Forecasting Stock Prices Using News Headlines

DANIEL FELBERG

GEORGE WASHINGTON UNIVERSITY

Literature Review (1)

- Marketing Research
  - *Corporate Reputation*
    - *How well a firm/company is perceived by the mass public in general*
    - *Paid vs. Owned vs. Earned Media*
  - *Earned Media*
    - *Any public exposure a company may receive from media coverage*
- Problem → existing research focuses on *Corporate Repuation* generally, less so on *earned media* specifically
  - Alloza, Carreras, and Carreras (2014), Fernández-Gámez, Gil-Corral, Galán-Valdivieso (2016)
- **Goal: help demonstrate relationship between *earned media* and how it impacts a firm's finances (i.e. stock prices)**

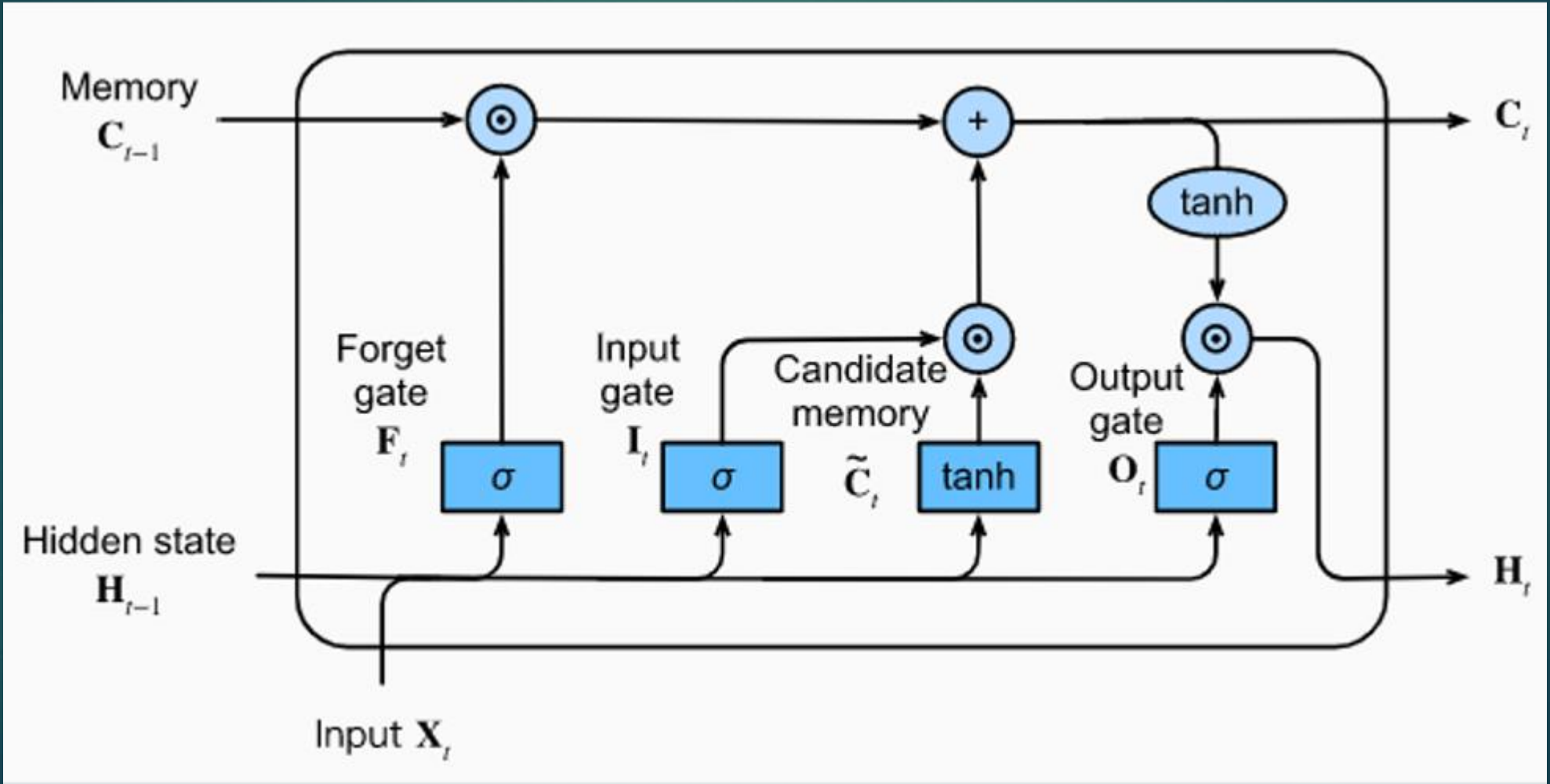
Literature Review (2)

