Journal Club

FINSPID: A Comprehensive News Dataset in Time Series

Dr. Andrew Van Benschoten, 10/14/25

What is journal club?

- *Review and discuss a novel scientific paper*

- Presenter: 5-10 minute paper summary
  - Main hypothesis
  - Methods used
  - Key results & acknowledged limitations

- Everyone: discussion & critique!
  - Strengths? Weaknesses? Cool new ideas?

Why journal club?

1. Understand & track cutting-edge AI research
2. Develop the *scientific method* muscle (consume -> produce)
3. Inspire new lines of exploration
4. Develop community!

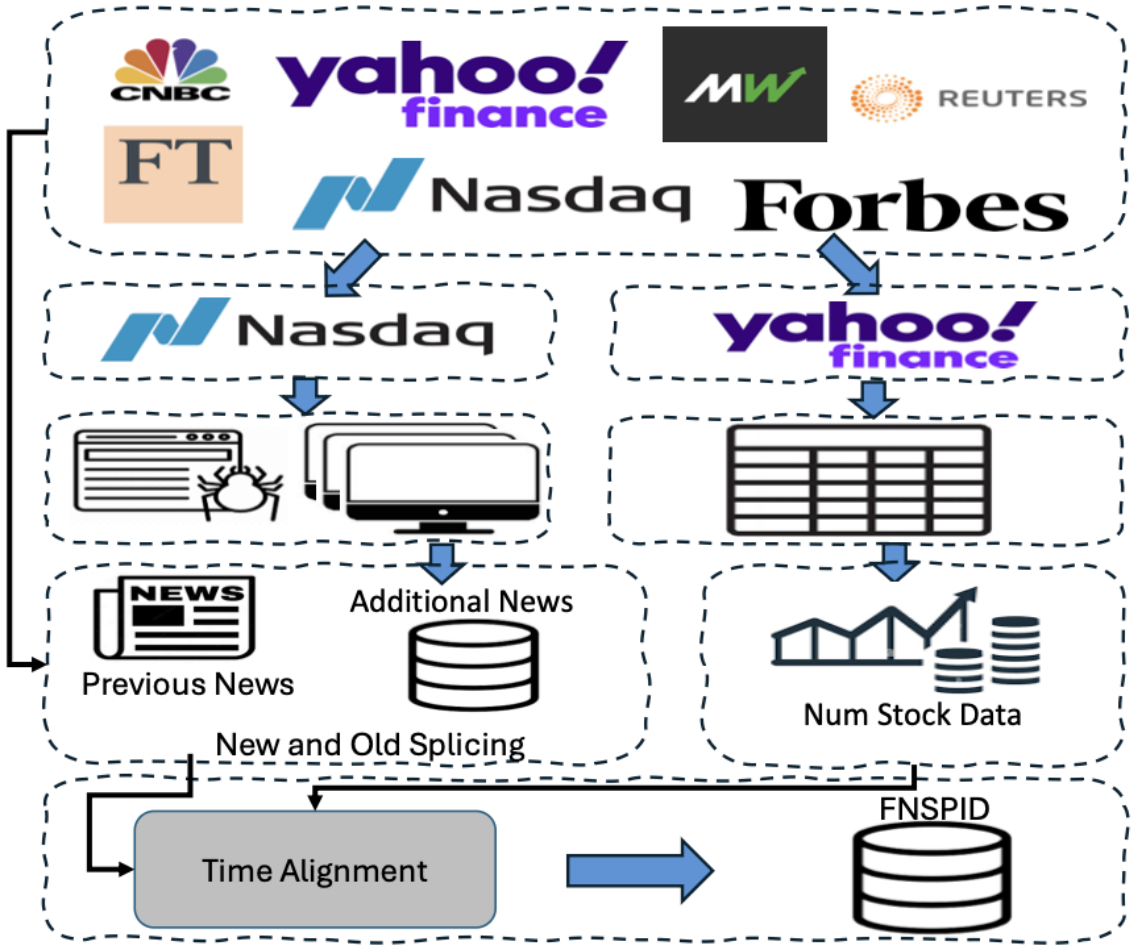
FNSPID: A Comprehensive Financial News Dataset in Time Series

