Master Thesis

Czech Technical University in Prague

F3

Faculty of Electrical Engineering
Department of Computer Science

Discovery of Causal Relationships from Textual Data

Jennifer Za Nzambi

Supervisor: Ing. Tomáš Mikolov, PhD.
May 2023

ZADÁNÍ DIPLOMOVÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Za Nzambi** Jméno: **Jennifer** Osobní číslo: **466091**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra počítačů**

Studijní program: **Otevřená informatika**

Specializace: **Umělá inteligence**

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Získávání kauzálních znalostí z textových dat

Název diplomové práce anglicky:

Discovery of causal relationships from textual data

Pokyny pro vypracování:

Cilem prace je nalezt kauzalni vztahy mezi ekonomickymi udalostmi, vyuzitim technik strojoveho uceni a pouzitim textovych dat z ruznych casovych obdobi. V prubehu projektu se predpoklada spoluprace s ekonomickym expertem, s jehoz pomoci bude navrzena metrika uspesnosti vytvoreneho systemu.
1. Seznamte se s literaturou v oblasti neuronovych siti, statistickeho jazykoveho modelovani a zpracovani prirozeneho jazyka
2. Seznamte se se zdroji textovych dat jako je Common Crawl, Wikipedie, Twitter a dalsi, kde je mozne ziskat data z ruznych casovych obdobi
3. Vytvorte statisticke jazykove modely s vyuzitim dat z bodu 2
4. Ve spolupraci s ekonomickym expertem (Filip Matejka, CERGE) vytvorte dataset, pomoci ktereho se budou automaticky vyhodnocovat vysledky modelu v ekonomicke oblasti
5. Vyhodnotte uspesnost modelu

Seznam doporučené literatury:

J. T. Goodman. A bit of progress in language modeling, extended version. Technical report MSR-TR-2001-72, 2001.
T. Mikolov. Statistical Language Models Based on Neural Networks, PhD thesis, 2012.
G. Zweig, C.J.C. Burges. The Microsoft Research Sentence Completion Challenge, Microsoft Research Technical Report MSR-TR-2011-129, 2011.
T. Mikolov et al. Advances in pre-training distributed word representations, LREC 2018.

Jméno a pracoviště vedoucí(ho) diplomové práce:

Ing. Tomáš Mikolov, Ph.D. RICAIP CIIRC

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) diplomové práce:

Datum zadání diplomové práce: **23.02.2022** Termín odevzdání diplomové práce: **26.05.2023**

Platnost zadání diplomové práce: **19.02.2024**

Ing. Tomáš Mikolov, Ph.D. podpis vedoucí(ho) práce	podpis vedoucí(ho) ústavu/katedry	prof. Mgr. Petr Páta, Ph.D. podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Diplomantka bere na vědomí, že je povinna vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

. _____ _____

Datum převzetí zadání Podpis studentky

Acknowledgements

I want to express my deepest gratitude to Ing. Tomáš Mikolov, PhD., for his exceptional guidance and supervision throughout this thesis and beyond. My sincere appreciation and many thanks go to Doc. RNDr. Filip Matějka, Ph.D., and his dedicated team, whose support and expert supervision in economic theory have been instrumental in weaving together the components of this thesis. I also wish to thank Ing. Jan Šedivý, CSc. and Bc. Tommaso Gargiani, whose invaluable feedback and collaboration on research beyond this thesis have significantly enriched the research presented.

I am indebted and thankful to JLJL for their love and support, without which this thesis would not see the light of day.

Declaration

I hereby declare that this thesis is the product of my independent work and represents my original research. All sources of information used in the composition of this thesis have been thoroughly cited and are included in the comprehensive list of references. This declaration is in full accordance with the methodological instructions for observing the ethical principles during the preparation of university theses.

London, May 2023

Abstract

Social media platforms are akin to an untapped gold vein harbouring a reservoir of public opinions, attitudes and sentiments which, if realised, could revolutionise how public opinions are gathered and interpreted. This thesis introduces a novel method for extracting opinions about economic indicators from social media texts by fine-tuning language models, specifically of the GPT-2 small, on Reddit comments. Through fine-tuning, language models can acquire the ability to understand and mimic the economic discourses within posts published on social media. This thesis' value is threefold. First, it presents carefully curated datasets through which the model can effectively learn domain-specificities and economic understanding on an advanced level. Second, it devises metrics based on perplexity comparisons of opposing statements which validate the model's comprehension of economic texts, thereby measuring the model's alignment with datasets it was fine-tuned on, and finally applies said models to Reddit datasets garnering results indicating that the model-based approach can rival, and in some cases outperform, survey-based predictions and professional forecasts in predicting trends of economic indicators. Beyond the scope of this study, the methods and findings presented could pave the way for further applications of language model fine-tuning as a complement, or potential alternative, to traditional survey-based methods.

Keywords: NLP, Language Models, GPT-2, Reddit, Social Media, Public Opinion, Transformers, Fine-tuning, Survey, Forecasting, Economics

Supervisor: Ing. Tomáš Mikolov, PhD. CIIRC, CTU in Prague

Abstrakt

Sociální sítě představují nedotčenou zlatou žílu, skrývající zásobu veřejných názorů, postojů a emocí. Pokud bychom ji dokázali plně využít, mohlo by to zásadně změnit způsob, jakým jsou veřejné názory shromažďovány a interpretovány. Tato práce představuje novou metodu pro extrakci názorů na ekonomické ukazatele z textů na sociálních sítích pomocí trénování jazykových modelů, konkrétně GPT-2 small, na komentářích z Redditu. Díky trénování mohou jazykové modely nabýt schopnost chápat a napodobovat ekonomické diskurzy na sociálních sítích. Přínos této práce je trojí. Zaprvé, představuje datasety, pomocí kterých může model efektivně získat poznatky specifické pro danou oblast a ekonomické znalosti na pokročilé úrovni. Zadruhé, práce prezentuje nové metriky založené na porovnání perplexity protichůdných tvrzení, které ověřují chápání ekonomických textů modelem a tím měří jeho naučení informací z datasetů na kterých byl trénován. Nakonec aplikuje dané modely na datasety z Redditu a prezentuje výsledky, které naznačují, že zvolený přístup může konkurovat a v některých případech dokonce i předčit predikce založené na průzkumech veřejného mínění a předpovědích vývoje ekonomických ukazatelů odborníky. Nad rámec této studie by metody a zjištění specifikované v této práci mohly razit cestu pro další aplikace trénování jazykových modelů jako doplněk nebo možnou alternativu k tradičním metodám založeným na průzkumech.

Klíčová slova: NLP, Jazykové modely, GPT-2, Sociální sítě, Veřejné mínění, Transformery, Ladění, Prognóza, Ekonomie

Překlad názvu: Získávání kauzálních znalostí z textových dat

Contents

Figures

Tables

Chapter 1

Introduction

Economics is a social science concerned with modelling the complex ways in which agents such as firms, governments, consumers, and foreign entities, amongst others, interact. Consequently, several variables in the economy and their changes are relatively predictable. For instance, if a left-leaning political party comes to power, it can be inferred with some degree of certainty that fiscal policy would become expansionary, and government social benefit spending will likely increase compared to a scenario where a right-leaning party took the helm. Similarly, suppose a country experiences a sharp decline in GDP. In that case, it is reasonable to anticipate that the central bank will step in with monetary policies geared towards stimulating the economy, such as by cutting interest rates or implementing quantitative easing measures to increase liquidity and encourage business investment, thereby bolstering economic activity.

However, despite the utility of historical precedents and economic findings in anticipating such events, understanding the deeply-rooted underpinnings of consumer opinions and behavioural patterns, as well as identifying their fears and plans, can prove to be a formidable challenge, as models striving for systematic ways of obtaining these are yet to produce generalisable results. Being able to predict events such as bank runs, panic-buying, boycotts, changes in tastes and fashions, and even far more nuanced daily decisions would be of great benefit not only to economists, who could bring to bear a deep understanding of the aforementioned in refining economic models and theories in their research, but could also be applied to studies in other domains.

The traditional tools used in social sciences to directly elicit public opinion are surveys and polls [94]. At the same time, these present challenges, such as high costs, biases in both sample selection and design and other factors hindering their predictive power. This sometimes leads to misestimations of public opinions, as exemplified by the 2016 US presidential election results or the Brexit referendum.

The central purpose of this thesis is to propose a novel method for uncovering opinions from the public domain, namely about economic variables such as interest rate, inflation, and unemployment. This is achieved by fine-tuning, a technique consisting of adapting a pre-trained model, which has learned general features of texts from a large dataset, to perform a more specific task using a smaller, task-specific dataset. This approach provides a better starting point than random initialisation and is particularly effective when training data for the specific task is limited. The model utilised was the Generative Pretrained Transformer 2 (hereinafter referred to as GPT-2) small, i.e. comprising 117 M parameters, [103] and throughout experiments within this thesis was fine-tuned on datasets consisting of economic textbooks, Investopedia articles, and comments from the popular social media platform Reddit.

It is posited that the knowledge of language gained through pre-training paired with domain-specific fine-tuning on public discourse can endow the model with the ability to grasp the public's opinions, beliefs and expectations effectively. Furthermore, given the fact that a considerable share of the global population engages in public online conversations, text corpuses extracted from social media platforms harbour a rich, untapped vein of data, which, if leveraged through capabilities of language models, could provide a potentially instrumental foundation in modelling human comportment and predicting future economic trends and decisions. If predictions retrieved align with actual economic outcomes, it would bolster the evidence for a potential causal link between public opinions and economic events, suggesting that opinions people hold, which then shape their actions, can be a driving factor in economic changes, thereby gleaning an insight into the interplay of public opinions and state of the economy, possibly allowing for forecasts of economic variables.

The key contributions of this thesis are the following:

1. **Curating domain-specific datasets**: The model was fine-tuned on domain-specific datasets extracted from economic textbooks and Investopedia articles to ascertain its capability of understanding economic jargon. Only once such capability was established was it applied to datasets of Reddit comments.

2. **Introducing perplexity-based metrics**: Perplexity-based metrics coined economic accuracy and perplexity ratio were developed to capture the fine-tuned model's alignment with the datasets it was trained on.

3. **Devising a framework for comparison with conventional tools**: In the experiments, it was shown that the results the fine-tuned models produced on public opinions of economic variables are comparable to those of the Survey of Professional Forecasters (SPF) and the Survey of Consumers, two renowned survey-based methods of gauging opinions on economics.

This thesis first presents a literature review focusing on survey methods and their limitations, applications of machine learning techniques and language models in economics and problems the methodology faces alongside similar approaches.

Next, the thesis delves into the foundational concepts of natural language processing. It gives a high-level overview of the inner workings of the GPT-2 model and the reasons why it was chosen.

Subsequently, the thesis presents the datasets used for experiments, the metrics applied, and the hyperparameter tuning methods.

In the experimental and results sections, experiments of fine-tuning GPT-2 on corpora extracted from economic textbooks and the online platform Investopedia along with the actual application of the model to Reddit datasets spanning the years of 2020 through 2022, are demonstrated, and the results analysed.

In the Evaluation section of the thesis, the potential limitations of this research and avenues for future exploration are discussed, along with an assessment of the value the proposed approach could bring.

Chapter 2

Literature Review

In this chapter, several aspects of affiliated literature on the topic of fine-tuning language models on data from social media for economic inference are addressed. Unfortunately, given the specificity of this thesis, there are not many sources which would share its focus. Nevertheless, topics ranging from current methods gathering public opinion, their shortcomings, application of language models in economics and their limitations, and even potential biases in datasets from social media are discussed and evaluated.

2.1 Surveys and Their Limitations

Surveys are often the go-to method for most social scientists when tasked with measuring opinions on any matter. As a technique perfected and heavily implemented in the 20^{th} century, it allows its designer to carefully craft prompts often designed by experts [38], specify the response domain, and target specific demographics, thereby ensuring representativeness [94].

Extending beyond the confines of respondent selection and question design, measuring public opinions through surveys also lends itself to utilising additional tools that can sway the results. Variables such as response time, question framing and specificity [19], and even the choice between multiple-choice or open-ended questions can significantly predetermine the results [75]. For instance, scientists in the 1980s discovered that when US respondents were faced with the question of what the most pressing issue the country is facing is, and subsequently a question of how good of a job the president is doing in his position, most assessed the leader's track record through the lens of the predicament they identified earlier [63]. This phenomenon, known as the priming effect [64], highlights how easily the order of questions can shape responses. This, along with other heuristics, biases, and even spatio-temporal and demographic discrepancies, has led many scientists to argue that creating a controlled environment for polls and obtaining accurate opinions without spurious conclusions is challenging [38].

One of the most widely known US surveys used to assess consumer behaviour is the Index of Consumer Sentiment (ICS) from the University of Michigan Surveys of Consumers [128]. Even though some papers have claimed the survey contains structural biases [127], others [37] countered these claims, demonstrating that the ICS remains an essential source of public opinions on economic variables. However, the survey's results still show signs of inconsistency and irrationality [107]. Whether these predicaments result from survey design, sample selection, or an inherent human trait remains a topic of debate. However, ICS is believed to hold great predictive power for inflation and other economic variables [119][29].

Nevertheless, demographically-linked biases indicate that less affluent individuals forecast differently than their wealthier counterparts and that these biases, albeit prevalent, are not consistent over time and are hard to capture [120]. Further research into how belief distortions change over time and affect macroeconomic fluctuations concerning the business cycle showed compelling evidence that agents are over-optimistic during economic expansions and over-pessimistic during contractions, suggesting that the implications of these tendencies are so powerful that they create causal self-reinforcing feedback loops between beliefs and macroeconomic outcomes [20]. The theme of an interplay between opinions causing economic variables to change and vice-versa is a widely accepted concept in economics [89]. Pessimism can amplify the impact of negative shocks, and analogously, optimism can multiply positive effects. This information only reinforces the importance of opinions for the economy and the potential limitations of surveys, suggesting that obtaining public opinions from social media could be a valuable alternative in establishing such causal links[119].

