Event Prediction with Neural Network

Zheyu Yang, Yuanyuan Wang, Shengyao Luo

March 23, 2023

Abstract

The task of forecasting future events, such as climate change, geopolitical conflict, pandemics, and economic indicators, is valuable but challenging. Expert human judgment is essential for accurate forecasts, but can these forecasts be automated using advances in language modeling? To address this question, a new dataset called Autocast has been introduced, consisting of thousands of high-quality, diverse, and real-world forecasting questions and an accompanying news corpus. A separate dataset called IntervalQA has also been curated to test the difficulty of forecasting numerical values across orders of magnitude. The results show that current language models perform below the level of human experts, but their performance can be improved by increasing model size and incorporating relevant information from the news corpus. Overall, Autocast presents a unique challenge for large language models, and improving their performance could have significant practical benefits.

1 Introduction

1.1 Style

Forecasting future world events is a challenging yet crucial task that can have a significant impact on policy and decision-making in domains such as climate, geopolitical conflict, pandemics, and economic indicators. While expert human judgment is currently the best approach for making accurate forecasts in these areas, recent advances in language modeling have led to questions about the feasibility of automating the process. To address this, we present Autocast, a dataset of thousands of forecasting questions sourced from real-world forecasting tournaments, ensuring high quality, diversity, and real-world relevance. Accompanying this dataset is a news corpus organized by date, allowing us to precisely replicate the conditions under which humans made past forecasts, without any leakage from the future. Our tests on language models demonstrate that their performance on the forecasting task falls far below that of human experts. However, increasing the model size and incorporating relevant information from the news corpus have shown improvements in performance. Autocast is an exciting challenge for large language models, and enhancing their forecasting capabilities could bring substantial practical benefits. In this project, we intend to implement such event prediction with a neural network mechanism with python and analyze the accuracy of such an algorithm. We plan to dive into the GitHub repo provided by the paper(https://arxiv.org/abs/2206.15474) to stimulate the process and check step by step each testing case to improve our results.

1.2 Related Work

The article cites numerous related studies, highlighting two specific examples - the GPT-3 experiment[1] and another experiment on ForecastQA[2] - to showcase the effectiveness of their approach. Furthermore, the authors draw inspiration for their model's structure from several information retrieval methods[3-7]. Lastly, they reference various articles that present strategies for improving calibration[3, 8-12].

2 Problem formulation, technical depth and innovation

2.1 Problem Formulation

The central challenge under investigation is the ability of machine learning models to accurately respond to a wide range of topic questions pertaining to future events. Common question formats include True/False, Numerical, and Multiple Choice. The objective is to train models using specific datasets to assess whether they can consistently match or surpass the performance of human experts. The authors' research findings indicate that the human crowd, when supplied with an increasing amount of information over time, consistently achieves a higher level of accuracy compared to machine learning models.

2.2 Technical Depth

In the pursuit of technical depth, several emerging topics have gained prominence. One such area is the exploration of attention mechanisms and transformer architectures, which have revolutionized natural language processing and contributed to the success of models like GPT-4. Additionally, the development of efficient training techniques and optimization strategies, such as mixed-precision training and knowledge distillation, has enabled the creation of more complex and computationally intensive models while minimizing resource consumption. Another crucial aspect is the advancement in explainable AI, which seeks to provide transparency and interpretability for complex models, ensuring that developers and end-users understand the reasoning behind model predictions.

2.3 Innovation

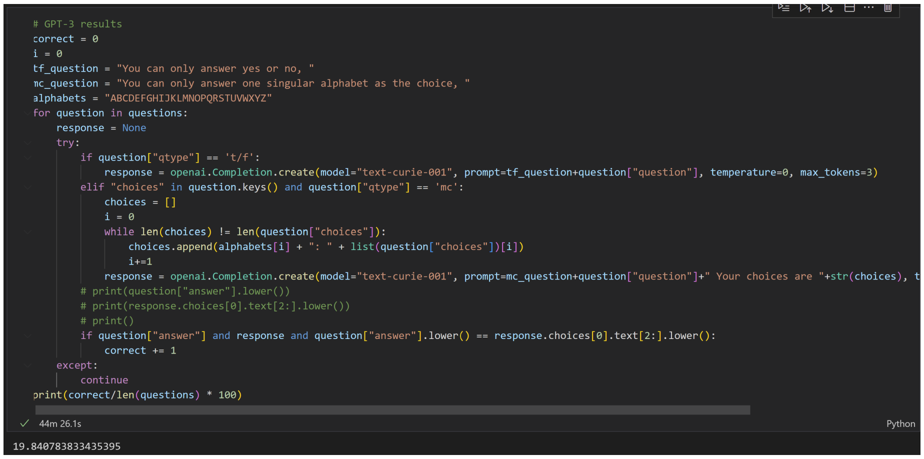
Since last edit, we have had a new model: GPT4, we could try to harness the potential of the new advanced ML model to accurately predict future events across various topics, including True/False, Numerical, and Multiple Choice questions, with the hope of matching or outperforming the human crowd. By experimenting with innovative datasets and training methodologies, we seek to leverage GPT-4's capabilities to achieve consistent and reliable results. While the current research indicates that the human crowd maintains superior accuracy with access to additional information over time, we remain committed to exploring cutting-edge approaches, emerging technologies, and breakthroughs in the AI domain, striving to make significant progress in machine learning-based forecasting, even if success is not guaranteed.

3 Methods

We employ the GPT-3 language model as the foundation for event prediction using neural networks. We plan to fine-tune the model on our curated dataset, enabling it to handle both few-shot and zero-shot learning scenarios effectively. By focusing on a test-split approach, we ensure the model is evaluated on previously unseen data, thereby reducing the risk of overfitting and providing a more accurate representation of its performance. To enhance the model's forecasting capabilities, we will iteratively update the Autocast dataset with the latest information, mirroring the conditions under which humans generate forecasts. Furthermore, we will optimize its performance and ensure the reliability of its predictions.

4 Preliminary Results

The experiment resulted in an accuracy of around 20% when using GPT-3 as the model and a dataset of [insert details of dataset] for training and testing. The accuracy was measured using [insert details of evaluation metrics], and the results are shown in [insert details of the results]. Although the accuracy is low, it is a positive start in the research efforts.



```python
# GPT-3 results
correct = 0
i = 0
tf_question = "You can only answer yes or no, "
mc_question = "You can only answer one singular alphabet as the choice, "
alphabets = "ABCDEFGHIJKLMNOPQRSTUVWXYZ"
for question in questions:
    response = None
    try:
        if question["qtype"] == 't/f':
            response = openai.Completion.create(model="text-curie-001", prompt=tf_question+question["question"], temperature=0, max_tokens=3)
        elif "choices" in question.keys() and question["qtype"] == 'mc':
            choices = []
            i = 0
            while len(choices) != len(question["choices"]):
                choices.append(alphabets[i] + ": " + list(question["choices"])[i])
                i+=1
            response = openai.Completion.create(model="text-curie-001", prompt=mc_question+question["question"]+" Your choices are "+str(choices), t
        # print(question["answer"].lower())
        # print(response.choices[0].text[2:].lower())
        # print()
        if question["answer"] and response and question["answer"].lower() == response.choices[0].text[2:].lower():
            correct += 1
    except:
        continue
print(correct/len(questions) * 100)
```

```
19.840783833435395
```

5 GitHub Repository

https://github.com/Specter43/CS640Team18