Zihan Dong*
zdong7@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Xinyu Fan
SiChuan University
SiChuan, Sichuan, China
fanxinyu@stu.scu.edu.cn

Zhiyuan Peng
North Carolina State University
Raleigh, North Carolina, USA

- *Sentiment information boosts stock market forecasting ability*
- *FINSPID dataset size & quality exceeds current benchmarks & creates more accurate forecasts*



1. Collect stock data from Yahoo Finance & sentiment from NASDAQ/Bloomberg/Reuters/Lenta/Benzinga.

2. “Ethics” applied.

- 29.7 million stock prices and 15.7 million financial news records for 4,775 S&P500 companies from 1999 to 2023, gathered from four stock market news websites.

Date	Open	High	Low	Close	Adj close	Volume	Sentiment_gpt	News_flag	Scaled_sentiment
2023-03-09 00:00:00+00:00	93.68000030517578	96.20999908447266	92.18000030517578	92.25	92.25	56218700	3.3157894736842106	1.0	0.5789723684210526
2023-03-10 00:00:00+00:00	92.66999816894533	93.56999969482422	90.25	90.7300033569336	90.7300033569336	69827500	3.125	1.0	0.5312749999999999
2023-03-13 00:00:00+00:00	89.97000122070312	94.0199966430664	88.12000274658203	92.43000030517578	92.43000030517578	72397100	3.588235294117647	1.0	0.6470838235294119

Date string	Article_title string	Stock_symbol string	Url string	Publisher string	Author null	Article null	Lsa_summary null	Luhn_summary null	Textrank_summary null	Lexrank_sum null
2020-06-05…	Stocks That Hit 52-Week…	A	https://www.benzinga.com/news/20/06/16190091/stocks-that-hit-52-week-highs-on-friday	Benzinga Insights	null	null	null	null	null	
2020-06-03…	Stocks That Hit 52-Week…	A	https://www.benzinga.com/news/20/06/16170189/stocks-that-hit-52-week-highs-on-wednesday	Benzinga Insights	null	null	null	null	null	
2020-05-26…	71 Biggest Movers From…	A	https://www.benzinga.com/news/20/05/16103463/71-biggest-movers-from-friday	Lisa Levin	null	null	null	null	null	
2020-05-22…	46 Stocks Moving In…	A	https://www.benzinga.com/news/20/05/16095921/46-stocks-moving-in-fridays-mid-day-session	Lisa Levin	null	null	null	null	null	
2020-05-22…	B of A Securities…	A	https://www.benzinga.com/news/20/05/16095304/b-of-a-securities-maintains-neutral-on-agilent-…	Vick Meyer	null	null	null	null	null	
2020-05-22…	CFRA Maintains Hold on…	A	https://www.benzinga.com/news/20/05/16095163/cfra-maintains-hold-on-agilent-technologies-lowers-…	vishwanath@benzinga.com	null	null	null	null	null	

Date	2022-06-03 00:00:00
Symbol	AAPL
Headline	Consider Alphabet Stock Even in a Recession
Text	After six straight red weeks, the bulls may rejoice with two consecutive green days. This is where the fear of missing out kicks in for most investors and they blindly jump back in. Today, we will contemplate the prospects of doing so with Alphabet (NASDAQ:GOOG,NASDAQ:GOOGL) stock. But first, we should discuss the bigger...
URL	https://www.nasdaq.com/articles/consider-alphabet-stock-even-in-a-recession
LSA Sum	But investors will be shy about risking money if they think a big recession is coming.7 Overlooked Value Stocks to Buy Before Wall Street Catches On ...
Luhn Sum	Today, we will contemplate the prospects of doing so with Alphabet (NASDAQ:GOOG,NASDAQ:GOOGL) stock.Judging by their statements, they...
TexRank Sum	The reason why experts are now calling for disaster is the rhetoric from the Fed. Ticker Company Price GOOG Alphabet Inc. $2,202.40 GOOG Stock....
LexRank Sum	These are conditions that Wall Street deems as recessionary. Current investors of GOOG stock have realistic expectation..