2.2 Natural Language Processing Methods in Economics

There were several instances when machine learning architectures were used to analyse textual data of economic importance. Some of the earliest applications of language models in finance involved mining financial news [80] and their classification, performing named entity recognition (NER) to identify people, organisations, locations, and other entities from textual data [5] [138], and other finance applications such as natural language-based finance forecasting [139] and accounting [45]. Other insightful applications [74] employed Long Short-Term Memory (LSTM), a type of recurrent neural network, to evaluate ad-hoc announcements made by firms and investigated whether the model could infer stock market fluctuations from them, which outperformed other ML techniques. In addition, various studies have utilised language models for financial tasks, such as employing the doc2vec technique to generate sentence embeddings paired with multi-instance learning to foresee stock market prices [87] or leveraging LSTM along with sentence simplification for analysing attitudes, resulting in the creation of the PhraseBank [88]. Additional research

has focused on using binary classifiers on tweets, demonstrating that Support Vector Machine (SVM) classifiers exhibit higher accuracy than Decision Trees or Naïve Bayes classifiers [131].

Soon after applications of NER [108], sentiment analysis, which stands for extracting opinions and sentiments from textual data and labelling them as positive, negative, or neutral [92], attained considerable traction among machine learning researchers and social scientists alike [140] [8] [73].

The study of sentiments and opinions over time using machine learning techniques is still an emerging field. An increasingly significant portion of sentiment analysis research focuses on spatio-temporal factors, examining how sentiments evolve and the shocks they are susceptible to. Curme, Stanley and Vodenska explored lagged correlation-based networks and their use in sentiment analysis based on index returns and news sentiment data [39]. Even though presented results showed a great promise, the method relied on annotated data, which severely impacts the scalability of the methods used, as they rely on expert-annotated datasets. Other notable papers analysed temporal dynamics of emotional appeals in 2016 Russian campaign communications in the US [97]. Some researchers believe that incorporating temporal dynamics and applying sentiment analysis in various industries is the way forward in furthering understanding sentiments holistically, however argue that these tools, in their present form, are unable to sufficiently establish causality or provide richer data beyond simplistic binary sentiment categorizations [108].

A great predicament researchers using automated methods in analysing sentiments and opinions face is that there is no well-defined procedure for data interpretation. Algorithms concerned with sentiment analysis are often based on the labelling of certain words as bearers of either positive or negative connotations and, through them, determining the polarity of sentiments [36]. An example of this method is exemplified by models analysing financial sentiments through carefully crafted financial lexicons, such as the ones presented by [86]. Such approaches heavily depend on word counting and can easily overlook the semantic nuances of sentences. On the other hand, they are easy to implement, adaptable through domain-specific lexicons that capture intricacies of each domain's jargon and do not require labelled data for prediction.

Despite the valuable insights sentiment analysis offers regarding public opinions, current literature acknowledges its limitations as conventional sentiment analysis needs to provide more detail and fully flesh out complex relationships of analysed variables [108] for which the mere determination of sentiments is not expressive enough. When extracting more nuanced opinions on open-ended and more convoluted questions, public views on statements and their confidence in their resolve are sparsely covered in literature, as there does not seem to be a unified approach when it comes to measuring these aspects.

2.3 Economic Applications of Large Language Models

The lack of expressiveness when applying sentiment analysis methods could potentially be tackled with other architectures. According to some researchers, large language models (hereinafter also referred to as LLMs) inherently possess a comparative advantage in an array of tasks ranging from summary creation, code writing, research, ideation, or even mathematical computation [71]. Recently published models, such as OpenAI's ChatGPT and GPT4 [72], DeepMind's Chinchilla [57], and Anthropic's Claude [7], have significantly influenced the perception and utilisation of language models. Moreover, with their potential for automation and the emergence of numerous businesses based on or related to language models, LLMs are expected to and, to an extent, have already revolutionised the economy [110] [121]. This versatility lends itself to being widely used and adopted by researchers in economics.

A relatively novel application of LLMs in economics is represented by their use to create synthetic data. A widely known concept in economics is the so-called *homo economicus*, which, similarly to concepts of perfect gas or frictionless surface in physics, describes a perfectly rational entity many economic models are founded on [124]. In their 2022 publication, Argyle et al. [9] coined the term *algorithmic fidelity*, describing the extent to which a model emulates a person based on a given prompt. First, they conditioned a GPT-3 model on 'silicon samples' using backstories of real-life human participants from extensive surveys. They then observed how closely the LLM's responses mirrored those of people they were 'fine-tuned' on. Given the considerably nuanced and multifaceted similarity between answers of the language model and that of the respondent, they gave rise to a new tool for creating implicit computational models of humans - a *homo silicus* [59].

Another application of language models in economics, this time in a conventionally microeconomic direction, is presented in [59], where language models were used to answer as part of behavioural experiments such as those conducted by Kahneman et al. in 1986 [66]. In the study, GPT-3 model was presented with the following dilemma: A store is selling shovels for USD 15 and, after a snowstorm, raises its prices to USD 20. The model was then asked how fair this price increase was, which was overwhelmingly deemed unfair according to respondents in the 1986 study. The results of this and other experiments bolstered the author's resolve that this is an affordable method with high validity for emulating human behaviour. Some argue that when endowed with social preferences and further information such as political allegiance or stance on fairness, the results become even more accurate, sparking further debate and interest in $in - silico$ experiments [71].

Further research on the topic continues to emerge. For example, [34] suggest that language models, when trained on the media diets of respondents, can

predict respondents' beliefs and, when scaled, could effectively mirror public opinions. In their experiment, a pre-trained BERT model was used and fine-tuned on media diets of survey respondents. Authors then adapted the model to generate probability distributions of words that would fill in blanks in the prompts given. Comparing the words chosen by the model to those selected by respondents, the results made authors confident that a similar media-diet approach could supplement surveys and forecasts and even serve as a retroactive tool to analyse public sentiments.

Despite the plethora of various novel applications of LLMs, such as extraction of opinions from media diets or $in-silico$ experiments, studies applying LLMs specifically to distill opinions from social media platforms, which would then contrast findings obtained with survey data, have not been found in literature. This conspicuous absence accentuates the unique opportunity this thesis' research has to contribute to the underexplored area of scientific inquiry and bridge this identified gap in literature.

2.4 Fine-Tuning Pre-Trained Models

In spite of potentially more sophisticated architecture, the issue of labelled data remains a challenge even for LLMs. While some researchers work with labelled data [74], a significant portion of language models trained on domain-specific texts, like social media posts, do not have access to data labellings, which makes the pursuit of inference even more complicated in the context of analysing sentiments, as even with pre-trained values in the first (word embedding) layer, not enough labelled data makes the task of learning complex interactions rather precarious. Therefore, many researchers believe that a more promising application would be using already pre-trained language models and fine-tuning them in relation to classification tasks [8].

Some even argue that Transformers are superior to other methods and can achieve results comparable to those of finance experts [92]. According to the authors of the Transformer architecture, it offers an alternative to RNNs (recurrent neural networks), as it bears advantages such as self-attention and the ability to bypass sequential computation, which is particularly useful when dealing with large datasets such as the ones extracted from social media platforms [130]. In addition, state-of-the-art Transformers trained on large corpora can be leveraged through further fine-tuning. Through the process of domain adaptation, as popularised by [60], the model can be then adapted to smaller datasets, provided they do not differ greatly from the original ones, and boost its performance through deep understanding of specialised domains [43], such as economic opinions expressed through social media. This allows scientists to focus on fine-tuning and thereby improving domain adaptation of models rather than starting training LLMs from scratch [102] [26].

OpenAI's GPT, released in 2018, featured 117 million parameters, the same

amount as the GPT-2 model used within this thesis [102]. Since then, both OpenAI and other companies have introduced newer, larger models, with some boasting billions of parameters [102]. Given the inaccessibility of some of the largest and most potent models, one could think that any research with a smaller model is inherently inferior. However, studies have demonstrated that when appropriately fine-tuned, smaller models can perform just as well and sometimes even outperform large few-shot learning models, which are models that can grasp new concepts with very few prompt examples [122] [57], suggesting that not pre-training but rather fine-tuning influences the model's effectiveness for inference tasks.

These findings were instrumental for this thesis and the decision to employ an already trained language model and fine-tune it instead of trying to garner knowledge through training from the ground up and for the contention of using a smaller model in hopes of achieving comparable results to those obtained through larger and newer ones.

2.5 Social Media as a Source of Textual Data

When it comes to what kind of data would be suitable for analysing public opinions, social media emerged as a popular source economists welcome and consider a valuable complement to traditional survey data [126]. Several studies report that whilst both social media and polls may exhibit biases researchers should tackle and try to minimise, social media offers distinct advantages regarding accessibility, scale, and efficiency [38]. The sentiments expressed in social media posts can be readily converted into quantifiable scores, thereby offering a way of supplementing survey findings [36].

A wealth of publications substantiates the existence of clear links between data in economic surveys and sentiments found in social media. A 2010 study established high correlations between Twitter sentiment scores and both The Gallup Poll and the Michigan Surveys of Consumers, both of which assess public opinion on a diverse array of topics, spanning politics, economics, and social issues [94]. Even less convoluted methods, such as those premised on counting word frequencies in Twitter data, demonstrate correlations between Twitter-derived information and US unemployment insurance claims [6]. Some researchers not only underscore the efficacy of Twitter data in predicting French election results but also identify their high fidelity when compared to survey data [31], which further reinforces the notion of social media in some instances potentially serving as a proxy for traditional surveys and polls, potentially capturing public opinions better.

Nevertheless, some studies caution that since social media data, unlike surveys, lack essential benchmarks and checks, it can only partially supplant polls and survey methodologies [6]. The benefits language models trained on social media datasets bring to various fields have been extensively documented

[10] [79]. However, some researchers maintain that the predictive power of social media data can sometimes be overstated [49] [48]. Indeed, data from social media has intricacies that warrant consideration when used in scientific endeavours. It has been posited that only once these intricacies are addressed can social media data be deemed suitable for public opinion analysis.

Several studies report that demographic groups online are not representative samples of the broader population [14]. Research shows that due to the uneven distribution of internet access, young users and individuals from developed countries tend to be overrepresented [100]. This assertion is supported by statistics: In 2020, the World Bank reported that over 90% of Americans had internet access, whereas the figure was as low as 10% in many African countries [13]. Pew research reports significant disparities in internet access even within the US, spanning urban/rural divides, age groups, and racial demographics [30]. A 2019 study by Pew Research revealed that a mere 22% of US Twitter users were representative of the wider population, with young Democrats, whose opinions on political predicaments such as immigration or gender inequality often differ from that of the broader US population, were the majority [91]. Another potential cause for concern is that this demographic, specifically the 10% of the most active Twitter users, accounts for 80% of US-generated content on the platform [134]. In the case of Reddit, data indicates a similar overrepresentation of young males; however, when considering race, the platform's user base seems to mirror the racial group composition of the US more faithfully [30].

The value of social media data is contingent upon researchers' ability to overcome its limitations [14]. Factors such as the inability of social media to represent the opinions of all [12] lead some to view social media data as a nonprobability sample. Rather than considering it a robust stand-alone source of information, some researchers suggest either using a so-called calibration survey to verify whether a small sample of the demographic composing social media data would yield similar results [36] or to pair findings elicited by a language model with polls and present them as a heterogeneous result [41] [98]. However, surveys, as already mentioned, have many limitations of their own. Metadata, often included in data from Reddit or Twitter, can address some concerns regarding external validity, potentially discouraging researchers from opting for a combined survey-language model approach [6].

2.6 Biased Language Models

As established, some report that social media data may not be suitable for economic research, as the sample of online posters may not be representative of *vox populi* researchers may take an interest in [6]. Many believe that economists would much prefer analysing actions rather than statements [59]. However, relying solely on surveys only exacerbates this predicament.

Another more apparent reason why economists often take a dim view of using language models applied to social media data is the volume of biases language models often contain, such as gender bias and stereotypical associations [15], which are engrained into language models even on the word embedding level [78] [141] [25]. For example, some studies have shown that African American names' embeddings are closer to unpleasant words than European-sounding names [27], while other publications assert that words such as Islam are disproportionately associated with charged words such as 'terrorism' or 'violence' [33]. Such biases seem to persist even after large language models such as GPT-3 are fine-tuned [50] and sometimes can even be amplified [125]. On a more positive note, evidence suggests that the GPT-2 language model family used in this thesis exhibits less bias than many other language models such as BERT, ROBERTa or XLNet [93].

It is worth noting that biases are a normative concept researchers often identify inconsistently and struggle to mitigate [23]. Evaluating the extent to which biases may interfere with the results of language models is also crucial, as most online texts are not composed solely of makeshift personas, but reveal opinions and intentions language models can effectively learn and identify. The implicit biases they carry do not always impact the results [59]. Nevertheless, language model bias should be considered, especially that of pre-trained larger models [17].

Juxtaposing the biases of both language models and social media datasets can yield valuable insights, as political allegiance is not only observed in social media but in human-supervised language models as well [99] [54]. Some research publications discuss a general misalignment issue between opinions of the broader American population and language models, claiming that according to the text-davinci-003 of the GPT-3 family, the US President J. Biden would garner an approval rating exceeding 99%, which is far from reality [112]. Moreover, opinions of older, widowed, religious individuals are most likely misrepresented across several large language models. Several researchers further state that even if LLMs harboured prejudice aligned with that of the groups whose opinions they seemingly represent the most, these biases often lack nuance and consistency [112].

In light of these observations, concerns persist about language model training experiments reproducing the biases of their authors or the text they were trained on. Being aware of the biases embedded in language models, some researchers propose so-called Turing experiments to assess the reproducibility of LLM-elicited results in the context of microeconomic experiments. These experiments aim to discern the extent to which observed behaviour is attributable to the biased authors, as opposed to reflecting the behaviour of the broader population [2].

Furthermore, several attempts at a systematic mitigation of biases in language models, a process called debiasing [95], are focused on how prompt

design, i.e. design of text inputted into language models, can help with more balanced results. Steerability in language models pertains to the capacity to control the output of the model by stipulating specific attributes or characteristics of the generated text [112], with the goal of uncovering public opinion that may otherwise be obscured. One potential remedy proposes steerability-focused tactics to circumvent biases, which concerns careful prompt design. There are various techniques of 'prompt engineering', methods, which involve the deliberate crafting of prompts provided to reveal the most probable response to a given question or prompt [52].

