UCLAAnderson
SCHOOL of MANAGEMENT

MGMTMFE 431:

Data Analytics and Machine Learning

Topic 7b: Large Language Models

Spring 2025

Professor Lars A. Lochstoer

UCLAAnderson
SCHOOL of MANAGEMENT

Large Language Models

Up until now the analysis has essentially been of the 'bag-of-words' variety

- Tokenize unigrams, bigrams, etc
- Collection of individual words

That is, the models have not evaluated context in any serious sense, far from the way we as humans generate and understand text

The recent developments in *Large Language Models* seek to overcome this

UCLAAnderson
SCHOOL of MANAGEMENT

Two main models

Today: sentiment using (1) BERT and (2) ChatGPT

From https://powerbrainai.com/bert-vs-gpt/

- BERT and GPT are two significant models in natural language processing, each presenting unique advantages and challenges.
- BERT stands out with its bidirectional design, enabling a deeper understanding of context by considering both preceding and following words. It's particularly suitable for tasks related to context-aware understanding, like entity recognition or question answering.
- Key features of BERT include multilingual support, adaptability through fine-tuning on small datasets, and the ability to handle long-term dependency. However, BERT demands significant computing power and memory and is not suitable for text generation.
- GPT, conversely, excels in text generation tasks due to its left-to-right, unidirectional design. Despite being less effective in capturing dependencies compared to BERT, it shines in generating cohesive and human-like texts.
- Notable aspects of GPT include its scalability, ability to handle large datasets, and proficiency in preserving consistency across extended texts. Yet, training requires considerable computational resources and doesn't officially provide multilingual support.
- Both models can effectively tackle out-of-vocabulary words and excel in transfer learning. Their selection depends on each project's unique needs and applications, with BERT leading in context-aware tasks and GPT in text generation.

UCLAAnderson
SCHOOL of MANAGEMENT

FinBERT

https://huggingface.co/ProsusAI/finbert

https://github.com/ProsusAI/finBERT

- FinBERT is a pre-trained NLP model to analyze sentiment of financial text. It is built by further training the BERT language model in the finance domain, using a large financial corpus and thereby fine-tuning it for financial sentiment classification. Financial PhraseBank by Malo et al. (2014) is used for fine-tuning. For more details, please see the paper FinBERT: Financial Sentiment Analysis with Pre-trained Language Models and our related blog post on Medium.

- The model will give softmax outputs for three labels: positive, negative or neutral.

UCLAAnderson
SCHOOL of MANAGEMENT

FinBERT - Tone

https://huggingface.co/yiyanghkust/finbert-tone

FinBERT is a BERT model pre-trained on financial communication text. The purpose is to enhance financial NLP research and practice. It is trained on the following three financial communication corpus. The total corpora size is 4.9B tokens.

- Corporate Reports 10-K & 10-Q: 2.5B tokens
- Earnings Call Transcripts: 1.3B tokens
- Analyst Reports: 1.1B tokens

More technical details on FinBERT: Click Link

This released finbert-tone model is the FinBERT model fine-tuned on 10,000 manually annotated (positive, negative, neutral) sentences from analyst reports. This model achieves superior performance on financial tone analysis task. If you are simply interested in using FinBERT for financial tone analysis, give it a try.

If you use the model in your academic work, please cite the following paper:

Huang, Allen H., Hui Wang, and Yi Yang. "FinBERT: A Large Language Model for Extracting Information from Financial Text." Contemporary Accounting Research (2022).

UCLAAnderson
SCHOOL of MANAGEMENT

Negative, Neutral and Positive

```
from transformers import BertTokenizer, BertForSequenceClassification
import numpy as np

finbert = BertForSequenceClassification.from_pretrained('yiyanghkust/finbert-tone',num_labels=3)
tokenizer = BertTokenizer.from_pretrained('yiyanghkust/finbert-tone')

sentences = ["there is a shortage of capital, and we need extra financing",
             "growth is strong and we have plenty of liquidity",
             "there are doubts about our finances",
             "profits are flat"]

inputs = tokenizer(sentences, return_tensors="pt", padding=True)
outputs = finbert(**inputs)[0]

labels = {0:'neutral', 1:'positive',2:'negative'}
for idx, sent in enumerate(sentences):
    print(sent, '----', labels[np.argmax(outputs.detach().numpy()[idx])])

'''
there is a shortage of capital, and we need extra financing ---- negative
growth is strong and we have plenty of liquidity ---- positive
there are doubts about our finances ---- negative
profits are flat ---- neutral
'''
```