Figure 3: Sentiment Data: Where 'Symbol' represents the stock code (e.g., AAPL for Apple Inc.); 'LSM Sum', 'Luhn Sum', 'TextRank Sum', and 'LexRank Sum' encapsulate the summarized news information generated by three different algorithms.

```
System: Forget all your previous instructions. You are a financial expert with
stock recommendation experience. Based on a specific stock, score for range from 1
to 5, where 1 is negative, 2 is somewhat negative, 3 is neutral, 4 is somewhat
positive, 5 is positive. 10 summarized news will be passed in each time, you will
give score in format as shown below in the response from assistant.

User: "News to Stock Symbol -- AAPL: Apple (AAPL) increase 22% ### News to Stock
Symbol -- AAPL: Apple (AAPL) price decreased 30% ### News to Stock Symbol -- MSFT:
Microsoft (MSTF) price has no change"
ASSISTANT: "5, 1, 3"
User: "News to Stock Symbol -- AAPL: Apple (AAPL) announced iPhone 15 ### News to
Stock Symbol -- AAPL: Apple (AAPL) will release VisonPro on Feb 2, 2024"
Assistant: "4, 4"

USER: ### News to Stock Symbol -- {symbol}: {text}
```

Figure 2: Example ChatGPT Prompt: The first section is the system prompt, defining constraints and specifying the task for ChatGPT. In the second section, two examples are included to guide ChatGPT on the desired content for the response. Subsequently, the summarized news is fed into ChatGPT for sentiment score labeling. *{symbol}* is the stock symbol variable input and *{text}* is the news variable input.

- "To meet the requirement, we incorporated a small dataset of news articles collected from 50 prominent US stocks from S&P 500 with sentiment labels (Task 3)"

- 10 data entries at a time, 1-5 sentiment score rating (more stable)

Date	Open	High	Low	Close	Adj.	Volume
2023-12-28 00:00:00	194.14	194.66	193.17	193.58	193.58	34014500
2023-12-27 00:00:00	192.49	193.50	191.09	193.15	193.15	48087700
2023-12-26 00:00:00	193.61	193.89	192.83	193.05	193.05	28919300
...	...	...	...	...	...	...

Table 2: Stock Numerical Data: 'Open' represents the opening stock price, 'High' indicates the highest price within the day, 'Low' signifies the lowest price within the day, 'Adj Close' represents the close price adjusted for dividends, and 'Volume' denotes the number of shares traded.

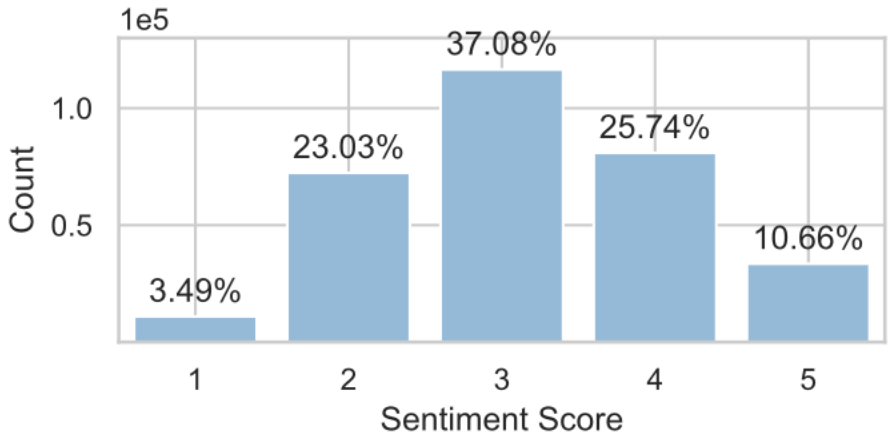


Figure 4: Sentiment Distribution: 1 is negative, 2 is somewhat negative, 3 is neutral, 4 is somewhat positive, 5 is positive