The few-shot prompt method, a popular yet not an infallible approach, entails presenting a prompt alongside several illustrative examples to a language model and anticipating a consistent response [26]. However, like human respondents, LLMs may misinterpret or misconstrue prompts, resulting in unintended answers. Hence the urgency of carefully choosing formulations of probes used on LLMs.

A chain-of-thought prompt is, according to some publications, a more effective way of providing models context aiding in the subsequent unpacking of a potentially complex question and answering it in a less biased fashion. The method recommends crafting prompts that mimic the process of thinking out loud, articulating step-by-step reasoning to guide the LLM through the proposed line of thought [132]. The method is aligned with literature claiming that giving models more context promotes improved results [116] [9].

Both circumspect prompt design and other methods of result improvement, such as self-debiasing [113] and careful consideration of potential effects of demographic groups which created datasets trained on, and inherent biases of LLMs concerning their application, are warranted and were carefully considered during the experimental phase of this thesis. At the same time, understanding the intended outcome of carefully crafted prompts is a crucial concern to address [112] and further highlights the need for caution when analysing results such as those presented within this thesis before making generalisations and inferences.

According to some publications, developing a methodology capable of genuinely grasping the essence of the wider public's opinions remains elusive, yet merits great focus of future research [36]. Several papers also indicate that while some older architectures have been extensively applied to sentiment analysis, such as Bayesian networks or SVMs, many of the newer ones, such as Transformers or even LSTMs, are underutilised. Some argue this is due to affordability and hard approachability for multidisciplinary teams. Nevertheless, expectations of the future of the field gravitate towards applications of large pre-trained models to opinion analyses [92] [108].

Despite challenges related to mitigating model and text biases, procurement of appropriate datasets, automating result extraction, and the unavailability

of validation tests and benchmarks that would aid with result interpretation, the literature still lacks practical applications of language models that would focus on assessing opinions about real-world variables and their impact on economies. That is why it is believed that research presented in this thesis can provide a key stepping stone for a new kind of research transcending AI and Economics, which uses LLMs to discern patterns and predict shifts in behaviour to the extent polls, economic models or sentiment analysis cannot, thereby bridging a gap in literature.

Chapter 3

Technical Background

This chapter outlines the fundamental components of the Transformer architecture. Namely the GPT-2 small model key to this thesis. The model employed consists of 117 million parameters, 12 layers (decoder blocks), 12 attention heads and an embedding dimension of size 768 and was accessed through the Hugging Face platform [42].

The chapter delves beyond a mere architectural overview, explicating the foundational concepts which underpin the model's inner workings, such as perplexity, tokenisation, self-attention, among others, as they are integral to understanding the research methods employed in subsequent chapters.

Even though providing a comprehensive exploration of alternative architectures which could have been employed for similar NLP (natural language processing) tasks would exceed the scope of this thesis, this chapter offers a rationale for the selection of the GPT-2 small model for the research at hand.

3.1 Language Model

A language model is a probability distribution function over sequences of words [35]. It assigns a likelihood to a sentence or can predict the probable succeeding words given a sentence or fragments of it. The objective of many language models is to model human written and spoken word as closely as possible, continually refining their estimates during training.

3.1.1 N-gram Models

One of the simplest models attributing probabilities to sequences of words is called an n-gram. The 'n' in its name denotes the number of words in a sequence; hence a 2-word sequence of words such as 'please come' is a bigram, a three-word sequence like 'what a show', a trigram and so forth. Some of the most widely used n-grams are trigrams, as computational complexity

increases with the length of word combinations [65]. N-grams use the chain rule to decompose the joint probability of a sequence of words into a product of conditional probabilities, as per the Markov assumption [65]. This way, models do not have to work with entire sentences when determining their probability but merely parts of sentences n words long.

Consider the example of a word $w =$ 'cool' and its history $h =$ 'Gustav is'. In determining the likelihood P('cool'|'Gustav is') in a bigram model, the probability is approximated by P('cool'|'is'), as it considers only one preceding word. Analogously, a trigram would look two words into the past and an n-gram n-1 words into the past. The approximation can also be denoted as $P(w_n|w_{1:n-1}) \approx P(w_n|w_{n-1})$[1]. This means that the formula for computing the probability of a word sequence of n words is equivalent to the one in equation (3.1).

$$P(w_{1:n}) \approx \prod_{k=1}^{n} P(w_k|w_{k-1}) \tag{3.1}$$

The formula is subsequently estimated using the maximum likelihood estimation (hereinafter referred to as MLE). Parameters for the MLE estimation are gathered by counting each 2-word pair's occurrence in corpus $C(w_{n-1}w_n)$ considered and subsequently normalising obtained counts as follows:

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)} \tag{3.2}$$

Other models have superseded n-grams, as they are too simplistic, and to capture long-distance dependencies languages are known for, n-grams would prove inefficient.

3.1.2 Perplexity

Perplexity (often abbreviated as PPL), closely linked to entropy, measures how well a language model can predict given text and is often used to evaluate a language model's performance. The lower its value, the better the model is at correctly predicting the next word in a given sequence [65]. The range of values of perplexity is from 1, which is equivalent to a perfect prediction, a situation generally unattainable in practice, to infinity, as perplexity is equivalent to the weighted average branching factor, the average number of equally likely choices the model makes when predicting the next word. This factor typically increases as the model considers more words, thereby indicating the model's

[1] A note on notation, the author uses an abbreviated form $w_{1:n-1}$ to allude to a sequence of n words. The joint probability of words the n^{th} word's probability is conditioned on implies that the k^{th} word corresponds to the k^{th} position within a sequence of words, where k ranges from 1 to $n-1$.

growing uncertainty with more choices. In its mathematical form, it is defined as:

$$PPL(W) = \sqrt[n]{\frac{1}{P(w_1, w_2, \ldots, w_n)}} \tag{3.3}$$

i.e., the perplexity of a set of words, W, is the inverse probability of the set normalised by the number of words at hand [65]. From this, it is clear that the lower the model's perplexity is, the better it is at estimating the probability of a specific sentence.

Perplexity is particularly useful for language model evaluation, as it captures the uncertainty inherent in predicting the next word in a given sequence. It conveys a robust way of quantifying language models' predictive capabilities.

3.2 Tokenisation

Language models are concerned with analysing texts. To infer results, however, the texts must first undergo tokenisation, an indispensable step in the pre-processing pipeline of the GPT-2 model whereby raw text is converted into a sequence of input IDs.

A token is the smallest indivisible unit of semantic significance a model can understand and manipulate. Their use aids in representing original text in a structured format and enables language models to learn subword patterns and relationships more efficiently [114].

3.2.1 Word Tokenisers

There are several methods of tokenisation. Some are based on words, assigning each an input ID. This method, however, is not flexible when it comes to word conjugations and derivatives [43]. For instance, 'car' and 'cars' would have different tokens attributed to them, which could, at a large scale, lead to a rather long vocabulary, i.e. the total number of words used, straining computational resources [43]. A way of circumventing this predicament is using the so-called 'unknown ID' to denote words not present in the tokeniser's vocabulary. That way, only a certain number of the most frequently used words in datasets can be included in the vocabulary, the remainder labelled as unknown. This labelling is a compromise that, in some ways, resolves issues linked to a strain on computational resources elicited by a large vocabulary. However, at the same time, it could diminish the model's results, which would likely suffer from loss of information if important words were omitted from its vocabulary.

3.2.2 Character-Based Tokenisers

At the other end of the spectrum, a character-based tokeniser splits words into characters and can make do with less than 300 tokens, each corresponding to a letter or a special character used in English. A character-based tokeniser would split the word 'cars' into three tokens: 'c', 'a', 'r' and 's'. Its advantage over a word tokeniser is that it has fewer out-of-vocabulary words, as all can be expressed through tokens. It is beneficial for expressing misspelt words rather than discarding them [106]. On the downside, letters in English contain less semantic value than words, so to grasp a word's meaning, the tokeniser needs far more tokens than had it been based on words alone. This leads to longer sequences of tokens being processed by the language model, which in turn will cover less context given the limited context window, reducing the size of the text that can be used as input to the model.

3.2.3 Subword-Based Tokenisers

As the name of the category suggests, subword-based tokenisers represent the intermediary solution between character-based and word-based tokenisers. They are governed by the principle of splitting less common words into subwords and leaving the most common ones intact [123]. A subword-based tokeniser would possibly split the word 'cars' into the word 'car' and the letter 's'.

Most of the state-of-the-art language models use subword-based tokenisers such as WordPiece (Bert, DistilBert), Unigram (XLNet, ALBERT) or the one GPT-2 uses, the Byte-Pair-Encoding [115] (hereinafter referred to as BPE) tokeniser [43].

GPT-2-used BPE tokenisation method was developed from an iterative data compression algorithm, as it merges common pairs of characters or symbols to create new tokens. This approach is notably practical even when handling out-of-vocabulary terms.

Algorithm 1 Byte-Pair Encoding [115]

Procedure BPE S, k
$V \leftarrow$ all unique characters in S
while $|V| < k$ **do** ▷ Merge tokens until k-sized vocabulary
    $t_L, t_R \leftarrow$ Most frequent bigram in S
    $t_{\text{NEW}} \leftarrow t_L + t_R$ ▷ Concatenation creating a new token
    $V \leftarrow V \cup \{t_{\text{NEW}}\}$
    Replace each occurrence of t_L, t_R in S with t_{NEW}
end while
return V
End Procedure

As conveyed in Algorithm 1, BPE starts with a set of words (strings) D,

denoting the text it is trained on and a pre-determined vocabulary size k. The set of strings is created during pre-tokenisation, which splits training data into words. The algorithm starts by creating another set V comprising all unique characters in strings at hand, each representing a token in the sense of a character-based tokeniser. Then, until the size of the set V is lower than k, it finds the most frequent bigram in D and uses it to create a new token. For instance, if said bigram is the pair of letters' i' and 't', a new token will be 'it' and added to the set of tokens V. Subsequently, the new token replaces all occurrences of the two tokens which it is composed of in D and the procedure repeats until the desired vocabulary size is reached, i.e., until the condition $|V| = k$ is met.

To implement tokenisation with GPT-2, the raw text data must be fed through a tokeniser suitable for this model. This tokeniser will split the text into tokens, map tokens to their corresponding IDs, and create the input format for GPT-2.

Following the model's tokenised input processing, the output undergoes a decoding process, converting token IDs back into human-readable text. This step ensures that the model's predictions are interpretable and can be evaluated against the original text. The vocabulary size, comprising the sum of the number of merges and the initial vocabulary size, is a hyperparameter. In this case, it was set to 50,257, the default value set by the Hugging Face API's pre-trained tokeniser employed [103].

3.3 Embedding and Positional Encoding

Once the sentence is tokenised, all tokens, represented by their IDs, undergo embedding and positional encoding.

3.3.1 Embedding

The tokens are converted into a continuous vector representation in the pre-trained embedding layer. This vector becomes the semantic representation of the token [130]. For the GPT-2 small model used in this thesis, the default dimension of embedding, d_{model}, is 768 [103]. The size of the dimension of the resulting matrix is defined as the product of the *context_window* (maximal sequence size the model can process at once, i.e. 1024 for the GPT-2 small model employed) and d_{model} [103].

3.3.2 Positional Encoding

The architecture of the GPT-2 model is not inherently aware of the order of tokens it processes. That is why in the positional encoding step, each token is attributed a vector that specifies its position within the context window the model considers.

Positional encoding (PE) is created using sinusoidal transformations creates vectors of the same dimensionality as the embedding vectors, i.e. $context\_window \times d_{model}$, as follows:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{3.4}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{3.5}$$

Where pos is the position in the input sequence, and i is the dimension up to $d_{model}/2$; hence each dimension of the positional encoding corresponds to a sinusoid [130]. The transformation chosen uses sinusoid attributes, making it easier for the model to learn to attend to relative positions. For any fixed offset k, PE_{pos+k} can be seen as a linear function of PE_{pos}, thus helping the model better understand the relative order of tokens in a sequence. The transformation also can be easily adapted to longer input sequences, as sinusoidal functions output different and smooth values for each position, and as per authors of the Transformer, are easy to compute [130].

The positional encoding and embedding matrices are then added element-wise, and the resulting matrix thus possesses both positional and semantic information for each token. It is then further passed on through Transformer layers. Specifically, rows of the resulting matrix are vectors for each token in the sequence at hand.

3.4 The GPT-2 Architecture

This subsection provides an overview of the Transformer architecture used within the GPT-2 model. GPT-2 is composed of so-called Transformer decoder blocks, which are illustrated on the right side of Figure 3.1. The GPT-2 small variant employed has 12 blocks overall [103].

Each block comprises two sub-units: self-attention and a position-wise feed-forward neural network. There are, however, other steps, such as layer normalisation and residual connections. This section explains each part's involvement in the training pipeline and how the GPT-2 architecture differs from the original Transformer's.

3.4.1 Self-Attention

According to many, self-attention is the pièce de résistance of the Transformer architecture. It is a process that allows the model to weigh the importance of each token from the input sequence against all others, thereby capturing contextual information while avoiding sequential computing [130].

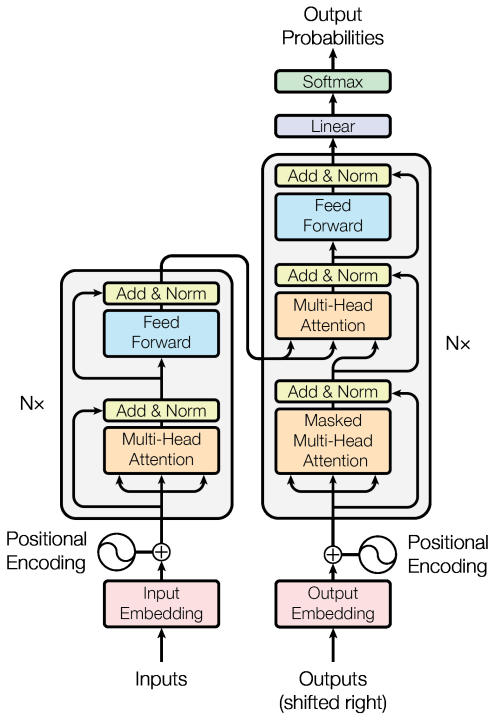


Figure 3.1: Original Transformer architecture composed of encoder (left side) and decoder (right) blocks. The GPT-2 model is based on N=12 modified decoder blocks only [130].