UCLAAnderson
SCHOOL of MANAGEMENT

FOMC minutes sentiment

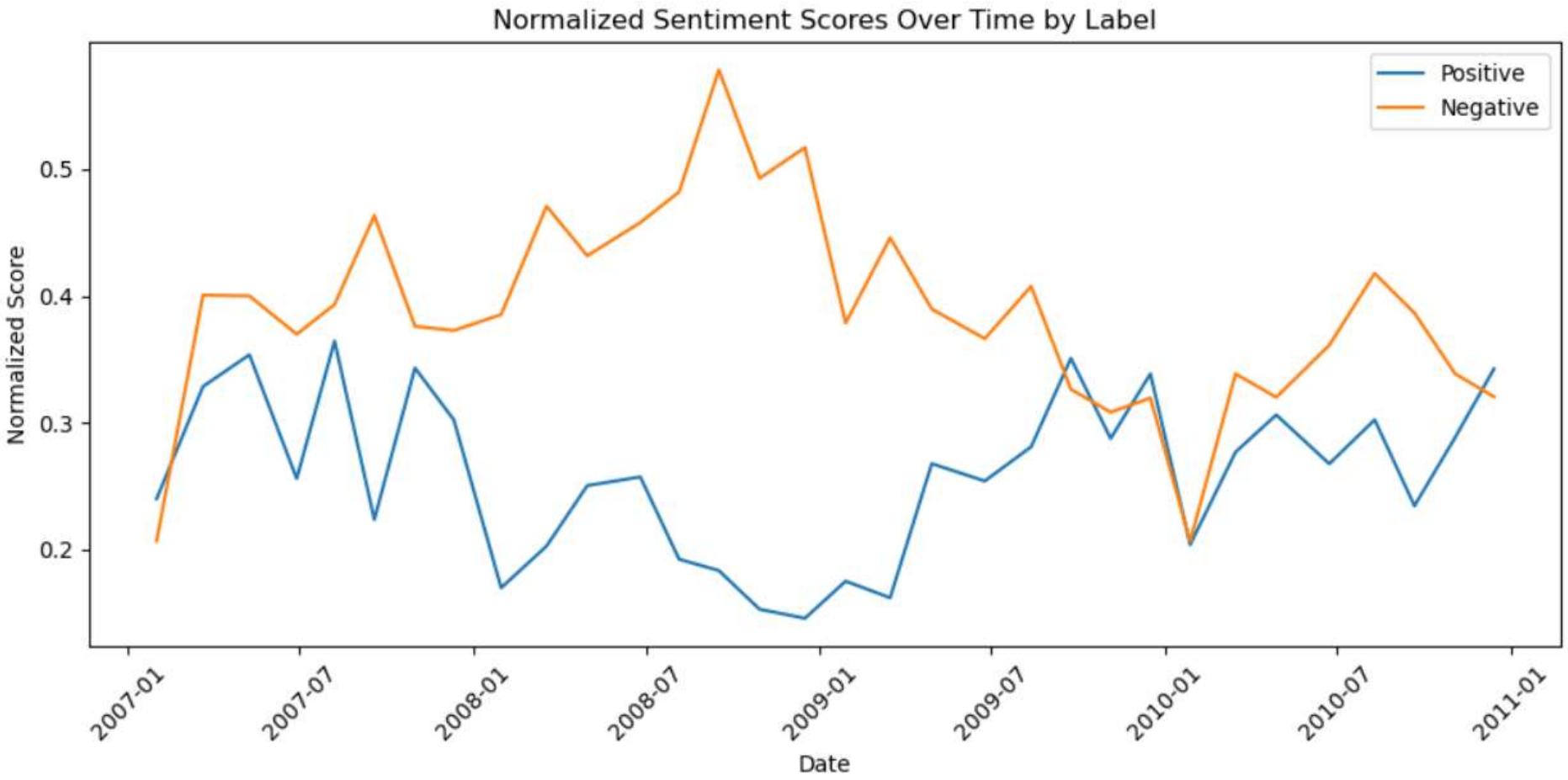
```
fomc_df
```

	DATE	LABEL	COUNT	SCORE	Count_Score	Total_Score_Per_Date	Score_Share
21	1/27/2010	Neutral	190	0.967935	183.907578	311.882023	0.589670
22	1/27/2010	Negative	71	0.907085	64.403026	311.882023	0.206498
23	1/27/2010	Positive	66	0.963203	63.571418	311.882023	0.203832
53	1/28/2009	Positive	55	0.947593	52.117632	297.820309	0.174997
51	1/28/2009	Neutral	136	0.977502	132.940321	297.820309	0.446378
...	...	...	...	...	...	...	...
38	9/21/2010	Negative	78	0.952751	74.314568	192.202896	0.386646
37	9/21/2010	Positive	49	0.919603	45.060569	192.202896	0.234443
80	9/23/2009	Negative	80	0.967107	77.368548	237.041923	0.326392
79	9/23/2009	Positive	86	0.966961	83.158613	237.041923	0.350818
78	9/23/2009	Neutral	79	0.968541	76.514761	237.041923	0.322790

96 rows × 7 columns

UCLAAnderson
SCHOOL of MANAGEMENT

Time series over financial crisis



UCLAAnderson
SCHOOL of MANAGEMENT

Relation to S&P500 returns

How does the sentiment relate to returns the 5 days around the release?

- Positive sentiment is related to negative return. Does this make sense?
- It might, as positive sentiment increases chance of a rate increase

OLS Regression Results

Dep. Variable:	Cumulative_RET	R-squared:	0.186
Model:	OLS	Adj. R-squared:	0.130
Method:	Least Squares	F-statistic:	3.306
Date:	Wed, 01 May 2024	Prob (F-statistic):	0.0509
Time:	13:14:07	Log-Likelihood:	67.529
No. Observations:	32	AIC:	-129.1
Df Residuals:	29	BIC:	-124.7
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Positive	-0.2293	0.095	-2.426	0.022	-0.423	-0.036
Negative	-0.0200	0.077	-0.259	0.798	-0.178	0.138
intercept	0.0744	0.046	1.600	0.120	-0.021	0.169

Omnibus:	3.917	Durbin-Watson:	2.119
Prob(Omnibus):	0.141	Jarque-Bera (JB):	2.561
Skew:	0.650	Prob(JB):	0.278
Kurtosis:	3.481	Cond. No.	23.1

UCLAAnderson
SCHOOL of MANAGEMENT

Recent paper using Chat GPT

Return Predictability and Large Language Models *

Alejandro Lopez-Lira and Yuehua Tang

University of Florida

First Version: April 6, 2023; This Version November 8, 2023

Abstract