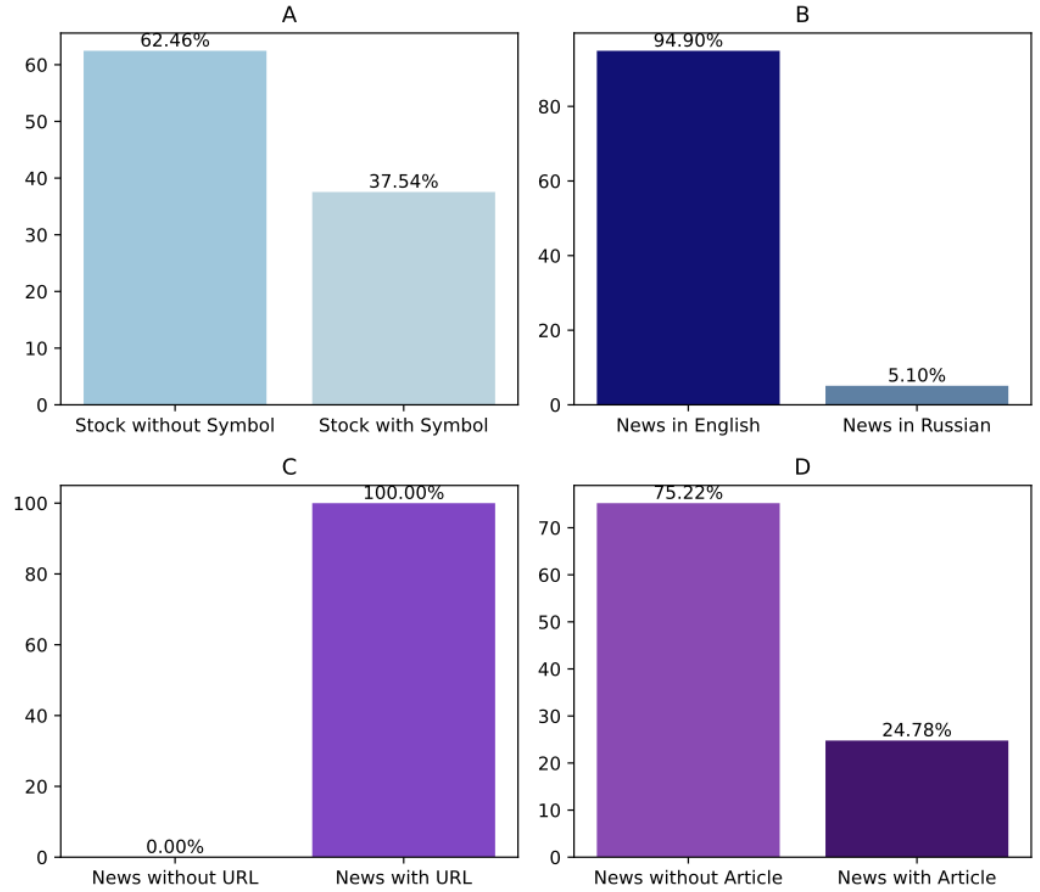
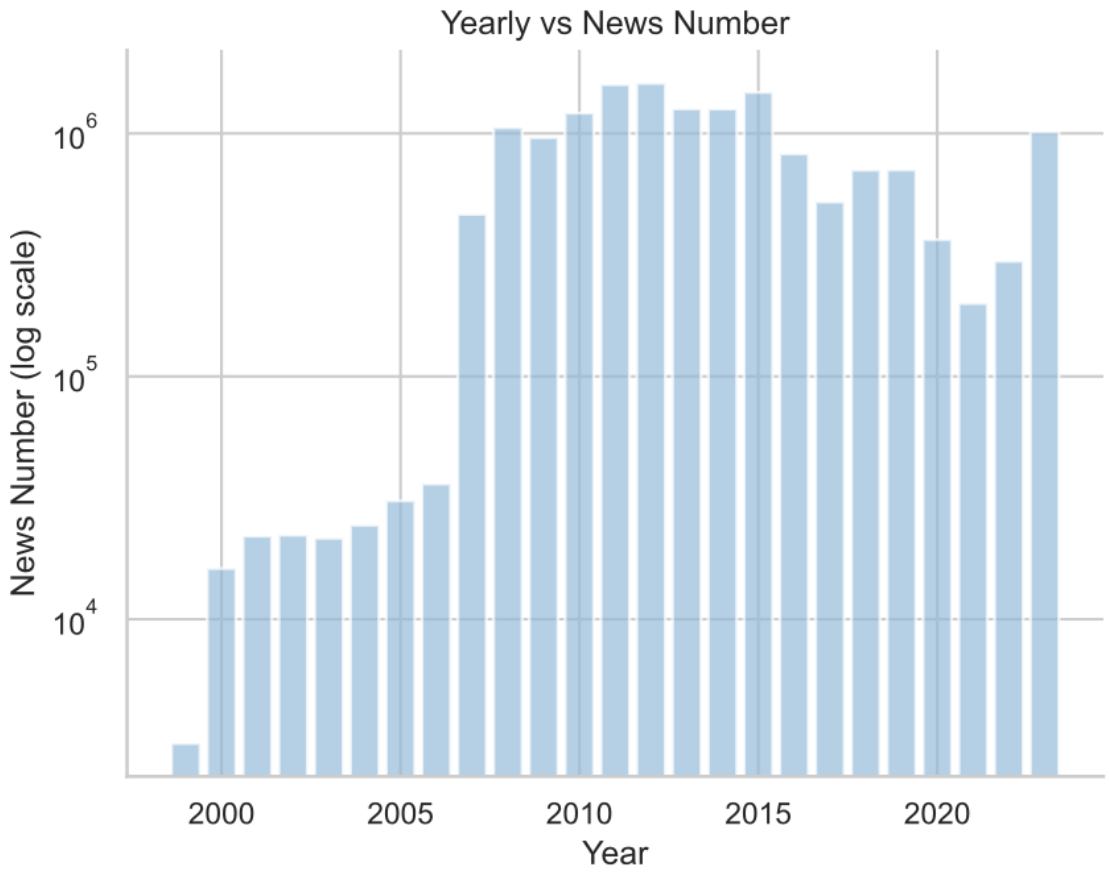