- Sentiment Analysis and Time-Series Forecasting
  - Hassan (2022) → forecasted directional movement of the S&P500 using Twitter data more successfully than baseline models
- Advancement of LSTM models
  - Bu, Li, Li, and Wu (2020) → successfully forecasted with high accuracy the next-day opening prices of the CSI 300 index using messages posted to *Eastmoney.com*, a Chinese stock message board
  - Asgarov (2023) → forecasted Tesla's and Apple's stock prices using Twitter data
- **Reiterating Goal: help model relationship between news media (i.e. *earned media*) and stock prices**

LSTM Models Explained

- Long Short-Term Memory (LSTM) is a deep recurrent neural network (RNN) method
- How they work:
  - **Highlighter**: marks important things
  - **Eraser**: deletes things that are no longer important
  - **Pencil**: writes down new things you learn
- You are reading a book one sentence at a time:
  - Day 1 → weather is sunny, highlight for predicting weather tomorrow
  - Day 2 → weather is cloudy, erase "sunny" and write "cloudy" instead
  - Day 3 → weather is raining, update book once again

LSTM Models Explained 2



Data Collection

- Three steps:
  - 1) Scraping news headlines
    - Custom scraper (`Beautiful Soup`) → *Financial Times*
  - 2) Perfom sentiment analysis
    - *`distilroberta-finetuned-financial-news-sentiment-analysis`*
  - 3) Fetch daily closing stock
    - Yahoo Finance API
    - Top 10 companies globally by market cap

Scraping News Headlines

- Using `Beautiful Soup` to scrape from *Financial Times* website
  - Scraper accesses "search" URL, and iterates through the pages

```
url = f"https://www.ft.com/search?q={company}&page={page_num}&sort=relevance&isFirstView=false"
```

- Returns headlines and their date of publication as an index:

index	Headline
2024-12-10	Nvidia's profit smorgasbord attracts an ant in...
2024-12-09	China launches antitrust probe into Nvidia
2024-11-21	Nvidia's revenue nearly doubles as AI chip dem...
2024-08-29	Nvidia brings out the crowd but not the fireworks
2024-08-29	Nvidia faces looming test on use of chips
	...
2024-09-04	FirstFT: Nvidia overtakes Apple and Microsoft ...
2024-09-05	Markets update: Asian currencies strengthen af...
2024-08-09	European stocks edge higher as Nvidia propels ...
2024-04-19	Management consultants find sweet spot in the ...
2024-06-19	European stocks hit one-month high in wake of ...

[144 rows x 1 columns]

Sentiment Analysis

- `distilroberta-finetuned-financial-news-sentiment-analysis`
  - Model was trained specifically on a dataset consisting of 4,840 financial news sentences from the English language
  - Ideal fit for our needs!
- For every headline, returns "Label" (positive, neutral, negative) and "Score" (respective label's score)
- Re-map labels to 1, 0, -1 (respectively)

Stock Data

- Yahoo Finance API → obtained stock data from top 10 global companies by market cap
  - NVIDIA
  - Apple
  - Microsoft
  - Amazon
  - Google
  - Aramco
  - Facebook
  - Berkshire Hathaway
  - TSMC
  - Tesla
- Returns daily closing stock data and date index for each value

Data Structures

Index	Headline	Label	Score	Numeric Label
'2024-12-09'	China launches antitrust probe into Nvidia	Neutral	0.998862	0
'2024-11-21'	Nvidia's revenue nearly doubles as AI chip dem...	Positive	0.999671	1
'2024-08-29'	Nvidia brings out the crowd but not the fireworks	Neutral	0.999871	0
'2024-08-29'	Nvidia faces looming test on use of chips	Neutral	0.998440	0
'2024-12-03'	The geopolitics of chips: Nvidia and the AI boom	Neutral	0.999856	0

```
In [9]: stock_values
Out[9]:
Date
2023-08-24     47.162998
2023-08-25     46.018002
2023-08-28     46.834999
2023-08-29     48.784000
2023-08-30     49.264000
                   ...
2024-12-03    140.259995
2024-12-04    145.139999
2024-12-05    145.059998
2024-12-06    142.440002
2024-12-09    138.809998
Name: Close, Length: 326, dtype: float64
```

Methodology (Develop LSTMs)

- **Four models were developed:**

LSTM1 (no headlines)