We examine the potential of ChatGPT and other large language models (LLMs) to predict stock market returns using news. Categorizing headlines with ChatGPT as positive, negative, or neutral for companies' stock prices, we document a significant correlation between ChatGPT scores and subsequent daily stock returns, outperforming traditional methods. Basic models like GPT-1 and BERT cannot accurately forecast returns, indicating return forecasting is an emerging capacity of more complex LLMs, which deliver higher Sharpe ratios. We explain these puzzling return predictability patterns by testing implications from economic theories involving information diffusion frictions, limits to arbitrage, and investor sophistication. Predictability strengthens among smaller stocks and following negative news, consistent with these theories. Only advanced LLMs maintain accuracy when interpreting complex news and press releases. Finally, we present an interpretability technique to evaluate LLMs' reasoning. Overall, incorporating advanced language models into investment decisions can improve prediction accuracy and trading performance.

UCLAAnderson
SCHOOL of MANAGEMENT

Recent paper using Chat GPT

ChatGPT training data ends in September 2021

- Authors consider sample of daily market returns from October 2021 through December 2022 so analysis is “out-of-sample”

Scrape news headlines from wide variety of sources

- Major news agencies, financial news websites, social media platforms
- Search for headlines containing company names or tickers
- 67k+ headlines for 4k+ companies
- Match with RavenPack data to ensure headlines are matched with stock returns in a correct manner

UCLAAnderson
SCHOOL of MANAGEMENT

Prompt engineering

This is the prompt they ask ChatGPT:

"Forget all your previous instructions. Pretend you are a financial expert. You are a financial expert with stock recommendation experience. Answer "YES" if good news, "NO" if bad news, or "UNKNOWN" if uncertain in the first line. Then elaborate with one short and concise sentence on the next line. Is this headline good or bad for the stock price of _*company_name*_ in the _*term*_ term?

