SentimentPulse: Real-Time Consumer Sentiment Analysis in Custom Language Models

Anonymous submission

Abstract

Large Language Models are trained on an extremely large corpus of text data to allow better generalization but this blessing can also become a curse and significantly limit their performance in a subset of tasks. In this work, we argue that LLMs are notably behind well-tailored and specifically designed models where the temporal aspect is important in making decisions and the answer depends on the timespan of available training data. We prove our point by comparing two major architectures: first, SentimentPulse, a real-time consumer sentiment analysis approach that leverages custom language models and continual learning techniques, and second, GPT-3.5-Turbo and GPT-4 (GPTs) which are both tested on the same data. Unlike foundation models, which lack temporal context, our custom language model is pre-trained on time-stamped data, making it uniquely suited for real-time application. Additionally, we employ continual learning techniques to pre-train the proposed model, and then use classification and contextual multi-arm bandits to fine-tune, enhancing its adaptability and performance over time. We present a comparative analysis of the predictions accuracy of both proposed architecture and GPTs models. To the best of our knowledge, this is the first application of custom language models for real-time consumer sentiment analysis beyond the scope of conventional surveys.

1 Introduction

Consumer sentiment is an important economic indicator because it shows how people feel about the health of an economy and affects both market trends and policy decisions. Conventional surveys are employed to measure public sentiments regarding the state of the economy. Surveys can yield valuable insights, which serve as data points in the decision-making process, yet conducting such surveys demands considerable time and financial resources. Further, these surveys are usually done at a specific frequency, so they can't show how people feel in real time.

We study the problem of consumer sentiment analysis with the help of a language model and continual learning. We conjecture that using a language model to capture consumer sentiment can be a viable and efficient compliment of existing surveys. As far as we know, this is the first time the consumer sentiment problem has been addressed in this way. We consciously refrain from employing the foundation models in the proposed model framework because the problem requires the model to be trained on data that includes specific time stamps. Foundation models are trained on internet corpora without time stamp information. To evaluate our proposed approach and its comparison to the foundation model, in this task, we set up extensive experiments for the proposed model and GPT-3.5-Turbo & GPT-4 (OpenAI, Accessed 2023) and compare the performance of all three models.

The paper presents three main contributions:

1. We propose a comprehensive consumer sentiment analysis framework that leverages news and S&P500 dataset (SP Dow Jones Indices LLC, Accessed 2023). Our framework does not only capture the consumer sentiment dynamics over time but can also provide feedback in a more timely manner, and it can be supplementary to traditional survey-based methods.

2. Our encoder-based model from scratch was pre-trained with a small dataset and showed good accuracy with a relatively small model size at a low cost. We use continual learning in our experiments and compare the results with GPT-3.5-Turbo and GPT-4. Our experiment results show that we can out-perform GPT-3.5-Turbo and GPT-4 on this task.

3. To the best of our knowledge, our framework is the first implementation to adapt the language model into economic consumer sentimental analysis. Our work establishes a baseline for future research.

This paper is organized as follows. In Section 2, we discuss the related work. Section 3 introduces the proposed model. The datasets for pre-training and fine-tuning are described in 4. Finally, Section 5 outlines the experimental setup and presents the results. Section 6 is the conclusion.

2 Related Work

The nature of the problem in the paper is related to different domains such as consumer sentiment, multiple choice question answering and GPTs models. In this section, we introduce some of the most recent related works in different domains.

2.1 Consumer Sentiment

(Kaur and Sharma, 2023) proposed a model for consumer sentiment analysis that uses long short-term memory (LSTM) and a hybrid feature extraction approach. The model proposes a technique for hybrid sentiment analysis aimed at improving the consumer review summarization. This method converts pre-processed reviews into feature vectors for sentiment analysis. Businesses can utilize the reviews to track product/service-specific reviews and analyze their competitors.

(Han et al., 2023) explored the relationship between consumer confidence index and web search keywords. The paper uses various machine learning models to predict the consumer confidence index with consumer confidence index data from China. The paper claims that the use of machine learning models has a better prediction on the consumer confidence index.