Layer	Parameter
Input	(1,1)
LSTM	25
LSTM	25
Dense	25
Dense	1

LSTM2 (no headlines)

Layer	Parameter
Input	(1,1)
LSTM	25
LSTM	25
Dense	1

LSTM3 (with headlines)

Layer	Parameter
Input	(1,2)
LSTM	25
LSTM	25
Dense	25
Dense	1

LSTM4 (with headlines)

Layer	Parameter
Input	(1,2)
LSTM	25
LSTM	25
Dense	1

- *For all models: sequence length/time steps = 1, epochs=20, batch size = 32*
- *Train/Test split set to 80/20*

Challenges Encountered: Joining Headline and Stock Data

- Problem: several dates had more than one news article published
  - Example (Microsoft): `The most frequent date is 2016-07-19 00:00:00 with 3 occurrences.`
- Solution: create “padding” so all lists are the same length (using netural value of `0`)
  - Disadvantage: forced to “compress” values back down using mean
    - “Mutes” potential noise in the data
- Something to keep in mind going forward!

Results 1

TABLE VI.

Company	Ticker	LSTM1	LSTM2	LSTM3	LSTM4
NVIDIA	NVDA	5.0781	12.6722	5.1487	11.6729
Apple	AAPL	5.7745	12.422	4.8278	9.7018
Microsoft	MSFT	5.5537	10.6744	6.6481	8.2138
Amazon	AMZN	3.9383	7.6315	4.3874	6.5533
Google	GOOG	6.0616	11.1451	4.5245	9.2089
Aramco	2222.SR	0.2355	0.2756	0.2431	0.2641
Facebook	META	16.447	47.2765	13.8866	27.283
Berkshire Hathaway	BRK-A	9066.7334	38368.7457	9150.6053	24210.8152
TSMC	TSM	7.6242	10.5543	7.9189	9.1132
Tesla	TSLA	10.25	11.0606	10.0546	10.6955

Fig. 6. LSTM modelling results – First iteration

Results 2

TABLE VII.

Company	Ticker	LSTM1	LSTM2	LSTM3	LSTM4
NVIDIA	NVDA	5.5762	8.5023	4.5493	4.8111
Apple	AAPL	5.3571	10.5778	4.9817	9.6739
Microsoft	MSFT	5.1183	13.6263	6.297	7.7407
Amazon	AMZN	4.0844	6.5961	3.5374	6.0456
Google	GOOG	5.5509	9.1127	6.0938	8.6193
Aramco	2222.SR	0.22	0.2811	0.2213	0.2538
Facebook	META	15.537	37.2793	10.969	32.3243
Berkshire Hathaway	BRK-A	15107.4927	38363.4021	10855.5346	33679.1282
TSMC	TSM	7.2026	14.4975	8.051	11.5392
Tesla	TSLA	10.9939	11.0201	10.8459	12.0311

Fig. 7. LSTM modelling results – Second iteration

Results 3

TABLE VIII.

Company	Ticker	LSTM1	LSTM2	LSTM3	LSTM4
NVIDIA	NVDA	4.8533	6.3436	4.9234	10.7417
Apple	AAPL	6.0234	10.7918	5.4215	9.3888
Microsoft	MSFT	5.1725	10.7472	5.9288	9.4753
Amazon	AMZN	3.8302	7.4784	4.1637	6.3794
Google	GOOG	5.7441	9.7828	7.3071	8.4759
Aramco	2222.SR	0.2299	0.2633	0.2904	0.2534
Facebook	META	16.7271	39.8547	11.7942	27.1295
Berkshire Hathaway	BRK-A	16465.6218	36633.0754	7663.6153	22896.0987
TSMC	TSM	7.4101	11.9626	8.9927	13.4486
Tesla	TSLA	10.3694	10.7385	10.7166	10.9173

Fig. 8. LSTM modelling results – Third iteration

Conclusion

- **Models are not conclusive enough to help establish a direct relationship between *earned media* and stock prices**
- BUT, results also show it may be possible to improve models using more complex LSTMs
  - More layers!

Next Steps and Areas of Improvement

- Go beyond `distilroberta-finetuned-financial-news-sentiment-analysis` → use other text classification tools
- Models with more layers, or additional hyperparameter tuning
- Introduce more news headline scrapers!
- Utilize less-obvious, financial metrics, such as goodwill share price, intangible assets
- Inconsistency in model performance:
  - Collected data changes based on new articles that are written, and new daily closing prices → unreliable!
  - *In the future, establish definite dates and extract news/stock data between those dates*