Headline: _*headline*_"

UCLAAnderson
SCHOOL of MANAGEMENT

Prompt engineering

Consider headline about Oracle: Rimini Street Fined $630,000 in Case Against Oracle.

The prompt then asks:

> Forget all your previous instructions. Pretend you are a financial expert. You are a financial expert with stock recommendation experience. Answer "YES" if good news, "NO" if bad news, or "UNKNOWN" if uncertain in the first line. Then elaborate with one short and concise sentence on the next line. Is this headline good or bad for the stock price of Oracle in the short term?
> Headline: Rimini Street Fined $630,000 in Case Against Oracle

And here is ChatGPT's response:

> YES
>
> The fine against Rimini Street could potentially boost investor confidence in Oracle's ability to protect its intellectual property and increase demand for its products and services.

UCLAAnderson
SCHOOL of MANAGEMENT

Empirical Strategy

We match the headlines to the next trading period. For headlines before 6 a.m. on a trading day, we assume the headlines can be traded by the market opening of the same day and sold at the close of the same day. For headlines after 6 a.m. but before 4 p.m., we assume the headlines can be traded at the same day's close and sold at the close of the next trading day. For headlines after 4 p.m., we assume the headlines can be traded at the opening price of the next day and sold at the closing price of that next day. We then run linear regressions of the next day's stock returns on the ChatGPT score, the sentiment score provided by the data vendor, and scores from other LLMs. Thus, all of our results are out-of-sample.

Specifically, we estimate the following regression specification:

$$r_{i,t+1} = a_i + b_t + \gamma' x_{i,t} + \varepsilon_{i,t+1}, \quad (1)$$

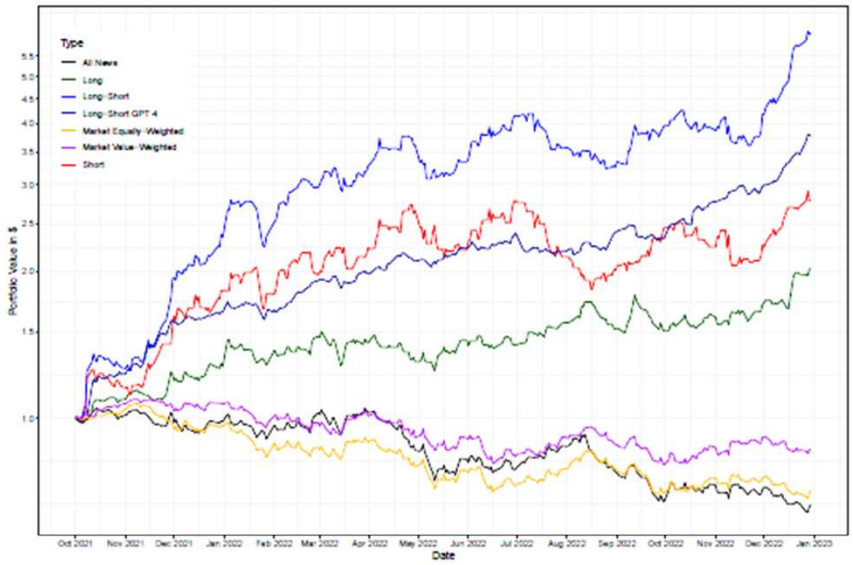
where the dependent variable, $r_{i,t+1}$, is stock i's return over a subsequent trading day as discussed above, $x_{i,t}$ refers to the vector containing the ChatGPT score or other scores from assessing stock i's news headlines, and a_i and b_t are firm and date fixed effects, respectively, which account for any observable and unobservable time-invariant firm characteristics and common time-specific factors that could influence stock returns. Standard errors are double clustered by date and firm.

I don't like the firm fixed effects! Why not run Fama-MacBeth regressions?

