SentimentPulse: Temporal-Aware Custom Language Models vs. GPT-3.5 for Consumer Sentiment

Lixiang Li[1*] **Bharat Bhargava**[1] **Alina Nesen**[1] **Nagender Aneja**[1]
[1] Department of Computer Science, Purdue University, West Lafayette, IN, USA
{li4256, bbshail, anesen, naneja}@purdue.edu

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, our proposed real-time consumer sentiment analysis approach that leverages custom language models and continual learning techniques, and second, GPT-3 which is 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 model, and then classification and contextual multi-arm bandits to fine-tune the model, enhancing its adaptability and performance over time. We present a comparative analysis of the predictions accuracy of both architectures. 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

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, and afterwards benchmarked with a foundation model. 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 [Ope23] and compare the performance of both.

The paper presents three main contributions:

1. We proposed a comprehensive consumer sentiment analysis framework that leverages news and S&P500 [SP 23] dataset. Our framework can not only capture the consumer sentiment dynamics over time but 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. Our experiment results show that we can out-perform GPT-3.5-Turbo (Zero-Learning learning) on this task.

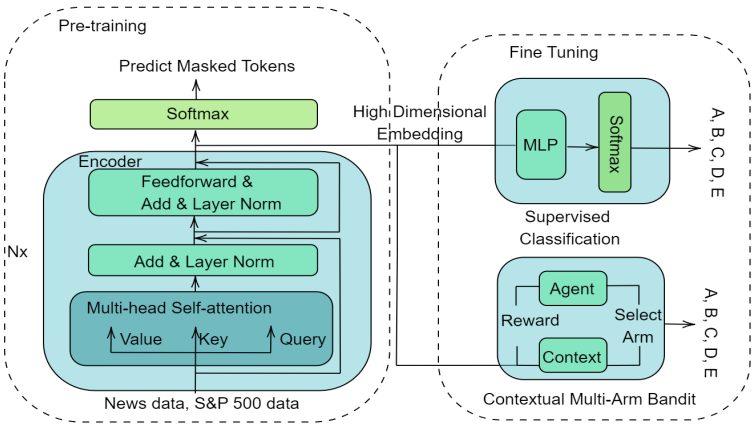


Figure 1: SentimentPulse: Two stages of training (Pre-training with Encoder; Fine-tuning with Supervised Classification and Contextual Multi-arm Bandit

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

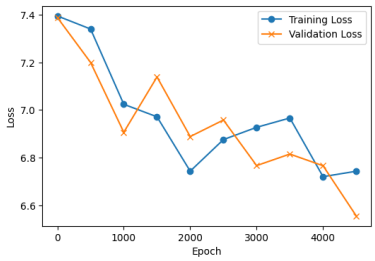
We provide a brief overview of some recent and related work on both consumer sentiment and multiple choice question answering methodology.

Consumer Sentiment Few works have been done on prediction on ecnonomic consumer sentiment using language model. [HLL23] 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.

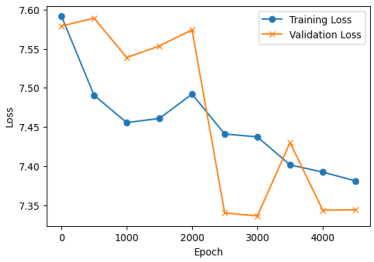
Multiple Choice Question Answering Methodology In this work, we try to predict the consumer sentiment by modeling it as a multiple choice question answering problem. Many of SOTA works are based on encoder-only architecture and encoder-based architecture has become a popular paradigm for MCQA problem [Hua+22]. Most recently, [Hua+22] uses transformer encoder-decoder architecture to generate a clue text input to an encoder-based MCQA block to enhance the performance. [RRW22] 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.

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.



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



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

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

3.1 Encoder

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+15]), 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 [New23], Guardian News API [The23], 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 and Guardian News by categories such as "Politics," "Economy," "Entrepreneurship," "International Business," "Automobiles," and "Business Day".(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. Table 3 in Appendix 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) [Uni23] for fine-tuning. UMCSI is 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 [Uni23]. 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. 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-GPUs setup, 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 5000 iterations before moving on to the next 12 months' news corpus and repeating the process. Figure 2 shows the training and validation accuracy with 5000 iterations using one 12-months of news corpus. We chose 5000 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).