Figure 5: Statistical Overview: In A, we provide information on news articles that include the stock symbol. The B displays the language distribution, encompassing English and Russian. In C, a comparison of the included URLs is presented. Finally, in the D, details are provided on the news text already incorporated in the dataset, along with potential expansions into additional text data.

	FNSPID (ours)	Reuters	Benzinga	Bloomberg	Lenta	Lutz's	Farimani's	SemEval*	SEntFiN 1.0
Time Stamp	Yes	Yes	Yes	Yes	Yes	No	No	No	No
Text Type	Article	Article	Article	Article	Article	Sentence	Sentence	Headline	Headline
Number of News	15698563	8556324	3252885	447341	800974	1000	21867	1142	10753
Symbol	Yes	No	Yes	No	No	No	No	No	No
Summarization	Yes	No	No	No	No	No	Yes	No	No
Sentiment Score	Integer	-	-	-	-	Integer	-	Real	Integer
URL	Yes	No	Yes	No	No	No	No	No	No
Language	Many	Eng	Eng	Eng	Ru	Eng	Eng	Eng	Eng
Stock Price	Yes	No	No	No	No	No	Yes	No	No

Table 1: Comparison of existing datasets for Time Series Financial Analysis. FNSPID stands out with the highest volume of news data and includes unique features not found in other benchmark datasets. In the label, SemEval* stands for SemEval-2017 Task5 dataset.