Query, Key and Value Matrices

The output of embedding and positional encoding (denoted as X) is multiplied with three learnable parameter matrices, W_Q, W_K, and W_V, to create the Query (Q), Key (K) and Value (V) matrices [40]. Each of the three plays an integral role in the self-attention mechanism.

The analogy of names similar to those found in dictionary components is not by accident. To provide a crude parable illustrating the principles that make up self-attention, one can think of a Q as a question one can try and solve. K represents the names of potential answers' labels, and V provide actual answers. However, unlike dictionaries, which return a single value, this case is rather probabilistic.

In terms of dimensionality, W_Q, W_K and W_V are determined by the embedding size (d_{model}) and d_k and d_v, where d_k is the second dimension for K and Q and d_v for V. Both d_v and d_k represent a fraction of d_{model} and often are set to equal each other [40]. However, it does not have to be always the case [130].

The following operations ensue [40]:

$$Q = XW_Q \tag{3.6}$$

$$K = XW_K \tag{3.7}$$

$$V = XW_V \tag{3.8}$$

The Dot Product

Once all three are obtained, the Q and $K^\top$ dot product is calculated and serves as a measure of similarity between Q and K, i.e. how well a query matches a particular key. Then, to follow up with an example consistent with the parable, it is an educated guess as to how relevant specific answers will be concerning the query at hand based on their names.

It represents the strength of the relationship between a specific token and all the other ones in a given sequence, and is given by the following:

$$\mathrm{Q}K^\top \tag{3.9}$$

Scaled Dot Product

Subsequently, the dot product is scaled by division by a square root of the K's dimension as follows:

$$\frac{QK^\top}{\sqrt{d_k}} \tag{3.10}$$

This is because once the values are used as inputs of a softmax function, which is sensitive to high numbers that are more likely with larger embeddings, later in the model, such high numbers could decrease the gradient substantially and hinder down learning [130].

Masking

In the next step, a mask is applied to the scaled dot product so that for each token, only the preceding tokens in the input sequence are 'paid attention to' by the model. This is done by assigning a large negative value to the positions above the diagonal line in the attention score matrix. This means each position can pay attention only to itself and the tokens to the left of it, as the tokens to its right are masked, and thus the leftward flow of information is prevented [130]. Finally, the mask is added to the scaled dot-product matrix as follows:

$$\frac{QK^\top}{\sqrt{d_k}} + M \tag{3.11}$$

It is worth noting that whilst masking is optional, it is widely used in practice [130], not only for the sake of including the so-called look-ahead mask stopping the model from paying attention to future tokens but also due to padded input, which is often shorter than a given context window. Another mask can be thus created, distinguishing between tokens of sentences and padding tokens, and mask M can thereby be the result of element-wise addition, minimum, logical OR or a different transformation depending on the format chosen for the masks that would ensure that positions outside the sentence tokens' positions and future tokens were masked-out.

Softmax Transformation

Following the masking operation, the resultant matrix is passed through a softmax function so that values at hand reside between 0 to 1, forming a probabilistic distribution [130].

$$\text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + M\right) \tag{3.12}$$

As a consequence of the masking, masked-out positions after softmax would yield values equal to zero.

Attention

Attention is then obtained as by the multiplication of the resultant matrix by the Value matrix as follows:

$$\text{Attention}(Q, K, V, M) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + M\right)V \tag{3.13}$$

For a visual representation of the aforementioned steps involved in the scaled dot-product self-attention, refer to Figure 3.2.

Note that the complexity of attention is quadratic, as each token's relationship with every other in the sequence is calculated [90]. Quadratic complexity can make training larger models computationally expensive, which is one reason why this thesis works with the smallest of GPT-2 versions, GPT-2 small, and is considered one of the most significant bottlenecks of scaling-up Transformer-based models. Solutions tackling this predicament, such as sparse attention mechanisms, have been proposed [32]; however, a detailed discussion of their inner workings is beyond the scope of this thesis.

Multi-Head Attention

The description above refers to a scenario where a single head of self-attention was used, and parameters d_v and d_k would be equal to d_{model}. Since Transformers often have several attention heads, the GPT-2 version employed

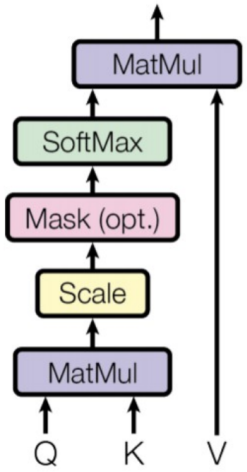


Figure 3.2: Scaled dot product [130]. As can be seen, first the dot product (MatMul) operation takes place, followed by scaling and addition of the mask. Subsequently the resultant matrix is transformed using the Softmax operation and finally multiplied with the V matrix.

used 12 in each of its 12 decoder blocks [42], this process happens in each head independently and simultaneously, showcasing that parallelisation is a significant advantage of this architecture, accelerating its training time.

After matrices Q, K and V are computed, they are partitioned into subparts[2] denoted as Q^i, K^i, and V^i, where i ranges from 1 to the number of self-attention heads.

Parallelisation entails partitioning each matrix by the number of self-attention heads along the d_{model} (embedding) dimension. Each new matrix has a size of d_{model}/h (where h denotes the number of attention heads) along the d_{model} (embedding) dimension [130]. Attention is calculated separately for each head, given each new matrix per head. This means that each attention head can focus on various positions at different representational subspaces of the input embeddings simultaneously, i.e. delve into different nuanced relationships between different permutations of tokens in the input sequence. The resulting attention matrices are then concatenated and multiplied with the parameter matrix W^O[3] as follows:

$$\text{Multihead}(Q, K, V) = \text{Concat}\left(\text{head}_1, \text{head}_2, ..., \text{head}_h\right) W^O \tag{3.14}$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V, M) \tag{3.15}$$

[2]In practice, the matrices are not explicitly partitioned into separate matrices. Instead, each attention head operates on a specific section of the matrix.

[3]According to [130], the matrix is square, dimensions are $hd_v \times d_{model}$, where d_v is set to d_{model} divided by the number of attention heads.

Where projection matrices W_i^Q, W_i^K and W_i^V are of dimensions $d_{model} \times d_k$, $d_{model} \times d_k$ and $d_{model} \times d_v$ respectively [130].

Ultimately, multi-head attention outputs a matrix of dimensions $context\_window \times d_{model}$, which encapsulates information from several representational subspaces. This ability enables the Transformer to learn different aspects of the relationships among tokens in the input sequence, leading to richer and more nuanced interpretations. For reference, Figure 3.3 depicts the scheme outlining multi-head self-attention.

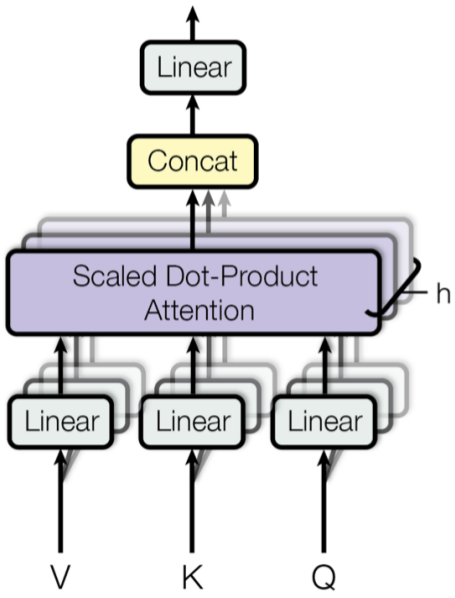


Figure 3.3: The Multi-Head Attention [130].

3.4.2 Residual Connections

Residual connections represent operations featured in Transformer models, which skip layers and work by adding to multi-head self-attention results the input initially passed to self-attention layers [55]. This method addresses vanishing gradients, as gradients often diminish in magnitude during backpropagation as their values come close to zero, thereby hampering the model's efficient learning. This alteration makes it easier for gradients to traverse deeper and shallower layers during backpropagation.

In practice, residual connections add the original input matrix X to the Multihead(Q, K, V) output of the multi-head self-attention, giving rise to a matrix of $context\_window \times d_{model}$, for explanation denoted as R [130].

An additional residual connection surrounds the feed-forward network (FFN), a component of the block explored later, which adds the input of the FFN to its output.

3.4.3 Layer Normalisation

As in many other neural networks, the magnitudes of outputs of each layer can become large, leading to unstable training of the entire model. To combat this issue and stabilise training, layer normalisation is applied. This entails calculating the mean and standard deviation for every token, i.e. for every row along the context window dimension of the matrix R. Each value in the row thus has the row-mean subtracted from it and is divided by the row values' standard deviation [11].

In mathematical terms, for each value r_{ij} within the R matrix, where i is a token number from 1 to the context window size and j an index corresponding to the embedding dimension, ranging from 1 to 768 in the model employed. The normalisation for each position is the following:

$$r_{ij} = \frac{r_{ij} - \mu_i}{\sigma_i} \tag{3.16}$$

where μ_i is the mean value of elements in the row pertaining to the token i, and σ_i is the corresponding standard deviation. Through this process, the normalisation centres the values in the matrix around zero, with a standard deviation of approximately 1 [103], thereby stabilising the model's training.

Layer normalisation is further enhanced with a linear projection step. Two parameters, γ and β, are subsequently introduced, which are learnable vectors of size d_{model}, thereby corresponding to each row of the matrix R.

Their existence enables the model to learn different standard deviations and means for every token. Once normalisation is finished, to acknowledge this transformation the matrix R will transformed into M, values of the output M will be:

$$m_{ij} = \gamma_i r_{ij} + \beta_i \tag{3.17}$$

In this equation, γ_i is the value at the i-th position of the γ vector, r_{ij} is the normalised value in row i and column j of R, and β_i is the value in the β vector at position i [82]. m_{ij} is the resulting normalised value at row i and column j of the matrix M.

Even though the layer normalisation is presented as a step following the residual connection after the first self-attention block, which is the case in the GPT model [102], the authors of the GPT-2 model adapted its placement at the input of every sub-block, i.e. before self-attention and following the residual connection, before the feed-forward network. Additionally, one layer normalisation was added after the last self-attention block. The changes in placements of the layers were to improve the model's scalability and

performance [103].

3.4.4 Position-Wise Feed-Forward Neural Network (FFN)

A salient feature of the GPT-2 model is the use of a position-wise Feed-Forward Neural Network (FFN) after the linear normalisation, which comprises two linear layers interspaced with a GELU (Gaussian Error Linear Unit) [56] activation function [103]. This step, unlike self-attention, is applied to each position separately and identically (hence position-wise), as conveyed by 3.18 [70].

$$\mathrm{FFN}(M) = \mathrm{GELU}(MW_1 + b_1)W_2 + b_2 \tag{3.18}$$

Here, M represents the entire matrix at the input of the FFN. W_1, b_1, W_2, and b_2 are two weight–bias pairs of learnable parameters of each of the two linear layers respectively.

The GELU activation function, which can be thought of as a smoother version of ReLU (Rectified Linear Unit) used in the original Transformer architecture [130], introduces nonlinearity into the model. In comparison to ReLU, it allows for easier interpretation of complex functions [56]. The GELU function for input z, is defined as:

$$\mathrm{GELU}(z) = zP(Z \leq z) = z\Phi(z) = z \cdot \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{z}{\sqrt{2}}\right)\right] \tag{3.19}$$

$$\text{where } Z \sim N(0, 1) \tag{3.20}$$

$\Phi(z)$ stands for the cumulative distribution function of the standard normal distribution [135]. As seen above, the distribution of Z values follows $N(0, 1)$, which was possible given the preceding normalisation operations.

The aforementioned represents the steps each block within GPT-2 consists of. Once the last block finishes its computations and outputs the resultant matrix, let us denote it F; it is then passed through a final linear transformation layer that outputs unnormalised logits as outputs [103].

In the case of the GPT-2 variant discussed, the matrix weights used to multiply the results in the last linear layer after the last decoder block are tied to the embedding matrix explained at the start of the Transformer [43]. The resulting matrix of dimensions corresponding to the sequence length and the size of the vocabulary represents the logits, which are unnormalised outputs of the model. Their values are then passed through a softmax function, normalising them into probability distributions. The resulting values represent probabilities for each possible next token in the sequence,

corresponding to each BPE-induced 'word' from the tokeniser's vocabulary, as across rows values sum up to 1.

Please note that the explanation provided abstracts from certain aspects of the model for the sake of clarity. GPT-2 includes several regularisations and intricate details. For example, the model scales the weights of residual layers at initialisation by a factor of $\frac{1}{\sqrt{N_r}}$, where N_r is the number of residual layers [103]. Other aspects, such as batching, regularisation or shifting labels one position to the right so that likelihood of a word being at the first position of the input sequence could be inferred, should not be omitted from the implementation of the model. However, for clarity and conciseness, they were omitted from this explanation.

3.5 GPT-2 vs Other Models

Given the elucidated intricacies of the GPT-2 variant of the Transformer model, it is worth reiterating why it was selected over different models and structures.

Numerous factors contributed to the choice of GPT-2. Transformers represent the current state-of-the-art architecture in NLP [130]. In terms of the GPT family, GPT-2 strikes an optimal balance between the model's size and its accessibility. Larger models such as GPT-3, albeit more powerful, are not freely available through APIs and require more computational resources. At the same time, GPT-2's predecessor, GPT, although trained on a smaller dataset, achieved the same or worse results as GPT-2 in experiments conducted for this thesis[4].

Another reason GPT-2 was selected over many others is that it supports causal learning, as given by its masking mechanism, it cannot 'peek ahead'. Unlike masked language modelling used in models such as BERT, it allows the model to learn about texts in a way that is more aligned with the relatively linear way humans learn and write, which is particularly suited for this thesis's objective of applying GPT-2 to social media post datasets.