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

for $data\ in\ (2018-2019, 2017-2019, 2016-2019, 2015-2019, 2014-2019)$ **do**
    encoder = Continual-pre-train(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	0.4835
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

Table 2: GPT-3.5 Answers Accuracy on Five Survey Questions

Q1(PAGO)	Q2(PEXP)	Q3(BUS12)	Q4(BUS5)	Q5(DUR)
0.2218	0.3687	0.2268	0.1843	0.3724

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 data. 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 600 samples after pre-training of each snapshot, and then test on the next month's 600 sample. Because there are five different survey questions, we fine-tune five different models. For each model, we fine-tune it using both a supervised classifier and a multi-arm bandits training algorithm (SC, EG, and AG).

To illustrate the effectiveness of continual pre-training, we run experiments with different number of continual pre-training. Each pre-training is using next 12 months' news corpus (we run 5000 iteration on each corpus). For every 5000 iterations, we save a snapshot of the encoder model and then fine-tune using survey data. We run 25000 iterations (5 continual pre-training with 5000 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 25000), 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 steadily decreases. 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 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), 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. 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)) and this can also be the fact that supervised classification updates the gradient in the encoder and better maps the survey takers profiles to answers.

5.3 Comparison with GPT-3.5 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 and compared the results. We also want GPT-3.5-Turbo to understand that it is acting as a person who can only choose answers based on a person's profile context and specific time stamp.

Table 5 (in Appendix) is an example of the text that was generated and fed to GPT-3.5-Turbo. For each survey taker's profile, we generate text similar to Table 5 (there are about 600 survey takers every month). We conducted Zero-Shot learning experiment and feed the 600 profile text to GPT-3.5-Turbo API 5 times each with different ordering of the answer options (GPT will give slightly different answers asking the question every time; 600 profiles * 5 runs = 3000 questions asked in total). Out of the 3000 answers that GPT provided, we calculate how many times GPT's answer matches the label. Table 2 shows the accuracy of GPT-3.5-Turbo's answer (accuracy is the mean of the 5 runs as described above). As we can see from the numbers in Table 2, GPT 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).

6 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 a language model to predict 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. Our experiment results show that we can outperform GPT-3.5-Turbo (Zero-Shot Learning) on this task.

7 Acknowledgement

The authors appreciate Gunika Verma and Shubham Pandey for downloading the dataset and helpful discussion in the project. The authors also appreciate the support from National Science and Engineering Council of Canada (NSERC) PGS-D scholarship.

References

[He+15] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[Hua+22] Zixian Huang et al. "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*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 3272–3287. DOI: 10.18653/v1/2022.naacl-main.239. URL: https://aclanthology.org/2022.naacl-main.239.

[RRW22] Joshua Robinson, Christopher Michael Rytting, and David Wingate. "Leveraging Large Language Models for Multiple Choice Question Answering". In: *ArXiv* abs/2210.12353 (2022). URL: https://api.semanticscholar.org/CorpusID:253098700.

[HLL23] Huijian Han, Zhiming Li, and Zongwei Li. "Using Machine Learning Methods to Predict Consumer Confidence from Search Engine Data". In: *Sustainability* 15.4 (2023). ISSN: 2071-1050. DOI: 10.3390/su15043100. URL: https://www.mdpi.com/2071-1050/15/4/3100.

[New23] New York Times Developer Network. *The New York Times APIs*. https://developer.nytimes.com/apis. Accessed 2023.

[Ope23] OpenAI. *ChatGPT3.5-Turbo*. https://platform.openai.com/docs/guides/gpt. Accessed 2023.

[SP 23] SP Dow Jones Indices LLC. *sp500 data*. https://fred.stlouisfed.org/series/SP500. Accessed 2023.

[The23] The Guardian News. *The Guardian News APIs*. https://open-platform.theguardian.com/. Accessed 2023.

[Uni23] University of Michigan. *Surveys of Consumers*. https://data.sca.isr.umich.edu/fetchdoc.php?docid=24774. Accessed 2023.

8 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: One of the Survey Questions Asked to GPT-3.5

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