2.2 Multiple Choice Question Answering Methodology

The state-of-the-art (SOTA) multiple choice question answering (MCQA) models are mostly transformer encoder/decoder based. Following are some of works that achieve SOTA results.

(Huang et al., 2022) uses transformer encoder-decoder architecture to generate a clue text input to an encoder-based MCQA block to enhance the performance.

(Chaturvedi et al., 2018) uses CNN to capture the "question, answer option candidate" tuples embedding and sentence embedding, and then use attention layer to attend on both embeddings to obtain scores for each answer option candidate.

(Jin et al., 2019) divide the task of MCQA into two stage (coarse tuning and multi-task fine tuning), in which coarse tuning stage uses Natural Language Inference(NLI) to enhance the model's entailment capability and multi-task fine tuning stage uses both in-domain source and the target dataset together to fine tune the model.

(Chen et al., 2019) uses Bi-LSTM and attention layer to capture an enriched representation of questions, answer option candidates and paragraph. And then a convolutional spatial attention is used to capture the final score of each answer option candidate.

(Huang et al., 2021) focuses on Scenario-based Question Answering (SQA) and proposes a Retriever-Reader framework, in which the retriever is trained on QA labels using a novel word weighting mechanism to output a sparse representation of the paragraphs S_{spa} and top-K paragraphs, and reader takes the top-K paragraphs as input to calculate dense representation S_{den} of each paragraph and a fused representation of paragraphs and answer option candidate S_{fus}. The final score of each answer option candidate comes from normalized and weighted sum of S_{spa}, S_{den} and S_{fus}.

(Robinson et al., 2022) experiments using LLM to do MCQA on 20 diverse dataset and claims that LLM that is not sensitive to answer options' order can largely close the gap of other SOTA MCQA models when prompted with multiple choice prompting instead of cloze prompting.

2.3 Prompt Engineering for Economy Studies

Proper generation of the survey results with LLMs requires carefully engineered prompts. Prompt engineering has been studied in a very rigorous manner by the scientific community. However, only a few works are related specifically to the economy studies. (Horton, 2023) provides an overview of the ability of LLMs to serve as a tool to pilot economic studies and for decision making in economy-related questions via simulations. The experiments were performed using GPT-3 and the overall conclusion was that considering the cost-effectiveness of LLMs compared to real humans, it is a promising research direction. (Zhou et al., 2023) create characters, social scenarios and goals in LLMs via prompts and discover that artificial

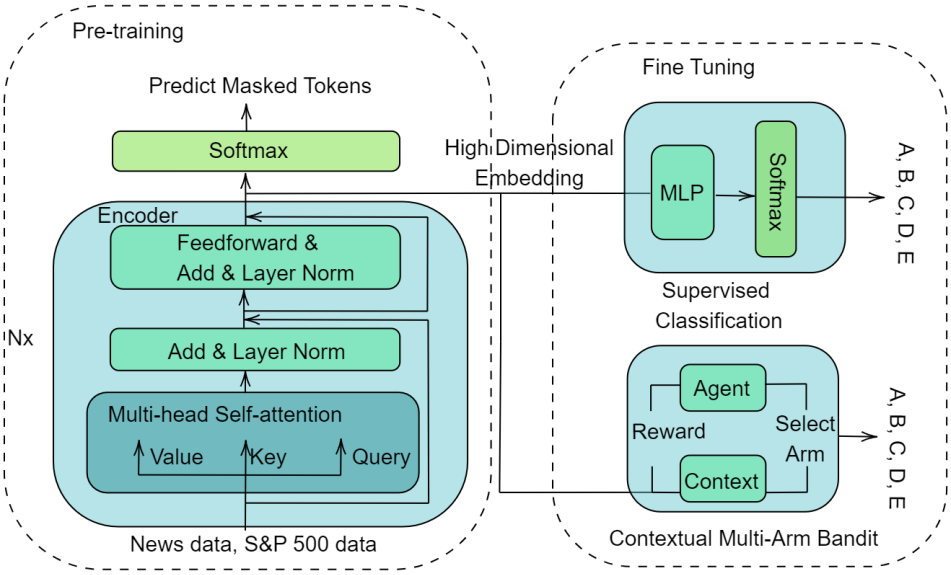


Figure 1: SentimentPulse: Two stages of training (Pre-training with Encoder; Fine-tuning with Supervised Classification and Contextual Multi-arm Bandit