Self-attention is one of the most significant advantages the Transformer architecture has over other deep learning models, such as RNNs. It allows the model to discern long-range dependencies and relationships between tokens concurrently, in contrast to RNNs' reliance on sequential processing. Since RNNs can only process a computation at a certain point in the input once all preceding tokens have been processed, its inference is considerably slowed.

[4]As a way of validating the steps taken during fine-tuning, most of the results presented within the thesis, which used a PyTorch script to fine-tune the GPT-2 model, another script employing TensorFlow was used in fine-tuning a GPT model variant. That is why this sanity check served as a reinforcement of results attained, as they were in some cases matched by the GPT model.

Another, perhaps an even more pertinent, predicament architectures such as RNNs have is that of vanishing gradients, which in the context of text processing means that by the end of a context window, the models may 'forget' what they analysed at its start [4]. In stark contrast to RNNs, Transformer architecture allows the self-attention mechanism to have access to all input tokens at once, as opposed to RNNs, which cannot revisit earlier tokens at the same time as the ones processed later. In addition, previously explained residual connections lessen the degree to which Transformer architecture-based models suffer from vanishing gradient issues.

Moreover, GPT-2 was trained on the WebText dataset, comprising 40 GB of data from over 8 million documents linked in Reddit posts with at least three upvotes [103]. Thanks to this vast dataset, the model can draw from the information it was trained on and then apply it by generating more contextually relevant and coherent texts.

One could argue, however, that CNNs (Convolutional Neural Networks), when applied to language modelling, can solve some of the issues RNNs suffer from with parallelisation. However, the predicament of long-range dependencies persists, as for tokens that are far apart, very large kernels would be needed to increase the model's computational complexity through the introduction of more layers [68].

However, Transformers have their limitations as well. They are generally very resource-heavy both in terms of computation and memory, especially when scaling up. In addition, the context-window length is, in some cases, a constraining hyperparameter and other issues, such as potential embedding biases or model interpretability [47], as it is hard to reverse-engineer algorithms learned by the model through the weights the model created, persist.

Chapter 4

Datasets

In natural language processing, it is seldom the case that data is prepared to be trained on right a way, but requires a series of pre-processing steps. This chapter elucidates the various facets of data preparation, encompassing data selection, cleaning, partition into training and validation sets, tokenisation and ultimately presents the evaluation metrics selected.

4.1 Data Selection

There was a clear intent behind the choices of datasets for the experiments within this thesis. Even though the primary focus was on assessing public opinions using Reddit comments, their heterogeneous nature made it challenging to discern whether the model accurately represents opinions within the dataset due to the absence of effective validation methods. Moreover, contradictory statements within Reddit discussions could introduce confounding variables to model verification. To mitigate these issues, it was essential first to test the GPT-2 model's inference abilities on domain-specific, easily verifiable texts less likely to contain conflicting information. Hence, two datasets were deliberately chosen for this task. The first comprised economic textbooks laden with expert-level economic vocabulary, and the other included articles from Investopedia.com [62], a source geared towards the general public. This two-step approach allowed for an initial examination of the model's alignment with datasets it was fine-tuned on, as well as its capability of comprehending varying levels of economic jargon and technical terms, prior to its application to Reddit data.

4.1.1 Textbook Dataset

The first corpus collected was created as an amalgamation of several economic textbooks [77] [28] [136] [53] [109] [84] [89] [51] [76] [133] [46] spanning curricula both on the undergraduate and graduate level. Texts that make up the dataset were carefully selected as a way of imbuing the model with

a comprehensive overview of economic concepts, theories, and terminology, thus ensuring the model has access to a wide range of explicit foundational knowledge of economics which might not have been sufficiently emphasised in the data it was trained on. Textbooks used span topics such as macroeconomic models, trade economics, development economics and even foundations of econometric theories and tests, thereby harbouring a diversity that could deepen the model's understanding of the domain. The dataset is particularly integral to the first experiment presented in the experimental chapter, concerned with whether a language model can grasp nuances engrained in economic terminology, thus reaffirming the hypothesis that applying language models to economic texts could draw out externally valid results. As conveyed in Figure 4.1, the dataset is filled with domain-specific terms, definitions and has an overall academic tone.

> *The basic premise underlying these models is utility maximization. The assumption that individuals make choices to maximize their well-being, subject to resource constraints, gives us a very powerful framework for creating tractable economic models and making clear predictions. In the context of consumption decisions, utility maximization leads to a set of demand equations. In a demand equation, the quantity demanded of each commodity depends on the price of the goods, the price of substitute and complementary goods, the consumer's income, and the individual's characteristics that affect taste. These equations can form the basis of an econometric analysis of consumer demand.*

Figure 4.1: An excerpt from the textbook dataset.

4.1.2 Investopedia Dataset

The second corpus comprises a compilation of economic and financial articles from Investopedia.com [62]. The dataset bridges the gap between academics and content capturing the public's interest. Investopedia contains various educational resources on finance and economics, such as articles on current events, practical applications of economic theories, and in-depth explanations of theories, which could provide the model with a broader understanding of economic topics and their context. It was selected as a juxtaposition to the format represented by the dataset comprised of textbooks, as it was posited that given the less convoluted and technical nature of articles from Investopedia, as can be observed from the excerpt in Figure 4.2, it would possibly prove easier for the model to align with a less formal dataset.

4.1.3 Reddit Datasets

The final corpus used is, in fact, a collection of smaller corpora extracted from comment sections on Reddit.com, covering every initial month in every quarter of the years 2020, 2021, and 2022, hence 12 in total, extracted from the API Pushshift.io [16].

he deposit multiplier is frequently confused with the money multiplier. Although the two terms are closely related, they are not interchangeable and are distinctly different. The money multiplier reflects the change in a nation's money supply created by the loan of capital beyond a bank's reserve. It can be seen as the maximum potential creation of money through the multiplied effect of all bank lending. Closing a credit card can impact your credit utilization ratio, potentially dinging your credit score. Credit utilization measures how much of your total available credit is being used, based on your credit reports.

Figure 4.2: An excerpt from the Investopedia dataset.

Several reasons informed the selection of Reddit as a source. First, as one of the most visited global social media platforms with a growing user base, Reddit offers a large and diverse set of discursive conversational text spanning numerous topics [105]. Unlike platforms like Twitter or Facebook, Reddit has a more relaxed data retrieval policy giving rise to APIs such as Pushshift.io, which is concerned with scraping data from social media platforms from which data was obtained [16]. Nearly 50% of Reddit users are based in the US [21]; consequently, when comparing opinions on topics such as unemployment or interest rate with actual values, the prevalence of US comments facilitates an easier match between inferred opinions and real-world data. Additionally, unlike many other datasets, Reddit data is ordered by time, an integral premise of the experiments comparing public opinion to economic indicators at specific time frames.

The time frame from 2020 to 2022 was selected due to the gradual expansion of Reddit's user base, meaning more recent crawls tend to be larger, and in part due to the COVID-19 pandemic, which provided a global external shock to economies [1] and as per our hypotheses, an opportunity to observe more easily discernible shifts in opinions. Given the focus on economic terms, often reported quarterly, only datasets corresponding to every initial month of every quarter (i.e. January, April, July and October) were extracted. In addition, due to computational constraints, the Reddit datasets include only comments referring to the economic keywords interest rate, inflation and unemployment.

Due to the idiosyncratic nature of the volume of data posted each month on Reddit, discrepancies can arise concerning frequencies of words across datasets. That is why the Reddit datasets underwent a normalization process, consisting of randomly selecting and discarding comments in larger datasets, ensuring each was approximately the same size, with a small tolerance of less than 0.3%.

Table 4.1 shows the total word count and vocabulary size for each dataset, with average figures provided for the 12 Reddit datasets. Note that in this

You answered OP's question and solved an upcoming problem. But I do have a simple counter-argument to you: inflation, especially more moderate, is always a signifier of someone posessing more money than before and spending them; in other words it is a signifier of a redistribution of resources in an economy. As the months roll on I'm more and more pleased with their style of was those damn liberal California winds! If people just want the federal government to enforce demonizing an entire industry that employs and provides for almost 2 million (direct and indirect) people I'm very much against. I work full time so I'm able to afford it thankfully.

Figure 4.3: An excerpt from the Reddit datasets, corresponding to January 2020. As can be observed, the tone and vocabulary is different to those in textbook or Investopedia excerpts.

context, a 'word' refers to a string separated by punctuation or spaces, rather than the conventional morphological definition of a word.

	Textbook	**Investopedia**	**Reddit (Mean)**
Train file word count	2,638,411	9,133,694	37,162,515
Train file vocab size	104,993	218,575	918,650
Validation file word count	139,516	91,142	768,232
Validation file vocab size	19,843	15,435	69,960

Table 4.1: Dataset Characteristics.

As shown in Table 4.1, the datasets were split into training and validation sets. The size of the validation sets was determined empirically, by monitoring the validation loss metric during preliminary tests to identify the optimal training-validation split. Given the extensive corpus on which GPT-2 was initially trained, the model performed effectively even with smaller validation sets.

4.2 Data Cleaning

Regarding data scraped from social media, as in this case, one ought to be wary of the substantial amount of noise it can contain. Data scraped from social media platforms are often littered with HTML tags, URLs, missing punctuation, grammatical errors, emojis, slang, and other elements that should be addressed before fine-tuning [118].

Filtering and cleaning data can significantly enhance the speed and accuracy of trained models, making it an indispensable part of pre-processing [52]. Therefore, several steps and heuristics were adopted to balance data quantity

and quality. Since the Reddit dataset required the most cleaning, the following text will refer merely to Reddit data cleaning, as it subsumes all steps cleaning of other datasets entailed.

4.2.1 Bots

Bots are applications often operating on public forums and social media platforms, performing automated assignments often with ulterior motives. They have been associated with the spread of fake news, theft of personal data, attempts to skew public opinion on issues such as COVID, and even election tampering [69]. Given that the datasets used were compiled to assess public opinion on topics of economic and political importance, it is crucial to realise the extent to which bot activity could have contaminated the data.

An issue with datasets from Reddit, as with other social media platforms, is the presence of comments generated by bots and how to remove them systematically. There are patterns to look out for, such as specific mistakes in punctuation or repetitiveness [85]. Unfortunately, whilst for the ordinary user of Reddit, random alphanumeric values for username, an oddly generic profile picture, a low profile rating or the content of bot-generated comments could be easy tell-tale signs to look out for; without the interface and only automatic bot detection tools, pieces of this vital information are not available.

Bots often exhibit other hallmark traits, such as being created simultaneously, exhibiting constant online activity, and reacting in groups to each other's or others' submissions and posts. In addition, other revealing metadata such as similar names, profile pictures or use of hashtags can also be a crucial clue in detecting artificially created posts [61].

When it comes to reproducing identical content, its impact on the datasets used was mitigated through deduplication. However, bots often generate merely semantically similar comments, thereby diminishing the effectiveness of deduplication. Additionally, bots frequently post links to third-party websites, which are irrelevant to the inference datasets were used for. The effect of this is abated with regular expressions removing URLs. The prevalence of bot-generated comments was also lessened through the limitation of using only comments from authors whose profiles were at least two weeks old at the time of each comment's publish date. Manual inspection revealed that a large share of accounts posting spam were only a few days or weeks old. Therefore, a heuristic was implemented, which filtered-out content written by accounts created less than two weeks before comments at hand were posted.

The number of sophisticated bots that exemplify evasive techniques is on the rise and, in some shape or form, affects every social media platform [85]. However, whilst systematic steps addressing their impact can be taken, thereby diminishing the probability of them negatively affecting datasets, the complete removal of their influence is an elusive task.

4.2.2 Named Entities

Another ramification of the use of data from social media is the potential inclusion of sensitive personal information, the misuse of which could be unethical. Several measures were taken to mitigate such possibility. First, only sentences pertaining to keywords were used in the model's testing, and no queries about a particular entity or a non-governmental figure were employed.

Second, a script utilising EntityRecognizer, a transition-based named entity recognition component by spaCY API [58], was used to detect potential named entities within the dataset. On a sample of randomly sampled 2500 lines of the January 2020 train set, there were no signs of personal information or names of entities that in some shape or form would not be connected to the public domain. However, it should be noted that such automated tools have their limitations and suffer from both false positive and negative detections. For instance, Lockheed Martin, a Maltese bank, and the word 'Pokémon' were mistakenly recognised as people.

Third, the administrator of the Pushshift API cleans sensitive information upon request from the users who posted it daily [111].

Considering the negligible likelihood of encountering sensitive personal information within the extracted datasets, combined with the focus of experiments on macroeconomic issues which are not concerned with individuals or corporations, and the fact that datasets are not publicly available, it is posited that all necessary steps were taken to handle scraped data ethically.

4.2.3 HTML Tags, Emojis and Special Characters

Regular expressions were used to replace HTML tags such as '<', '>' and '&' with their corresponding characters '<', '>' and '&'. In addition, URL links and tags representing them (i.e.'[link]' segments of text), emojis, special and UNICODE characters which are of no consequence to prompts used in experiments, were removed using regular expressions.

4.2.4 Markdown Syntax

Reddit uses a special markdown syntax, a lightweight language that adds formatting such as bulleted lists, italicisation, background colours or headings to text-based data [104], an example of which is conveyed in Figure 4.4. The syntax proved to be a challenge when cleaning data in two regards.

On the one hand, the use of markdown syntax can prove to be problematic during tokenisation, as to make certain parts of words bold, the syntax dictates for the said word to be written as a sequence of letters and hashtags or underscores or asterisks, thus not fitting into the vocabulary of the pre-trained model. On the other hand, eliminating such symbols could detract from the meaning authors gave to specific parts of their comments.

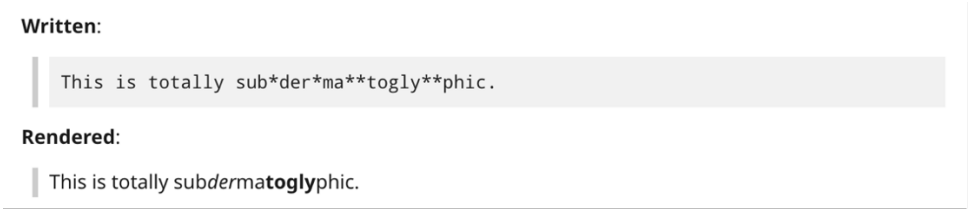


Figure 4.4: Example of the use use of markdown syntax in Reddit showing how added symbols change the visual aspect of text presented on the platform [104].