Date	Article_title	Stock_symbo	Url	Publisher	Author	Article	Lsa_summary	Luhn_summary	Textrank_summary	Lexrank_summary
2023-12-16 2	Interesting A	A	https://www.nasdaq.com/articles/inter			Investors in	Because the $125.00 strik	The current analytical	Below is a chart showin	At Stock Options Channel,
2023-12-12 0	Wolfe Resear	A	https://www.nasdaq.com/articles/wolfe			Fintel	Fintel reports that on Dece	T. Rowe Price Investme	Agilent Technologies De	The projected annual reven
2023-12-12 0	Agilent Techr	A	https://www.nasdaq.com/articles/agile			In recent	In recent trading, shares of	In recent trading, share	When a stock reaches t	When a stock reaches the t
2023-12-07 0	Agilent (A) En	A	https://www.nasdaq.com/articles/agile			Agilent				



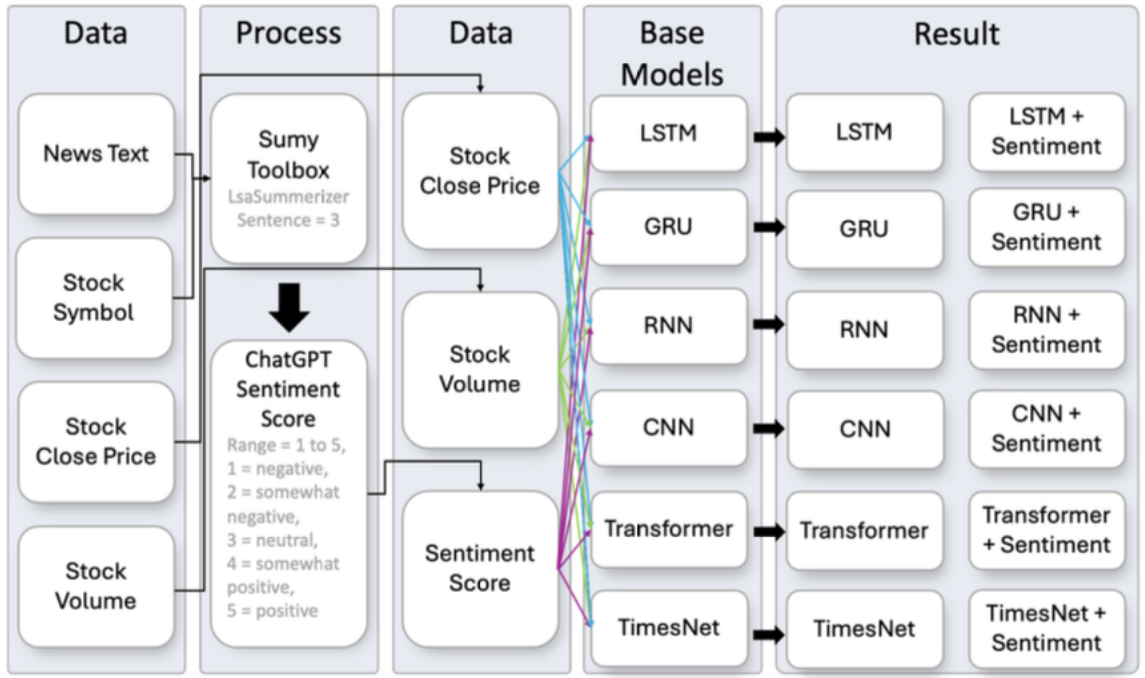


Figure 7: Experiment Procedure: The experimental setup involves utilizing news text, stock symbol, stock close price, stock open price, and stock volume as inputs to predict the stock's close price. The news text is processed through the Lsa-summarizer, followed by ChatGPT sentiment quantification. The obtained sentiment score, stock close price, open price, and volume are input into CNN, RNN, LSTM, GRUs, Transformer, and TimesNet. Concurrently, a reference group is established, incorporating only stock open price, close price and volume as input variables.