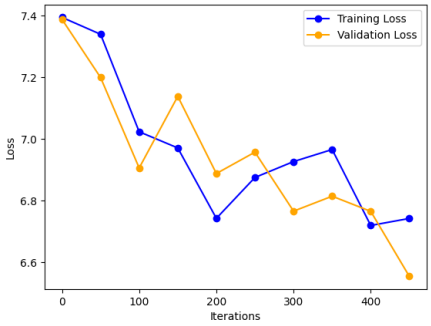
agents provided by LLMs do face difficulties in mimicking real-world social behaviors, and do so more under certain constraints. However, in certain scenarios, the authors find that agents created with GPT-4 achieve comparable results with human agents.

3 Model Framework

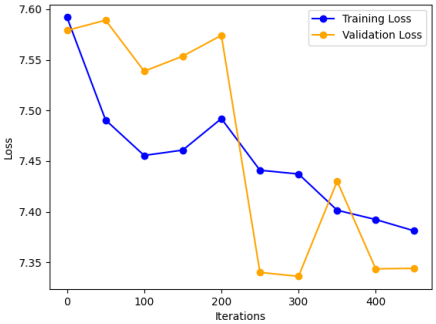
The proposed model framework is illustrated in Figure 1. It consists of two parts, namely, the Pre-training part and Fine-tuning part. To predict consumer sentiment, we treat it like a multiple-choice question-answering problem. This allows the proposed model to provide the closest answer based on the survey takers' information. We use a transformer encoder to unsupervised pre-train on news corpus and S&P 500 data. In fine-tuning, we use two strategies (supervised classification and contextual multi-arm bandit) to fine-tune the survey data independently.

3.1 Encoder

The encoder-decoder transformer architecture for machine translation tasks was first introduced in (Vaswani et al., 2017). Since then, various language models have been created using either encoder-only or decoder-only architecture for different language tasks. Multiple choice question answering is a long-standing research problem, and many of SOTA works are based on encoder-only architecture and



(a) Number of Parameter:732 million
Attention block dimension:160
Max input token allowed:150; Batch size: 16



(b) Number of Parameter:369 million
Attention block dimension:80
Max input token allowed:150; Batch size: 16

Figure 2: Cross entropy loss vs Number of iterations between the training set and validation set with two different settings of parameters of encoder

encoder-based architecture has become a popular paradigm for MCQA problem (Huang et al., 2022). In this work, we also use encoder-only architecture to build the proposed framework.

The left hand side of Figure 1 shows the encoder architecture. It is a standard transformer encoder that includes a multi-head self-attention layer, a normalization, a feedforward (with skip connection (He et al., 2015)), and a final softmax layer. During pre-training, we randomly mask tokens from a sentence in the news corpus, and then final softmax layer predicts the masked token of the sentence. During the fine-tuning stage, the encoder will generate high dimensional embedding using survey takers' profile text information as input. And then the high dimensional embedding will be fed into supervised classifiers and multi-arm bandit agents for fine-tuning independently.

3.2 Supervised Classification

In the fine-tuning phase, demonstrated on the right side of Figure 1, we employ a Multilayer Perceptron (MLP) and a softmax layer for the final prediction. During fine-tuning, the encoder will generate high dimensional embedding for each survey taker (how the encoder generates the embedding is discussed in details in section 5.2). The high dimensional embedding will be passed into MLP and subsequent softmax layer for supervised fine-tuning. Because we have information on which answer option each survey participant selected for specific years and months, these data will become our fine-tuning label (on the right hand side of Figure 1, it shows that there are "A, "B", "C", "D", "E" five labels, but the actual survey dataset contains questions with different number of labels/answer options). And it should be noted that, during supervised classification fine-tuning, the gradients are also backpropagated to the encoder to update its weights.