To enable the model to process and understand texts at hand efficiently and for the sake of consistency with other datasets, markdown syntax elements such as chained asterisks, underscores and other formatting text along with hashtags were removed, as they did not greatly benefit the flow of the text and despite highlighting specific portions of the text at hand, the message was comprehensible even without them.

At the same time, it is essential to note that in future experiments, applying tokenisers accustomed to markdown language processing, such as [83], is worth considering, as markdown syntax can contribute towards improved understanding and add a semantic dimension to texts at hand. However, given the early stages of experiments undertaken within this thesis and the fact the BPE tokeniser was not adapted to handle visually-rich texts containing markdown syntax, for the sake of clarity and practicality, markdown syntax was left out.

4.2.5 Sentence Tokenisation

Not all textual content was easily segmentable sentences, prompting the deployment of several heuristics to improve sentence detection in the texts under consideration. For example, instances in which a lowercase letter immediately followed a capital letter were, except for academic titles, deemed likely to be the result of an oversight by the author or data retrieval process. In such cases, a period and a space were inserted between the two letters, thereby enabling the formation of two separate sentences.

Subsequently, a sentence tokeniser *sent_tokenize* from the NLTK library [22] was employed to further distinguish between sentences, as either due to the informal character of the comments or scraping itself; in many cases, it was difficult to discern where sentences started and ended, as many lacked a space separating them. For illustration, an example of this particularity is depicted in Figure 4.5.

4.2.6 Different Languages

As a global platform, Reddit contains comments and submissions in several languages. Therefore, an English language detector was applied after the

. . . led to believe even a smaller amount of reclaim on there is pretty reasonable of a dose. I do also have 0 edible toleranceDo you have any advice for us CowbellyRedditers?Actually, I'm sick and tired of all the 'hard' champions. Every other jungler is a Lee Sin . . .

Figure 4.5: An excerpt from the January 2020 Reddit datasets, showing the lack of sentence delimeters in Reddit data.

forenamed cleaning steps to avoid noise in the data making the comments seem non-English. Two publicly available libraries for language detection were employed [117] [58], however, neither had a 100% success rate at detecting non-English texts, as sometimes, they suffered from false negative classifications, classifying particularly shorter sentences in English, as foreign. Furthermore, due to bilingual posters and parsing-elicited mistakes, some sentences were a blend of multiple languages, which was a particularly hard issue to solve automatically. Some of these mistakes were cleaned manually, but some were left within the datasets, as given their low numbers in comparison to the Reddit datasets' sizes (less than 0.3% of the text processed, as revealed through a manual test of a representative sample of 200 comments), the probability of them having a significant impact on inference was deemed as minimal.

4.2.7 Added Vocabulary

The dataset was then deduplicated to avoid repeating comments. Subsequently, since the original GPT-2 vocabulary does not account for sentence separators [42], the end-of-sentence delimiter '<|eos|>' was added to the tokeniser's vocabulary and used to separate sentences. The GPT-2 model by default does not use tokens to delineate sentences, as its original purpose is to generate text, which is a feature this thesis is not using. In order for the model to understand structures of sentences better, which was deemed important given prompts it was trained to attribute perplexities to, sentence separation was key.

4.3 Metrics

Creating metrics that would weigh the models' capabilities of understanding texts it was fine-tuned on was one of the research's focuses in this thesis. Realising the extent to which the GPT-2 model understood datasets could not be solely estimated through conventional metrics such as validation loss, as these do not offer practical insights into the model's capabilities when producing texts or producing perplexities of prompts given to them.

At the same time, merely comparing perplexity values of sentences can prove ineffective at capturing trends of opinions, as different formulations lead to different perplexities. For example, to see what a language model thinks about the weather, two sentences could be created in the following

formats: 'It is a sunny day.' and 'T'is a gloomy day riddled with cumuli indicative of impending showers.'. Trying to see which one is more likely merely through their perplexities would be inefficient, assuming the latter would elicit a higher perplexity score due to its intricate formulation alone, even if the text leaned more towards a cloudy day.

That is why perplexity-based metrics focusing on normalising the perplexities of sentences through their ratios rather than nominal values were created and explained within this section. Furthermore, Cross-Entropy Loss used to calculate perplexities, and the Economic Accuracy metric used to validate the model's understanding of verifiable economic texts is presented.

4.3.1 Cross-Entropy Loss

The GPT-2 model uses the cross-entropy loss during backpropagation [42], which serves as a metric used in experiments throughout this thesis to evaluate the perplexity of the model on several prompts and as a signal of model's learning in the form of validation loss during training. It measures the difference between two probability distributions—in this case, the predicted probabilities for each token in the context window and the actual target distribution of tokens. The loss is zero if the two distributions are identical, and the more the predicted probability of a token diverges from its label, the greater the cross-entropy becomes.

The loss is applied to token probability values obtained in the last layer of the Transformer. The formula for the loss is the following [65]:

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) \tag{4.1}$$

Where N is the number of tokens in the context window, y_i is the ground truth, i.e. the actual outcome of token i, p_i is the estimated probability of said token i as predicted by the Transformer. The summation is over all tokens in the context window.

Cross-entropy loss is calculated at the end of each epoch for training and validation datasets and indicates the model's performance. It can be subsequently used to calculate perplexity, as follows [65]:

$$PPL = \exp(L) \tag{4.2}$$

4.3.2 Perplexity Ratio

The perplexity ratio is used throughout this thesis as the ratio of perplexities of sentences with opposite semantic meanings. Customarily, the ratio is

between the sentence with negative connotations to the perplexity of its opposite as follows:

$$\text{Perplexity Ratio} = \frac{\text{Perplexity of a Negative Sentence}}{\text{Perplexity of a Positive Sentence}} \tag{4.3}$$

If the ratio equals one, perplexities are identical, and the model can predict either sentence equally well. If the ratio increases above 1, however, the perplexity of the sentence in the numerator is higher than that of the one in the denominator, meaning the positive sentence is more aligned with the model's predictions. The converse applies with it decreasing below 1, signalling that the negative sentence has a lower perplexity and is more in line with the model's predictions.

4.3.3 Economic Accuracy

Another perplexity-based metric used is economic accuracy. It is similar to the perplexity ratio in that it compares the perplexities of two opposing sentences. In the context of several pairs of opposing sentences, the economic accuracy measures the percentage of sentence pairs, where the model gave the correct one the lower perplexity.

Let us consider a hypothetical example with hypothetical values. We are given two pairs of sentences (a_1, a_2, b_1, b_2), where odd-numbered ones are the true ones, such as:

a_1 = GDP measures a country's economic output.
a_2 = GDP measures a country's economic input.
b_1 = A recession is two consecutive quarters of negative GDP growth.
b_2 = A recession is five consecutive quarters of negative GDP growth.

Let us assume that these are the sentences' perplexities elicited by the hypothetical model:

$$a_1 = \text{perplexity} = 15.44$$
$$a_2 = \text{perplexity} = 45.62$$
$$b_1 = \text{perplexity} = 77.3$$
$$b_2 = \text{perplexity} = 65.1$$

Since the model attributed a lower perplexity to the true statement in the case of the first pair, its prediction is correct. In the second one, however, the converse is true. This means that the economic accuracy of the model is 50%.

The selected datasets for fine-tuning represent various facets of resources harbouring economic knowledge. From clearly verifiable economic statements

and academic texts in textbooks to the less formal yet informative Investopedia entries and, ultimately, the public discourse captured in Reddit comments, these sources embody vital avenues through which the model's economic comprehension can be examined. Moreover, given the lack of annotated data or other means of external validation of public opinion, perplexity and its application in perplexity ratios and economic accuracy metrics provide powerful ways of quantifying the model's aptitude for aligning with the training datasets and grasping economic concepts.

Chapter 5

Experiments

This chapter first presents the way hyperparameters were selected. The second part is divided into two sections, each serving a different training purpose.

The first section of the second part consists of fine-tuning the model on economic datasets, namely the Investopedia and textbook datasets, which in part served as a benchmark for validation of results of economic meaning.

The second and final section is devoted to applying the model to the Reddit datasets, extrapolating public opinions on economic variables across time, and comparing results to survey data.

5.1 Hyperparameter Search

Hyperparameter search is an essential part of fine-tuning which enables the model to adapt to the given dataset, leading to improved results. Prior to undertaking experiments, hyperparameters such as batch size, learning rate (and schedulers of learning rate), regularisation methods (such as weight decay and dropout), number of epochs and different training termination conditions were evaluated for each dataset.

The hyperparameter search method adopted blended informed iterative grid search and random search [18]. In the iterative grid search method, every combination of hyperparameters' values in a pre-defined hyperparameter search space determined by values used in relevant literature was tested. In case of an optimal value being at the extremes of pre-determined ranges, said ranges were changed so that the previous best value was no longer an extreme and the search restarts with an expanded domain. Random search, which entails random sampling of hyperparameters from a bigger pre-defined hyperparameter space and determining which combinations perform the best given a specified metric [18], was then undertaken on hyperparameter

combinations not included in the iterative grid search method, which provided a finer granularity of hyperparameter testing.

5.1.1 Learning Rate

One of the critical hyperparameters affecting the model's convergence speed is the learning rate.

The learning rate governs the size of the optimiser's step while descending the gradient. A too-small learning rate may decelerate learning and cause the model to get stuck on local minima. However, a too-large one can lead to overshooting the minimum and, thereby, inconsistent results. Unfortunately, the original paper introducing GPT-2 does not explicitly state what learning rates were used, except for mentioning that they were selected manually [103].

The search space selected comprises values consistent with the literature [42] [81] and consists of 1e-4, 5e-5 and 5e-6.

5.1.2 Epochs

The decisive factor informing how many times the model will process the entire training dataset, the number of epochs used, was determined by considering the degree of convergence of validation loss during preliminary testing. Implementing three variants of stopping obviated the need for an extensive experimental search for the optimal number of epochs for each dataset. The stopping strategies tested included:

- **Early stopping with patience:** Training was halted after several consecutive epochs where the validation loss did not decrease. In this case, patience was manually set to 3.
- **Fixed number of epochs:** The number of epochs was set as an informed guess given preliminary runs.
- **Learning rate reduction with early stopping:** This approach integrated learning rate scheduling with early stopping. If the validation loss stopped decreasing, the learning rate was halved till the end of training. The end of the training was triggered when the validation loss stopped decreasing despite the reduced learning rate for more than two epochs.

Experimentally, the strategy which presented the best results given training time was learning rate reduction with early stopping and is the method used throughout the experiments. Whilst the early stopping with patience provided similar results, the lack of adjustment of the learning rate meant that the validation loss, metric governing the decision between the stopping criteria, did not decrease as much as with the third approach.

5.1.3 Batch Size

The batch size, i.e., the number of samples used in one iteration before each weight update, is a vital hyperparameter to consider. While larger batch sizes can produce more accurate gradient estimations, they may be strenuous both time and memory-wise. The batch size had to be kept low, given preliminary tests revealing that the architecture used did not provide enough resources for larger batches. The search space was thus limited to 4, 8 and 16, which were used in many applications of the GPT-2 small model [43].

5.1.4 Regularisations

Regularisation techniques implemented to counteract overfitting were weight decay and dropout.

Weight Decay

Weight decay is a technique applied to the weights in the Transformer by adding a small penalty function to the original loss, which is usually the L2 norm of the weights, encouraging the model to keep the weights on the lower side, preventing the model from fitting the data at hand too closely.

Despite the publication introducing GPT-2 omitting whether weight decay was used, in the paper on the GPT model [102] and in other applications [101], a weight decay of 0.01 was used.

The search space of weight decay thus consisted of 0.1, 0.01, and 0.001.

Dropout

Dropout is another way of mitigating over-fitting. The dropout value is the probability of any random neuron being disabled during training. This allows the Transformer to generalise and not over-rely on specific neurons. Dropout values experimented with were 0.1, the default setting given the Hugging Face library used [42], 0.3, and 0.05.

In the end, there is an equivalent of 81 combinations of hyperparameters to consider, which is not a significant amount, as the space of hyperparameters could further be expanded, but is expected to give a semblance of an insight as to which hyperparameters' values could improve training. This search was supplemented with 20 random sample searches, exploring hyperparameter values within, above, and below the previously specified ranges.

Other parameters, such as mixed-precision, warm up and learning rate schedulers, were experimentally tested but did not lead to consistent or significant improvements in validation loss.

The Table 5.1 contains the parameter combinations which achieved the

lowest validation losses on a dummy excerpt of the January 2020 Reddit dataset. The dataset is substantially smaller than the one used later during the experiments, as it comprised 800 000 words; however, it was considered significant enough to be representative.

Learning Rate	Batch Size	Weight Decay	Dropout	Validation Loss
5e-5	4	0.1	0.05	2.7680
1e-4	4	0.01	0.05	2.7690
5e-5	4	0.1	0.3	2.7690
5e-5	8	0.1	0.1	2.7690
5e-5	4	0.01	0.05	2.7700
5e-5	8	1e-3	0.1	2.7710
1e-4	4	0.1	0.3	2.7730
1e-4	4	1e-3	0.1	2.7740
1e-4	4	0.1	0.05	2.7750
1e-4	4	1e-3	0.3	2.7770

Table 5.1: Table showing results from hyperparameter search on the dummy Reddit dataset - as can be seen, in spite of various hyperparameter values, results were very similar.

As can be seen in Table 5.1, the performance of several hyperparameters led to nearly identical results. This was in many ways a consolation; due to the scaling up from the selected representative dataset to larger ones, GPU constraints required substitutions of some hyperparameters in certain cases for ones less demanding in terms of computational resources. Utilising random search was instrumental in adjusting the values of hyperparameters used in the end. Finally, the hyperparameters selected for each dataset are in Table 5.2.

