the 11th International Conference in Central Europe* on Computer Graphics, Visualization and Computer Vision 2003 (2003), Plzen-Bory (CZ), 2003.

[6] Fang, Y., Xu, H., & Jiang, J. (2020). A Survey of Time Series Data Visualization ResearchIOP Conference Series: Materials Science and Engineering, 782, 022013.

[7] Kriglstein, Simone & Pohl, Margit & Smuc, Michael. (2013). Pep Up Your Time Machine: Recommendations for the Design of Information Visualizations of Time-Dependent Data. 10.1007/978-1-4614-7485-2_8.

[8] Riehmann, Patrick & Hanfler, M. & Froehlich, B.. (2005). Interactive Sankey diagrams. Proceedings - IEEE Symposium on Information Visualization, INFO VIS. 233 - 240. 10.1109/INFVIS.2005.1532152.

[9] Kaggle. 2020. https://www.kaggle.com/