3.3 Contextual Multi-Arm Bandit

We have also taken a second approach to the classification task by using Contextual Multi-Arm Bandit, which is also depicted on the right-hand side of Figure 1. As shown in Figure 1, an agent will select an arm(one answer option) given a context information and a reward associated with the arm will be awarded to the agent. The context information is coming from the encoder (the same high dimensional embedding for supervised classification). During fine-tuning, every time the agent pick the arm that matches the reward table, the reward will plus 1; otherwise there is no reward. The training algorithm will try to maximize the total reward and minimize the regret. We experimented different training algorithms including Upper Confidence Bound , Epsilon Greedy and Adaptive Greedy. The detail experiment results and discussion are in section 5.2.

4 Dataset

4.1 News Corpus and S&P500 Data

For pre-training encoder, we use news corpus from New York Times News API (New York Times Developer Network, Accessed 2023), Guardian News API (The Guardian News, Accessed 2023), and S&P 500 data. We do not build proposed model framework on top of the existing pre-trained encoder because it lacks time stamp information. Our goal is to capture the economic sentiment from the news corpus and S&P 500 data, so we extract news based on various categories. We extract the news from the New York Times News API by categories such as "Politics," "Economy," "Entrepreneurship," "International Business," "Automobiles," and "Business Day." Likewise, we filter the news from the Guardian News API by categories such as "Money," "Politics," "Business," and "Society" within "USA-News"(both Guardian news and New York Times news archive their news based on categories).

After filtering out the news corpus by different categories, we divide them by different time stamps. Following Table 3 shows a snippet of the news corpus (extracted from Guardian news with time stamp of 2014 January) that is used to pre-train the encoder.

4.2 UMCSI Survey Data

We use survey data from the University of Michigan Consumer Sentiment Index (UMCSI) (University of Michigan, Accessed 2023) for fine-tuning. Since 1978, UMCSI has been monitoring consumer sentiment, making it one of the most closely followed economic indicators in the United States. It releases monthly consumer sentiment index reports. According to the University of Michigan, the survey accurately predicts the country's future economic path.

The questions posed to every survey taker are shown in Table 4 (in the Appendix). There are five questions in the survey, which aim to gather

consumers' opinions on different aspects of the economy, such as personal finances, business conditions, and buying conditions. Each question has several answer options, and survey takers choose the one that best reflects their attitude toward the current or expected changes in the economy. As shown in Table 4, Q1 and Q2 poses question about comparing family financial condition to one year ago and one year in the future respectively; Q3, Q4 poses question about comparing business economic condition to one year and five years in the future respectively; Q5 poses question on household purchase power. For details of the questions, please refer to Table 4 in the Appendix.

Additionally, participants need to provide their personal information, such as income, residence region, political affiliation, education level, number of adults & children in the household, birth year, and home ownership status.

5 Experiments

We conducted both pre-training and fine-tuning experiments on dual-GPU setups, each with 24GB of memory. Various model sizes were explored for encoder pre-training. All experiments were completed within a 12-hours window on this hardware configuration.

5.1 Unsupervised Pre-training of Encoder

The pre-training accuracy plots of two encoders (with different model parameters) are shown in Figure 2. During pre-training, the news corpus was divided by monthly time stamp, and the encoder was trained continuously using corpus with different time stamps. For every 12 months of news corpus, we trained the model for 500 iterations (each iteration is training on 10 batches) before moving on to the next 12 months' news corpus and repeating the process. Figure 2 shows the training and validation accuracy with 500 iterations using one 12-months of news corpus. We chose 500 iterations to avoid overfitting because it can occur with too many iterations. As shown in Figure 2(a) and Figure 2(b), both the training and validation loss decrease steadily without overfitting. The larger model size of 739 million parameters (compared to 369 million parameters) allowed for faster convergence of the pre-training loss, as seen in Figure 2(a) and Figure 2(b).

Continual Learning We specifically divide the corpus every 12 months to avoid overfitting during pre-training. We have also experimented with pre-training the encoder using 60 months' news corpus all together(12 months x 5 years), and the encoder overfits after a small number of iterations. And if the encoder is trained on 12 months of news corpus 5 times continually, the encoder's loss steadily decreases. This is because when the encoder is trained on a larger text corpus, the encoder is tuned toward a specific narrow distribution of the corpus data whereas dividing the corpus into five and training on them continually and individually can make the model generalize much better.