	Learning rate	Batch size	Weight decay	Dropout
Economic sentences	1e-4	4	0	0.1
Textbook dataset	1e-4	4	0	0.1
Investopedia	1e-4	4	0.01	0.1
Reddit datasets	5e-5	2	0.01	0.1

Table 5.2: Default Hyperparameters.

5.2 Fine-Tuning the GPT-2 Model on Economic Texts

The objective of the first experiment was to assess the model's capacity to comprehend economic concepts within the textbook and Investopedia datasets and to ascertain how representative the model's results could be of the dataset on which it was trained. This understanding is crucial for the second experiment, where the model's alignment with texts is harder to

examine due to more variability and overall complexity of data comprising opinions instead of definitions, as well as for evaluating how would GPT-2 be able to grasp technical economic contexts.

For this experiment, two sets of opposing statements were drafted, one from the Investopedia dataset, and the other from the textbook dataset. Each of the sets contained sentence pairs, in which one of the sentences was a paraphrase of a definition or a general statement from within the text, and the second was its negation.

A crucial aspect to consider was how the model understands underlying economic principles laid out in the datasets, which was evaluated using the economic accuracy metric. It was posited that a high score could suggest that the model can apply its acquired knowledge to a broader range of economic situations. Conversely, a low score could indicate that the model harbours inherent biases, favouring particular types of statements and views, or is incapable of generalising within the economic domain.

Furthermore, including the textbook and Investopedia datasets allowed for a more balanced perspective, as economic textbooks provide formal and didactic content. In contrast, Investopedia features reworded definitions and articles designated for people with non-economic backgrounds. This approach was a crucial precursor to training the model on data characterised by heterogeneous and informal economic content.

The statement drafting process and illustrative examples proceeded as follows.

1. Selection of a statement presenting an economic fact or a definition from within each dataset [1].
    - *'The policymaker can expand aggregate demand to lower unemployment and raise inflation.'*
2. Reformulation of the statement so it would not follow the same structure as the original and creating an opposing statement.
    - *'Increasing aggregate demand can increase inflation and have a downward effect on unemployment.'*
    - *'Decreasing aggregate demand can increase inflation and have a downward effect on unemployment.'*
3. Evaluation of the sentences' perplexities after each training epoch, with economic accuracy (EA) representing the percentage share of true statements with lower perplexities.

[1] This process was done under the supervision of economic experts, in order to select statements which truly represent true economic principles and definitions.

To view a larger sample of statements used to calculate economic accuracy, please refer to Appendix B.1 and Appendix B.2 for the textbook and Investopedia datasets, respectively. The hyperparameters used for this experiment were the ones mentioned in Table 5.2.

5.3 Reddit Experiments

The final series of experiments pertained to fine-tuning the GPT-2 model on datasets extracted from Reddit comment sections from the initial months of each calendar quarter from 2020 to 2022.

The purpose of this experiment was to address the primary motivation of this thesis, which was to ascertain whether public opinion can be extrapolated from social media data through a fine-tuned GPT-2 model and to see whether inference changes over time, possibly due to changes in public opinions.

The section was split into four parts:

1. **Perplexity ratios of economic variables:** In the first part, attention was focused on analysing how perplexity ratios of sentences pertaining to three economic variables differ across time (datasets), thereby gaining an insight into how well the model captures public opinions on these variables.

2. **Robustness of perplexity ratios:** The second subsection investigated the consistency of findings from the first subsection, examining whether the trends of perplexity ratios remained stable when averaged across several paraphrases of the sentences used and how they compared to actual economic variables.

3. **Short-term predictions:** The third subsection examined predictions in a short-term context and compared them to real-life values of economic indicators they were designed to predict and predictions mentioned in the Survey of Professional Forecasters.

4. **Long-term predictions:** The final subsection focused on one-year-ahead predictions, comparing the results not only with actual values of economic variables but also with predictions from the Survey of Professional Forecasters and the Michigan Survey of Consumers.

5.3.1 Perplexity Ratios of Economic Variables

The approach to this experiment, in many ways, resembles that of the preceding one. A common sentence structure was devised as shown in Figure 5.1.

As conveyed by the sentence structure, the selected economic variables

'In the economy, [economic variable] is [low/high].'

Figure 5.1: Template for constructing sentences about economic variables, where economic variable was one of 'interest rate', 'inflation' or 'unemployment'.

were interest rate, inflation, and unemployment. Their selection, similarly to the design of the template sentence structure, was consulted with an expert in the field of economics. The rationale behind this choice was threefold.

First, given the scope of this thesis, it was warranted to focus on a selected few variables in detail as opposed to considering a multitude of terms eliciting heterogeneous opinions amongst Reddit users, when choosing which aspects of economic opinions held by the public to focus on.

Second, the three variables – interest rate, unemployment, and inflation - are of particular importance to the public. This is in contrast to other economic terms such as tariffs or repo rates, which may not be as fervently discussed, as the public may be less sensitive to them. Interest rates are of great importance not merely because of the price of borrowing they determine, but also given the negative relationship between interest rates and wages, thereby affecting a broad swath of the population. The importance of unemployment is straightforward; especially in dire socioeconomic situations, individuals may harbour fears concerning job security and may discuss them online. Lastly, inflation determines the rate of increase in prices, a critical concern for households as rising inflation erodes their real incomes.

Third, the three variables were, to some extent, subjects of both the Survey of Professional Forecasters and the Michigan Survey of Consumers, thereby lending these two sources of opinions as suitable tools for comparisons further explored in the third subsection of this experiment.

For these reasons it was hypothesised that the selected variables would be more prevalent in public discourse and thus more suitable for the presented research.

For each of the three variables, a sentence including the word 'low' and another with 'high' were created as per the template in Figure 5.1. Subsequently, the ratio of their perplexities was measured across the Reddit-trained models.

The models used in this section were then used in the upcoming sections within the second experiment. Hyperparameters used throughout were specified in Table 5.2.

5.3.2 Robustness of Perplexity Ratios

The second part of the experiment followed the structure of the first, except for reformulating the sentences used and working with aggregate perplexity ratios

corresponding to averages of semantically similar sentences and comparing the trends in their perplexity ratios with actual interest rate, inflation and unemployment values over time.

For each sentence containing one of the three economic variables used constructed according to the Figure 5.1 template, three new versions were created; hence four versions of the same overall were tested, and their perplexities were subsequently averaged, as conveyed in Figure 5.2.

'In the economy we can see that [economic variable] is [low/high].'
'When it comes to the economy, [economic variable] is [low/high].'
'With respect to the economy, the [economic variable] is quite [low/high].'

Figure 5.2: Templates for constructing paraphrases of sentences about economic variables, where economic variable was one of 'interest rate', 'inflation' or 'unemployment'.

This means that when considering the keyword *unemployment*, for instance, a sentence according to the template in Figure 5.1 was created: 'In the economy, *unemployment* is *high*.' Then three additional paraphrases were created according to the respective templates in Figure 5.2 as follows: 'In the economy we can see that *unemployment* is *high*.', 'When it comes to the economy, *unemployment* is *high*.' and finally, 'With respect to the economy, the *unemployment* is quite *high*.'

These sentences' perplexities were then measured. Subsequently, in order to create their respective opposing statements, the word *high* was replaced in each of them with *low* and these sentences' perplexities were also measured.

Finally, perplexities for the two sets of paraphrases were averaged and perplexity ratio was calculated, as per equation 5.1, which is in line with the explanation of the perplexity ratio in section 4.3.

$$\text{Perplexity ratio} = \frac{\text{Mean perplexity of paraphrases containing '}low\text{'}}{\text{Mean perplexity of paraphrases containing '}high\text{'}} \quad (5.1)$$

The reasons for this kind of experimenting were to an extent due to the lack of external validation mechanism, as to whether perplexities attributed to the data the model was trained on were true indication of the model's alignment with the content of the dataset. In order not to overestimate the value of each sentence tested, four semantically very similar sentences, paraphrases, served as a more robust estimation of the model's capabilities. This experiment was designed specifically to examine the extent to which the model was sensitive to rephrasing, so that inference would not overstate its capabilities judging by cherry-picked sentences.

In the end, trends obtained from the perplexity ratios of paraphrases were

compared to trends in each of the three economic variables respectively, over the course of 2020, 2021 and 2022.

The values for the economic indicators that perplexity ratios were compared against were extracted from FRED [96], the Federal Reserve Economic Data, a comprehensive database by the Federal Reserve Bank of St. Louis. FRED provides access to economic data from various sources and is widely used for the analysis of economic trends and indicators. The variables focused on were interest rate, inflation, and unemployment rate. Unfortunately, not all three are conveyed by economic indicators directly, as some are approximated by proxy measures. The statistics used as proxies for economic variables were the following:

1. *Unemployment rate* as a measure of unemployment, defined and measured as the percentage of the entire labour force actively seeking employment yet remaining unemployed. The data is released monthly by the Bureau of Labour Statistics and was averaged for every three months in a year to obtain quarterly data and seasonally adjusted to account for periodic changes to unemployment caused by seasonal variations [129].

2. $10-year\ Treasury\ bonds$ served as a proxy for interest rates. Treasury bonds are debt securities issued by the US Department of Treasury with a 10-year maturity when the face value of the bond is repaid. This metric is commonly used in economics as an indication of the sentiments of investors about the future and is widely used as a measure of interest rates. The values of the metric are updated daily based on their current values in the secondary trading market and are published quarterly [24].

3. The change over the past year in the *Consumer Price Index* (*CPI*) was employed as a measure of inflation. Since there are several measures of changes in price, CPI was chosen as it is one of the most used ones in economic analyses. CPI measures the average change in the prices paid by urban consumers for a specific basket of goods and services over time. As a substantial overlap is assumed between urban consumers and Reddit users, the metric aptly reflects the demographic whose opinions were analysed. The value of CPI is published by the Bureau of Labor Statistics monthly [96] [2].

The aim of this comparison was to see, whether there was a possibility of the metric designed to weigh opinions on whether economic variables are on the rise, or the decline, could show similar trends to those exemplified by economic variables paraphrases pertained to. This is another form of external verification, anchoring the results off real-world data.

[2]The author is deeply grateful for the guidance received from Doc. RNDr. Filip Matějka, Ph.D., Associate Professor of economics with Tenure at CERGE, Charles University, Prague, and his team. Their expertise significantly contributed to the selection of the appropriate metrics and other economic aspects of this research.

5.3.3 Short-Term Predictions

In this part of the experiment, the aim was to gauge the accuracy of the paraphrases' perplexity ratio trends in comparison to actual real-world economic indicators they forecast, as well as the Survey of Professional Forecasters' results over the course of quarters spanning from 2020 to 2022. Whilst the previous part of the experiment focused solely on comparisons with real economic variables, this section explores the extent to which trends in perplexities align with those of economic variables better than short-term predictions by professionals.

The source of data representing opinions of professionals was the Survey of Professional Forecasters (hereinafter also referred to as SPF) [44], one of the most renowned and oldest quarterly survey of macroeconomic indicators in the US. It periodically gathers predictions on economic variables from professionals both in the long-term and short-term, addressing economists, consultants, and academic institutions for their predictions [44]. The values presented are means of predicted values provided by survey respondents with respect to the unemployment rate, 10-year Treasury bond returns which serve as a proxy for interest rate, and core CPI percentual change, which is a measure of inflation based on the CPI index adjusted to represent only goods and services which values are not volatile, such as those of food, fuel prices or financial investments [89]. It was postulated that public opinions are not as sensitive to anticipate seasonal effects on economic metrics and therefore using code CPI as opposed to a more volatile headline CPI was deemed more appropriate.

The shortest forecasts featured in the SPF relate to the same quarter in which the survey occurs. Similarly, Reddit datasets were based on a snapshot of public opinions from the first month of every quarter. Consequently, neither the professional forecasters nor the authors of the Reddit comments used were aware of the actual values of the economic variables they were discussing when alluding to the present state of the economy. This method allows for an effective comparison between the perplexity ratios of paraphrases and professionals' predictions for the current quarter's unemployment, inflation, and interest rate values. These short-term predictions serve to demonstrate whether Reddit comments are susceptible to external factors that can affect economic variables in a similar way professionals can be and whether their predictions' trends could be compared to those of professionals. Through this comparison, the question of whether Reddit posters' opinions on the state of economic variables align with those of experts was addressed.

5.3.4 Long-Term Predictions

The last part of the experiment altered the tense of the template sentence in Figure 5.1 and of paraphrases used in the previous experiment sections in Figure 5.2 from present to future, as follows:

'In the economy we can see that [economic variable] will be [low/high].'
'When it comes to the economy, [economic variable] will be [low/high].'
'With respect to the economy, the [economic variable] will be quite [low/high].'

Figure 5.3: Templates for constructing sentences about economic variables in the future, where economic variable was one of 'interest rate', 'inflation' or 'unemployment'.

Initially, the aim of the part of the experiment was to assess whether the model could be specific in its predictions of time and change in quantity of economic variables one year in the future. However, preliminary results had shown that perplexity ratio trends were very similar across different formulations referring to various points in future, such as months, years and even decades and percentual changes. An illustration of the model's insensitivity to changes in specific numerical quantities of economic variables and specific points in time is shown in subsections C.1.1 and C.1.2 of the Appendix. The reasons why the model was unable to refine its estimates to a period one year in the future was attributed to the data fine-tuning was undertaken for not including enough temporal details, as upon manual inspection, it was revealed that in some of the datasets, less than 100 sentences were referring to a point 12 months (or the equivalent of a year) in the future. Out of these, even fewer were semantically similar to the sentences used for perplexity ratio calculations, as shown in Figure 5.3.

For the sake of comparing future predictions, however, in the end, sentences used referred to the future in general and did not specify a time frame, as seen in Figure 5.3. It was posited that even though not specific, perplexity ratios obtained could, in theory, show trends similar to those either of the economic variables paraphrases used pertained to, or possibly to results by the SPF [44]. To correspond to these predictions, economic indicators referring to quarters from the years 2021 and 2022 were used. This adjustment accounted for the shift between predictions and actual values a year later. This way, predictions from January 2020 could be compared to the actual values from January 2021, a year in the future. Consequently, only models from 2020 and 2021 were used, as the actual values of future year-ahead predictions from 2022 have yet to be determined.

