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

Lixiang Li, Bharat Bhargava, Nagender Aneja, Alina Nesen
Department of Computer Science, Purdue University

Section 1: Abstract and Contribution

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 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.

Section 2: Model Framework

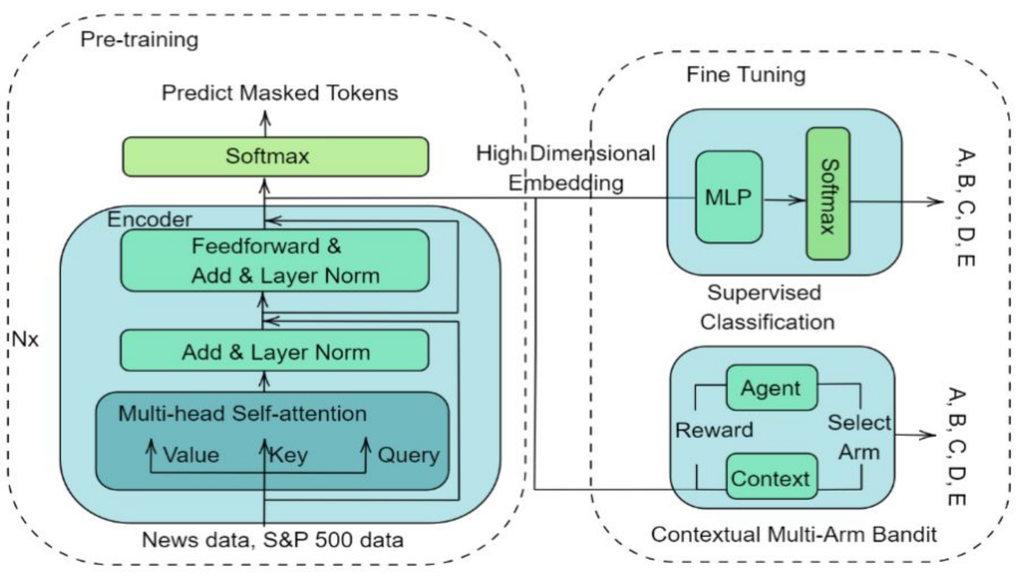

Figure 1: SentimentPulse: Two stages of training (Pre-training with Encoder; Fine-tuning with Supervised Classification and Contextual Multi-arm Bandit

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.

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 1: News Corpus Example

Section 3: Dataset

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 2: Survey Questions on Consumer Sentiment

For pre-training encoder, we use news corpus from New York Times News API, Guardian News API, and S&P 500 data. 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" (similar categories for "Guardian News").

We use survey data from the University of Michigan Consumer Sentiment Index (UMCSI) 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.

Section 4: Experiment Results

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)
```

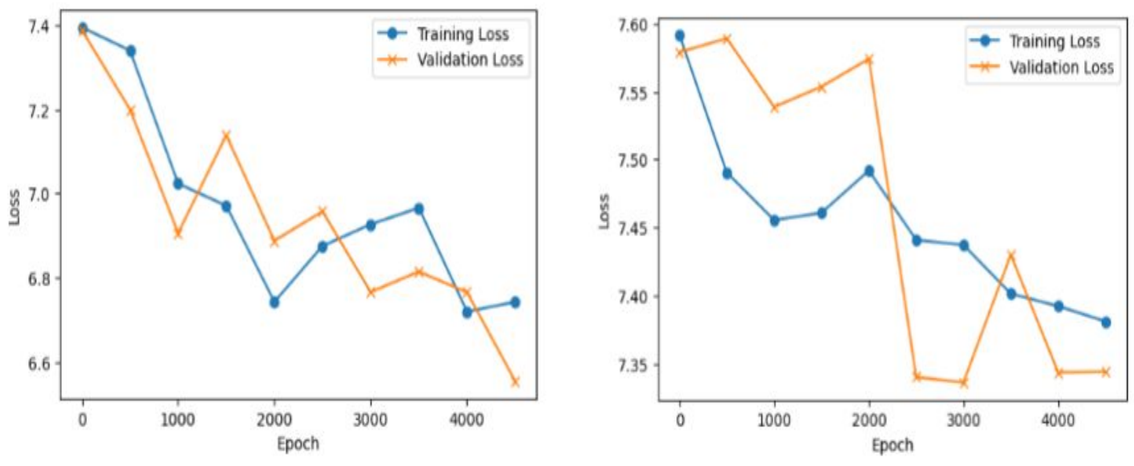

Figure 2: Cross entropy loss vs Number of iterations between the training set and validation set with two different settings of parameters of encoder

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 pre-trained continuously using corpus with different time stamps. For every 12 months of news corpus, we pre-trained the model for 5000 iterations (one snapshot) before moving on to the next 12 months' news corpus and repeating the process. The encoder undergoes pre-training on 12-months of news corpus continually. The training procedure is illustrated in Algorithm 1. The fine-tuning results of all five snapshots of the encoder are shown in Table 3. We run supervised classification (SC), UCB, EG, AG on all five questions (denoted as Q1 to Q5 in Table 3) on final fine-tuning.

We conducted experiments using GPT API and asked the same survey questions to GPT-3.5-Turbo and compared the results. Table 4 shows the accuracy of GPT-3.5-Turbo's answer (accuracy is the mean of the 5 separate runs). As we can see from the numbers in Table 3, 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).

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

Table 3: Test Accuracy Using Different Training Strategies in Supervised Classification and Contextual Multi-Arm Bandit

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

Table 4: GPT-3.5 Answers Accuracy on Five Survey Questions

Section 5: Summary

- We proposed a comprehensive consumer sentiment analysis framework that leverages news and S&P500 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.
- 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 on this task.
- 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.