The encoder undergoes continual pre-training on 12-months of news corpus continually. The training procedure is illustrated in Algorithm 1. The encoder is pre-trained in line 2 and then connected to a MLP to create "model1" (line 3) for future fine-tuning. A contextual bandit instance is also initiated in line 4 to create a reward table and action table for each training algorithm (Upper Confidence Bound (UCB), Adaptive Greedy (AG), Epsilon Greedy (EG)) of the multi-arm bandit problem. In line 6, a high dimensional embedding is generated for each survey taker (which will be discussed in subsection 5.2). The models are then fine-tuned for each training algorithm in line 8, 9, 10 and 11 (supervised classifier(SC), UCB, EG, and AG).

5.2 Fine-tuning

As discussed in Section 4.2, each survey taker provides information about their income, residence region, political affiliation, and education level, etc. We generated the survey taker's profile in a text format using these historical data. The Table 5 (in the Appendix) shows an example of text that we generated for a survey taker's profile. For the UMCSI dataset, there are around 600 survey takers every month (which might vary between months), and we fine-tune the models with these 600 samples. The fine-tuning procedure is as follows. We fine-tune the last month's samples after 60 months of continual pre-training (12 months x 5 years), and then test on the next month's sample.

Because there are five different survey questions, we fine-tune five different models. For each model, we fine-tune it using a supervised classifier and multi-arm bandits training algorithm (SC, EG, and AG).

Algorithm 1 Continual Learning on News corpus and S&P 500, and fine-tuning on Survey Data

for $data\ in\ (2014-2015, 2015-2016, 2016-2017, 2017-2018, 2018-2019)$ **do**
    encoder = pre-train(encoder, data)
    model1 = MLP(encoder, classifier)
    model2 = ContextualBandit(encoder)
    for $each\ surveyQuestion$ **do**
        Context = GenerateContext(encoder, surveyData)
        for $each\ in\ (Supervisedclassification, UCB, EG, AG)$ **do**
            Supervised_classifier(model1, Context)
            UCB(model2, Context)
            EG(model2, Context)
            AG(model2, Context)
        end for
    end for
end for

Table 1: Test Accuracy Using Different Training Strategies in Supervised Classification and Contextual Multi-Arm Bandit

Fine Tuning Methods	1st Snapshot	2nd Snapshot	3nd Snapshot	4th Snapshot	5th Snapshot
SC(Q1)	0.4458	0.5432	0.5543	0.6082	0.6875
SC(Q2)	0.5435	0.5242	0.5239	0.6143	0.6574
SC(Q3)	0.5389	0.5525	0.5356	0.5579	0.6485
SC(Q4)	0.5053	0.5342	0.5425	0.5932	0.6485
SC(Q5)	0.4564	0.5456	0.5982	0.6352	0.7034
UCB(Q1)	0.3821	0.4348	0.4854	0.5822	0.6252
UCB(Q2)	0.3245	0.3934	0.4354	0.5150	0.5152
UCB(Q3)	0.4023	0.4381	0.5208	0.5423	0.5396
UCB(Q4)	0.3831	0.4287	0.4929	0.5823	0.6349
UCB(Q5)	0.4564	0.5034	0.5723	0.6583	0.7083
EG(Q1)	0.3356	0.4345	0.4967	0.5242	0.5475
EG(Q2)	0.3113	0.392	0.4203	0.4345	0.4543
EG(Q3)	0.3564	0.3953	0.4422	0.4453	0.5334
EG(Q4)	0.4243	0.4035	0.4534	0.4563	0.4930
EG(Q5)	0.4564	0.5034	0.4835	0.5732	0.6359
AG(Q1)	0.3345	0.3852	0.4425	0.5435	0.6045
AG(Q2)	0.3054	0.3367	0.4035	0.4564	0.5135
AG(Q3)	0.3356	0.4253	0.4593	0.5103	0.5823
AG(Q4)	0.4501	0.4462	0.5024	0.6325	0.6823
AG(Q5)	0.4691	0.5409	0.5923	0.6832	0.7035
Average(Q1)	0.3745	0.4494	0.4947	0.5645	**0.6162**
Average(Q2)	0.3711	0.4116	0.4458	0.5051	**0.5351**
Average(Q3)	0.4083	0.4528	0.4894	0.5139	**0.5759**
Average(Q4)	0.4407	0.4531	0.4978	0.5661	**0.6146**
Average(Q5)	0.4596	0.5233	0.5616	0.6375	**0.6878**