	Dataset	A-Sen.	A-Sen.	A-Sen.	A-Non.	A-Non.	A-Non.	B-Sen.	B-Sen.	B-Sen.	B-Non.	B-Non.	B-Non.
#	Name	MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE	R^2
5	LSTM	.02599	.00157	.87115	.02530	.00148	**.88016**	.02677	.00160	**.86811**	.02523	.00142	**.88181**
	CNN	.06180	.00712	.48205	.04913	.00475	.61811	.04236	.00354	.71668	.04522	.00398	.66687
	GRU	.02474	.00143	.88588	.02494	.00141	.88302	.02631	.00154	.86756	.02470	.00139	.87746
	RNN	.04152	.00355	.72957	.03353	.00251	.81128	.04315	.00339	.54265	.03898	.00291	.65470
	Transformer	**.01801**	**.00058**	**.87260**	**.01883**	**.00060**	.86659	**.01700**	**.00060**	.84659	**.01007**	**.00021**	.94629
	TimesNet	.02847	.00148	.63407	.02225	.00089	.81824	.03441	.00194	.51742	.02697	.00129	.69189
25	LSTM	.02569	.00155	.87040	.02482	.00141	.87627	.02569	.00146	.86889	.02706	.00178	.86401
	CNN	.04520	.00402	.69021	.04271	.00371	.71418	.04201	.00365	.71958	.04161	.00354	.72290
	GRU	.02696	.00178	.86873	.02484	.00145	.88233	.02848	.00192	.86129	.02523	.00142	.87175
	RNN	.03829	.00311	.73611	.03426	.00277	.76536	.03828	.00293	.68064	.03975	.00280	.58985
	Transformer	**.00757**	**.00008**	**.98304**	**.00711**	**.00008**	**.98178**	**.00943**	**.00013**	**.96811**	**.00763**	**.00009**	**.97948**
	TimesNet	.02347	.00093	.79670	.02364	.00093	.77555	.02412	.00104	.77040	.02319	.00091	.78261
50	LSTM	.02493	.00170	.85585	.02510	.00145	.87988	.02772	.00168	.83983	.02590	.00154	.86678
	CNN	.03550	.00289	.73355	.04126	.00343	.73344	.04092	.00346	.74825	.04129	.00343	.73457
	GRU	.02769	.00209	.82767	.02612	.00166	.87071	.02671	.00160	.85643	.02587	.00150	.86944
	RNN	.04154	.00389	.61744	.03343	.00243	.78635	.03849	.00317	.75238	.03658	.00289	.74494
	Transformer	**.00544**	**.00005**	**.98785**	**.00615**	**.00006**	**.98592**	**.00488**	**.00004**	**.99109**	**.00614**	**.00007**	**.98527**
	TimesNet	.02577	.00106	.73819	.02181	.00084	.80573	.02119	.00077	.82663	.02551	.00118	.72460

Table 3: Experiment Evaluation via 50 epochs of training, A-Sen. is ChatGPT labeled sentiment dataset result, B-Sen. is the TextBlob labeled sentiment dataset, A-Non., and B-Non. are the numerical data only dataset for experiments A and B. # is the number of stocks used in training for 5,25,50.

- In conclusion, we summarize 3 points from the experiment based on FNSPID:
- 1. Both the quality and quantity of the dataset largely affect the stock price prediction.
- 2. High-quality sentiment information has a positive effect on transformer-based training.
- 3. The transformer-based model surpasses traditional time series models and novel methods like TimesNet in stock price prediction.

FNSPID: A Comprehensive Financial News Dataset in Time Series

Zihan Dong*
zdong7@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Xinyu Fan
SiChuan University
SiChuan, Sichuan, China
fanxinyu@stu.scu.edu.cn

Zhiyuan Peng
North Carolina State University
Raleigh, North Carolina, USA

- *Sentiment information boosts stock market forecasting ability*
- *FINSPID dataset size & quality exceeds current benchmarks & creates more accurate forecasts*

AVB's thoughts

- Findings seem to contradict the importance of sentiment for forecasting?
- TextBlob out of left field! Don't agree with the general conclusion…
- Interesting that they only used/processed/forecasted with part of their dataset (not every stock has a GPT sentiment score)
- Weightscores (A.2) are rather obtuse…
- The exponential decay is a *big* assumption that isn't tested
- Research feels a bit like a grab-bag…many "connections" don't flow together and are due to pragmatism (Russian language?)