UCLAAnderson
SCHOOL of MANAGEMENT

Empirical Strategy: trading strategies

Figure 1: Cumulative Returns of Investing $1 (Without Transaction Costs)



This figure presents the results of different trading strategies based on ChatGPT 3.5 and ChatGPT 4 without considering transaction costs. If a piece of news is released before 6 a.m. on a trading day, we enter the position at the market opening and exit at the close of the same day. If the news is released after 6 a.m. but before the market close, we enter the position at the market close price of the same day and exit at the close of the next trading day. If the news is announced after the market closes, we assume we enter the position at the next opening price and exit at the close of the next trading day. All the strategies are rebalanced daily. The "All-news" black line corresponds to an equal-weight portfolio in all companies with news the day before (regardless of news direction). The green line corresponds to an equal-weighted portfolio that buys companies with good news, according to ChatGPT 3.5. The red line corresponds to an equal-weighted portfolio that short-sells companies with bad news, according to ChatGPT 3.5. The light blue line corresponds to an equal-weighted zero-cost portfolio that buys companies with good news and short-sells companies with bad news, according to ChatGPT 3.5. The dark blue line corresponds to an equal-weighted zero-cost portfolio that buys companies with good news and short-sells companies with bad news, according to ChatGPT 4. The yellow line corresponds to an equally weighted market portfolio. The purple line corresponds to a value-weighted market portfolio.

UCLAAnderson
SCHOOL of MANAGEMENT

Empirical Strategy: trading strategies

Table 3: Regression of Next Day Returns on Prediction Scores from More Advanced LLMs

This table reports the results of running regressions of the form $r_{i,t+1} = a_i + b_t + \gamma' x_{i,t} + \varepsilon_{i,t+1}$. Where $r_{i,t+1}$ is the next day's return in percentage points, a_i and b_t are firm and time fixed effects, respectively. $x_{i,t}$ corresponds to the vector containing prediction scores from different models. The main regressors include scores from three advanced LLMs: (i) ChatGPT 3.5, (ii) ChatGPT 4, and (iii) BART Large. We include the event sentiment score from the data vendor for comparison purposes. We provide an overview of the different LLMs in Appendix A of the paper. The corresponding t-statistics are in parentheses. Standard errors are double clustered by date and firm. All models include firm and time fixed effects. The sample consists of all U.S. common stocks with at least one news headline covering the firm.

	(1)	(2)	(3)	(4)	(5)	(6)
GPT-3.5-score	0.259***	0.243***				
	(5.259)	(4.980)				
event-sentiment-score		0.058		0.038	0.118*	
		(1.122)		(0.683)	(2.272)	
GPT-4-score			0.176***	0.167***		
			(5.382)	(4.768)		
bart-large-score						0.142***
						(4.653)
Num.Obs.	60 755	60 755	60 755	60 755	60 755	60 176
R2	0.184	0.184	0.184	0.184	0.184	0.185
R2 Adj.	0.121	0.121	0.121	0.121	0.121	0.121
R2 Within	0.001	0.001	0.001	0.001	0.000	0.000
R2 Within Adj.	0.001	0.001	0.001	0.001	0.000	0.000
AIC	370 534.7	370 534.9	370 534.8	370 536.1	370 560.5	367 175.7
BIC	409 811.3	409 820.5	409 811.4	409 821.7	409 837.2	406 374.6
RMSE	4.75	4.75	4.75	4.75	4.75	4.76
Std.Errors	by: date & permno	by: date & permno	by: date & permno	by: date & permno	by: date & permno	by: date & permno
FE: date	X	X	X	X	X	X
FE: permno	X	X	X	X	X	X

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

UCLAAnderson
SCHOOL of MANAGEMENT

Trading strategy Sharpe Ratios

Table 12: Sharpe Ratios by News Type

This table reports the annualized Sharpe ratio of the long-short portfolio implied by different models. The first column reports results using all news. The second column reports results using only news articles, and the third column reports results using only press releases.

Model	All	News Articles	Press Releases
Gpt-4	3.8	4.74	2.84
Gpt-3.5	3.09	2.4	1.45
Distilbart-Mnli-12-1	1.9	4.8	0.73
Event-Sentiment	1.74	0.5	0.69
Bart-Large	1.26	3.74	1.37
Gpt-2-Large	0.82	0.83	1.34
Finbert	0.77	2.85	<0
Gpt-1	0.62	0.1	0.06
Bert-Large	0.2	0.82	0.02
Bert	<0	<0	<0
Gpt-2	<0	0.93	<0

To good to be true? Or, perhaps this will be competed away?

UCLAAnderson
SCHOOL of MANAGEMENT

Epilogue

Exciting new tools are being developed to analyze text data

Possibly very useful for investment strategies

Lots more to learn, this is only to whet your appetite!