To illustrate the effectiveness of continual pre-training and how the size of training new corpus affect accuracy, we run experiments with different number of continual pre-training. Each pre-training is using next 12 months' news corpus. For every 500 iterations (training on 12 months of news corpus), we save a snapshot of the encoder model and then fine-tune using survey data. In Table 1, snapshot 1 is pre-trained on 12 months news data, and snapthot 2 is pre-trained on 24 months (2 continual pre-training of 12 months news data), and so on. We run 2500 iterations (5 continual pre-training with 500 iterations on each) and save 5 snapshots of the encoders in total. The fine-tuning results of all five snapshots of the encoder are shown in Table 1 (the fine-tuning results are done based on 739 million parameters pre-trained encoder).

We run supervised classification (SC), UCB, EG, AG on all five questions (denoted as Q1 to Q5 in Table 1) on final fine-tuning. As shown in the Table 1, in the 5th snapshot(as the number of iterations increases up to 2500), some of the questions' accuracy can reach around 70% (for example, SC(Q5), UCB(Q5), and AG(Q5); accuracy is measured by "number of correct prediction"/"total number sample"). The increase of accuracy from 1st snapshot to 5th snapshot is due to the fact the pre-training loss decreases with more iteration. But it should also be noted that some questions' (such as Q2) accuracy does not increase in the same rate as others, and this is because the pre-traning corpus might not be diverse enough for the model to generalize

Table 2: GPTs Answers Accuracy on Five Survey Questions

	Q1(PAGO)	Q2(PEXP)	Q3(BUS12)	Q4(BUS5)	Q5(DUR)
GPT-3.5-Turbo	0.2218	0.3687	0.2268	0.1843	0.3724
GPT-4	0.2710	0.5143	0.0691	0.1625	0.2778

well.

It can be observed from the table that some of the questions have better accuracy than others with the same amount of iterations (for example, Q5 is generally better than Q2 regardless of which fine-tuning algorithm is used to train), and this is because there might be in-balance/bias in the pre-trained dataset and some categories of news are more than others and it leads to different accuracy. And also Q2 is generally a harder question than Q5 for custom model because Q2 requires the mapping from current data to prediction that forcast one year later (for the details of Q2, please refer to Table 4). From Table 2, we can also observe that supervised classification is generally better than most multi-arm training algorithms in most questions (except for UCB(Q5) and AG(Q5)). We speculate that this is because for relative small model (in the size of hundred of millions of parameters), gradient updates in supervised classification in the encoder can better maps the survey takers' profiles to answers.

Because different fine-tuning methods yield different accuracy, we compute the mean accuracy of four different fine-tuning methods for five of the survey questions (Q1 - Q5) and use those as final accuracy. As can be seen from Table 1, the mean accuracy of five survey questions ranges from 0.5351 to 0.6878 (highlited in bold font). As shown in Figure 3, we also plot the variance of the accuracy of five survey questions with different fine-tuning methods. It shows that the variance can be large within one question as well as across different questions. As shown in Figure 3, Q2 has the largest variance compared to all the other four questions and Q5 has the least variance.

5.3 Comparison with GPTs results

To further evaluate our proposed approach, we also conducted experiments using GPT API and asked the same survey questions to GPT-3.5-Turbo & GPT-4 and compared the results. We also want GPTs to understand that they are acting as a person who can only choose answers based on a person's profile context and specific time stamp. We give explicit prompt instructions to GPTs about not being able to look into the future.

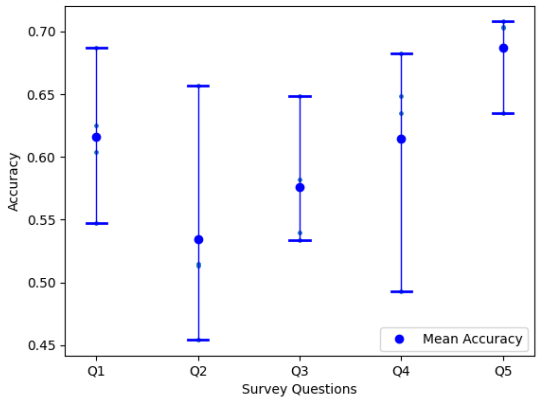


Figure 3: Plots that shows the accuracy variance of four different fine-tuning methods across five survey questions

Table 6 (in the Appendix) is an example of the text that was generated and fed to GPTs. For each survey taker's profile, we generate text similar to Table 6 (there are about 600 survey takers every month). We feed the 600 profile text to GPT-3.5-Turbo and GPT-4 API 5 times each with different ordering of the answer options (GPTs will give slightly different answers asking the question every time even with low temperature; 600 profiles * 5 runs = 3000 queries asked for each survey question for both GPT-3.5-Turbo and GPT-4). Out of the 3000 answers that GPTs provided, we calculate how many times GPT's answer matches the label. Table 2 shows the accuracy of GPTs answer (accuracy is the mean of the 5 runs as described above).

As we can see from the numbers in Table 2, GPT-3.5-Turbo has lower accuracy across all five questions than the proposed approach with the highest accuracy on Q5 being 0.3724, but it is still much less than the proposed approach (all four training algorithm including supervised classification, UCB, EG, AG have more than 0.6 accuracy on this question).

For GPT-4, it has significant higher accuracy on Q2 (0.5142) than the other four survey questions. And GPT-4's Q2 accuracy is comparable to proposed custom model's Q2 accuracy (0.5351). But in all four other survey questions, GPT-4's accuracy are still lower than proposed custom model.

Observing from results of both GPT-3.5-Turbo and GPT-4, it can be seen that GPTs have better accuracy in Q1 and Q2 compared to Q4. It makes senses because Q1 and Q2 post questions about economy conditions about last year and one year in the future whereas Q4 poses question about 5 years in the future. Therefore, Q4 is considered as "harder" questions for GPTs. GPT-4 has lower accuracy than GPT-3.5-Turbo in Q3, Q4 and Q5, and it validates the point that larger language model does not necessary guarantee better performance in prediction related to temporal data.

6 Limitation and Future Work

It should be noted that the proposed model (732 million parameters) that we evaluated is still a relative small model compare to modern large scales of language models. The news dataset that we train on is also small compared to the dataset that GPTs models are trained on.

The proposed model can always be scaled up push fine-grained performance. But with current experiment and comparisons, we are able to demonstrate that a small custom language model trained on temporal data (such as news data) can outperform GPTs model in downstream tasks such as real-time consumer sentiment prediction.

For future work, a larger news dataset source can also be used to make sure that the model captures the economic conditions in different regions of the country more comprehensively.

7 Conclusion

In this paper, we design a model framework for economic consumer sentiment prediction. To the best of our knowledge, this is the first work to use language model to predict the economic consumer sentiment using UMCSI data. We train a custom language model with subsequent classifier and Multi-arm bandit agent using news corpus, S&P500 data and UMCSI survey data. Our encoder-based model was pre-trained from scratch with a relatively small dataset and showed good accuracy with a relatively small model size at a low cost. We use continual learning in our experiments and compare the results with GPT-3.5-Turbo and GPT-4. Our experiment results show that we can out-perform both GPT-3.5-Turbo and GPT-4 on this task.

References

Akshay Chaturvedi, Onkar Pandit, and Utpal Garain. 2018. CNN for text-based multiple choice question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 272–277, Melbourne, Australia. Association for Computational Linguistics.

Zhipeng Chen, Yiming Cui, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. Convolutional spatial attention model for reading comprehension with multiple-choice questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6276–6283.

Huijian Han, Zhiming Li, and Zongwei Li. 2023. Using machine learning methods to predict consumer confidence from search engine data. *Sustainability*, 15(4).

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

John J. Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus?

Zixian Huang, Ao Wu, Yulin Shen, Gong Cheng, and Yuzhong Qu. 2021. When retriever-reader meets scenario-based multiple-choice questions. In *Conference on Empirical Methods in Natural Language Processing*.

Zixian Huang, Ao Wu, Jiaying Zhou, Yu Gu, Yue Zhao, and Gong Cheng. 2022. Clues before answers: Generation-enhanced multiple-choice QA. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3272–3287, Seattle, United States. Association for Computational Linguistics.

Di Jin, Shuyang Gao, Jiun-Yu Kao, Tagyoung Chung, and Dilek Hakkani-tur. 2019. Mmm: Multi-stage multi-task learning for multi-choice reading comprehension.

Gagandeep Kaur and Amit Sharma. 2023. A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis. *Journal of Big Data*, 10(1).

New York Times Developer Network. Accessed 2023. The New York Times APIs. https://developer.nytimes.com/apis.

OpenAI. Accessed 2023. ChatGPT3.5-Turbo. https://platform.openai.com/docs/guides/gpt.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2022. Leveraging large language models for multiple choice question answering. *ArXiv*, abs/2210.12353.

SP Dow Jones Indices LLC. Accessed 2023. sp500 data. https://fred.stlouisfed.org/series/SP500.

The Guardian News. Accessed 2023. The Guardian News APIs. https://open-platform.theguardian.com/.

University of Michigan. Accessed 2023. Surveys of Consumers. https://data.sca.isr.umich.edu/fetchdoc.php?docid=24774.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2023. Sotopia: Interactive evaluation for social intelligence in language agents.

Appendix

Table 3: News Corpus Example

Let's get real: the sharing economy won't solve our jobs crisis These days, everyone's talking about the so-called sharing economy. Newspaper columnists, pundits and tech reporters are – for the most part – enthusiastically explaining how new rental, resale and sharing services like Uber, Lyft, TaskRabbit and DogVacay are revolutionizing how we consume, and fostering entrepreneurship, conservation, cost savings and community spirit along the way. The prevailing narrative is that startups like these are the bright spots in an otherwise lackluster economy, and that if we could all learn to be better micro-entrepreneurs, our economy would recover faster.

Table 4: Survey Questions on Consumer Sentiment

Question	**Answer Options/Category Labels**
Q1(PAGO): Would you say that you (and your family living there) are better off or worse off financially than you were a year ago?	Better now; Same; Worse now; Don't Know (DK); Not Applicable (NA)
Q2(PEXP): Now looking ahead–do you think that a year from now you (and your family living there) will be better off financially, or worse off, or just about the same as now?	Better now; Same; Worse now; DK; NA
Q3(BUS12): Now turning to business conditions in the country as a whole–do you think that during the next twelve months we'll have good times financially, or bad times, or what?	Good times; Good with qualifications; Pro-con; Bad with qualifications; Bad times; DK; NA
Q4(BUS5): Looking ahead, which would you say is more likely–that in the country as a whole we'll have continuous good times during the next five years or so, or that we will have periods of widespread unemployment or depression, or what?	Good times; Good with qualifications; Pro-con; Bad with qualifications; Bad times; DK; NA
Q5(DUR): Generally speaking, do you think now is a good or a bad time for people to buy major household items?	Good; Pro-con; Bad; DK; NA

Table 5: Survey Taker Profile in a Text Format Example

Person information: income is 100000 dollars; income percentile is bottom 90%; home ownership status is renting; age is 31; birth year is 1984; living in the South of USA; gender is male; marital status is married/partner; number of adults is 2; education is Grade 0-8 without high school diploma; education is not a college graduate;

Table 6: One of the Survey Questions Asked to GPT-3.5 and GPT4

Acting as a person who is living in the year of 2020, month January. You can not see the future beyond 2020, January. Following is your information. Information: income is 100000 dollars; income percentile is bottom 90%; home ownership status is renting; birth year is 1984; living in the South of USA; gender is male; marital status is married/partner; number of adults is 2; education is Grade 0-8 without high school diploma; education is not a college graduate; Answer the following question and only pick one of the answer options. Just reply with the option that you pick. As can be seen, the GPT's answer accuracy is much lower than the proposed approach. Now looking ahead, do you think that a year from now you will be better off financially, or worse off, or just about the same as now? 1: Better now; 3: Same; 5: Worse now; 8:Don't Know; 9: Not Applicable