Furthermore, data from the SPF for predictions one year ahead were used for comparison.

An additional dataset, to which perplexity ratios were compared, was extracted from the Michigan Survey of Consumers [128]. It is a survey measuring consumer sentiment and expectations in the US conducted by the University of Michigan's Survey Research Center. The survey addresses a random sample of US citizens from telephone records. The survey encompasses questions from several domains, such as personal finances, buying conditions, and economic indicator outlooks. Data collected each quarter from 2020 to

2022 was gathered. The following measures were used to represent consumers' views of unemployment, interest rates, and inflation [128]:

1. **To assess participants' forecasts on unemployment, they were asked the question:** 'How about people out of work during the coming 12 months – do you think there will be more unemployment than now, about the same, or less?'

2. **For interest rates:** 'No one can say for sure, but what do you think will happen to interest rates for borrowing money during the next 12 months – will they go up, stay the same, or go down?'

3. **For inflation, the questions were:** 'During the next 12 months, do you think that prices, in general, will go up, or go down, or stay where they are now?' and 'By what percent do you expect prices to go up, on the average, during the next 12 months?'

4. For each of the three questions, respondents answered with one of the three options. Their responses were used to form an index of their aggregate opinions on unemployment, interest rate, and inflation.

The resulting indexes of consumer beliefs, and estimates of variables, all pertaining to values one year in future, were tallied, and their trends were assessed alongside paraphrases' perplexity ratios against actual economic values. This analysis sought to determine whether any predictions referring to a long-term, one-year period in advance could demonstrate external validity, indicating that the results from the fine-tuned models could have far-reaching applications.

Chapter 6

Results

This section mirrors the structure of the Experiments chapter and analyses results obtained.

First, it fleshes out results from fine-tuning the language model on the Investopedia and textbook datasets, suggesting that the degree of paraphrasing is essential in determining the model's economic accuracy.

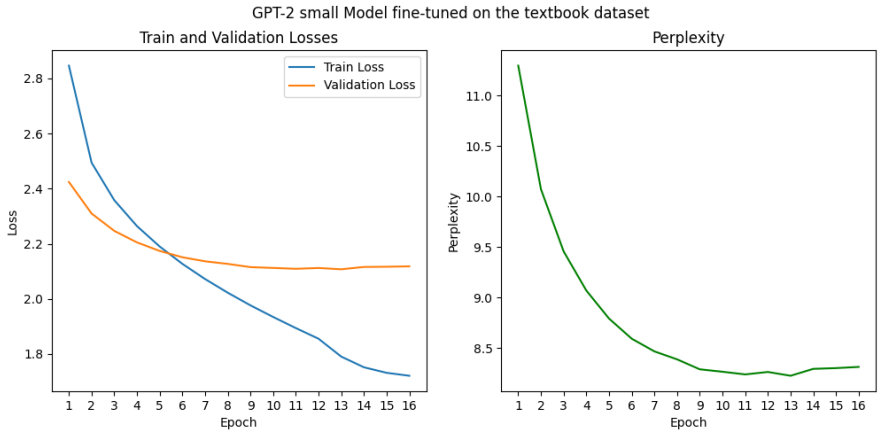
Second, it is concerned with assessing the extent to which results from the Reddit dataset convey a relationship between the models' findings, economic indicators' values, and survey data both in the long and short term. The analysis indicates that perplexity ratios extracted from the model show promising features.

Experiments were conducted using a HPC cluster with an Intel Xeon CPU E5-2690 v4 @ 2.60GHz with 56 cores between two CPUs. Architecture provided was x86_64 and was running a GNU/Linux OS with kernel version 3.10.0-1160.21.1.el7.x86_64. Training was accelerated using two NVIDIA GeForce GTX 1080 Ti GPUs, with CUDA 11.1.1. and cuDNN 8.0.5.39 for optimised GPU operations. All models used the PyTorch library ver. 1.10.0, compiled explicitly for CUDA 11.1.

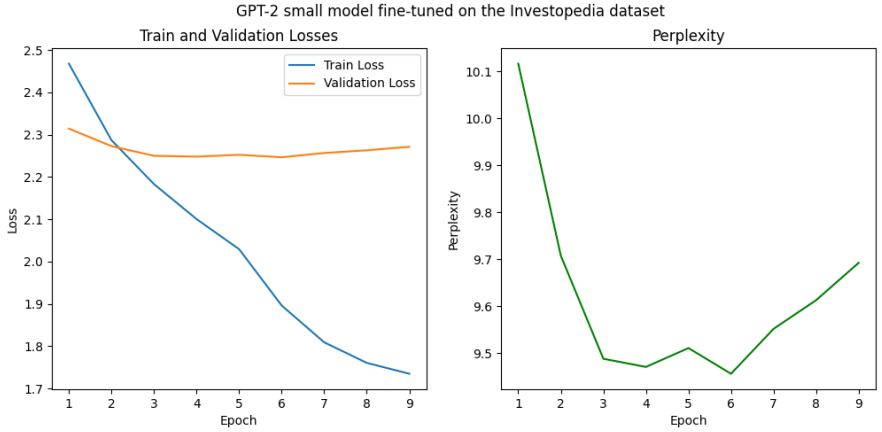
The implementation adapted and employed the pre-trained GPT-2 small model and the pre-trained BPE tokeniser from the Transformer library [135].

6.1 Investopedia and Textbook Dataset Fine-Tuning Results

The first two graphs of Figure 6.1a convey validation and training losses, as well as the model's perplexity during training on the textbook dataset. On the left side, the validation and training loss briskly decreased between the first two epochs and then assumed a more convex trajectory up to the



(a) : Fine-tuning results on the textbook dataset



(b) : Fine-tuning results on the Investopedia dataset

Figure 6.1: Graphs depicting training and validation losses, and perplexity of models trained on (a) the textbook dataset, and (b) Investopedia dataset.

12^{th} epoch. At this point, validation loss ceased to decrease, as it rose from 2.109 in the 11^{th} epoch to 2.112. Consequently, the learning rate was reduced by half. This prompted a further decline of all variables illustrated in the figure during epoch 13 when the learning rate attained its minimal value of 2.107. The final three epochs represented a period in training during which validation loss stopped decreasing, and training was terminated due to the early stopping criterion. As demonstrated by the perplexity graph, the model's lowest perplexity, 8.227, was also achieved during the 13^{th} epoch, after which it gradually increased and seemingly stabilised.

Figure 6.1b portrays the validation loss for the model trained on the Investopedia dataset, which reached its minimal value of approximately 2.25 after the 6^{th} epoch, coinciding with the lowest training perplexity of 9.45. This point also marked the model's highest economic accuracy on selected sentences at 74%. As highlighted by the graphs, this minimum is achieved after all three variables divot after the 5^{th} epoch, which is caused by the

change in learning rate elicited by the early stopping strategy employed. While the training loss continues to decline beyond the 6^{th} epoch, the validation loss and perplexity do not, leading the stopping criterion to halt training after the 9^{th} epoch. Experimentally, this study was replicated with various permutations of hyperparameters, fixed 30 epochs of training, and even with the division of learning rate by factors of 5 and 10, as opposed to 2, which was the default setting for this experiment. However, none of the aforementioned yielded improved results.

For both models, weights and parameters were saved following each epoch, allowing for the retrieval of those corresponding to the model's optimal performance upon completion of the training.

Despite the apparent similarities between the two experiments, there are notable distinctions between the results obtained.

A compelling observation when comparing the two sets of results is that whilst it took the model approximately 12 epochs to train on the textbooks dataset before it started overfitting, on the Investopedia dataset, which comprises over 190 000 lines of text as opposed to 128 259 lines that make up the textbook dataset, the effective training necessitated mere 6 epochs. Some could posit that the variation in jargon across the datasets could have better facilitated the model's adaptation to the Investopedia dataset, given its less domain-specific notation and language. On the other hand, OpenAI's source code release suggests that the WebText dataset on which the GPT-2 model was pre-trained on included thousands of samples from Investopedia [103]. It could be thus inferred that the model required less training on the Investopedia dataset due to its similarity with data within the WebText dataset, possibly accounting for the reduced duration of training [1]. However, other variables, such as sentence and batch shuffling, hyperparameters, and the size and complexity of each dataset, may also play a role in this occurrence.

Another interesting observation, which could potentially be in part explained by the familiarity of the GPT-2 model with Investopedia texts, is elicited by the distinct patterns exhibited by the perplexity curves of the two models. Perplexity in the Investopedia training started increasing sharply once overfitting began, while in the case of the textbook dataset, it rose only marginally. The model's potential pre-existing training on the Investopedia texts, combined with a higher learning rate of 1e-4 compared to 1e-5 used with the textbook dataset, may have contributed to the faster convergence and subsequent brisk over-fitting.

Both models were assessed using the Economic Accuracy (EA) metric, as illustrated in Figure 6.2. After the first epoch, the textbook and Investopedia-

[1] In spite of the same source, it should be noted that there is not an overlap of texts between the WebText and Investopedia datasets, as the former contained articles dated from January 2022 onwards, which was after the WebText dataset was created.

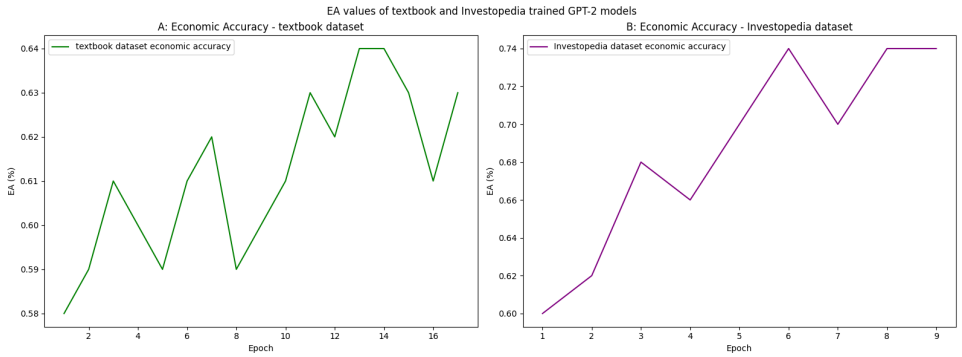


Figure 6.2: Economic accuracies of models trained on (a) the textbook and (b) Investopedia datasets. As can be observed, both curves' trends are rather volatile, with an overall increasing trend. The maximal score for each dataset coincided with the lowest validation loss observed and were 64% (after the 13^{th} epoch) and 74% (after the 6^{th} epoch) for the textbook and Investopedia datasets respectively.

trained models exhibited EA of 58% and 60%, respectively. For reference, the pre-trained model without fine-tuning yielded an EA of 54% and 58% on the textbook and Investopedia datasets, respectively, at zero-shot (i.e. when tasked with evaluating previously unseen prompts). As seen in the graphs, the progression of both curves appears somewhat erratic despite a positive overall trend. The volatility is partially attributable to the pairwise nature of the EA metric, which relies on comparisons rather than continuous functions. In instances of marginal differences between sentence pairs, this can lead to abrupt shifts, as exemplified in the graphs in Figure 6.2. Nevertheless, the maximum EA of both models coincided with their lowest validation loss: 64% for the textbook dataset after the 13th epoch and 74% after the 6^{th} epoch for the Investopedia dataset.

Although the textbook dataset was the smaller of the two, the economic accuracy of the model was 10% below that of the Investopedia dataset. Several factors may account for this discrepancy.

For one, a dataset's size is not the sole determinant of a model's perplexity concerning sentences at hand. The language utilised in textbooks, eloquent and replete with specialised terms and domain-specific notation, may be less familiar to the model than the Investopedia texts.

Furthermore, the sentences selected for evaluation can significantly influence the EA metric. The paraphrases of chosen statements for each dataset were created at random without adhering to a specific pattern, suggesting that their phrasing may have diverged from the dataset's original statements, especially in the case of the textbook dataset, as it was more technical and thus sensitive to wording. A more accurate prediction might have been possible had the sentences been more aligned with the textbook corpus. This conjecture is partially supported by a supplementary experiment, in which

sentences with higher perplexities elicited by a non-fine-tuned model were removed, which led to an increase in EA to 70% on the textbook dataset. Lastly, since the two models employed distinct sets of hyperparameters and training sets, expecting identical outcomes would be presumptuous.

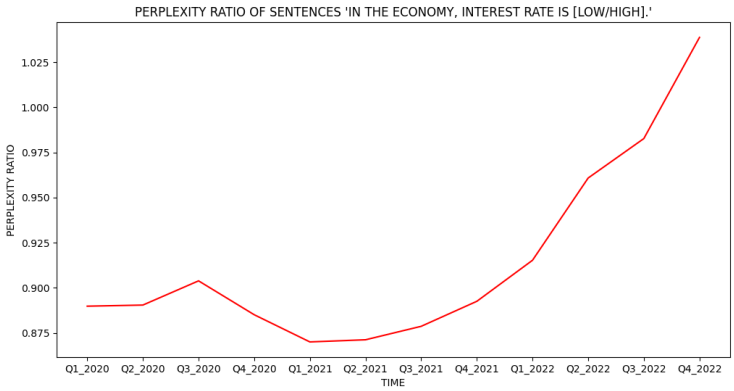
Another plausible reason explaining the suboptimal EAs is the presence of contradictory sentences within the corpus. A peculiar aspect of the textbook dataset is the inclusion of end-of-unit exercises, where students are tasked to label statements as true or false. It could thus be the case that some of the statements were either chosen as part of the EA evaluation set or may have hindered the model's training. However, it is unlikely that the textbook dataset would hold many opposing statements. Nonetheless, given the dataset's size, the inability to verify the potential influence of confounding statements warrants consideration.

Overall, the values of EA for both models set them apart from the results without fine-tuning. However, at the same time, the EA results could be improved. For one, increasing the size of datasets could reiterate overarching concepts, thereby facilitating better understanding and memorisation by the model. At the same time, despite a thorough hyperparameter search, it is possible that the set of hyperparameters used could further be improved and, perchance, even more hyperparameters, such as the number of attention heads used, could lead to improved results in future trainings.

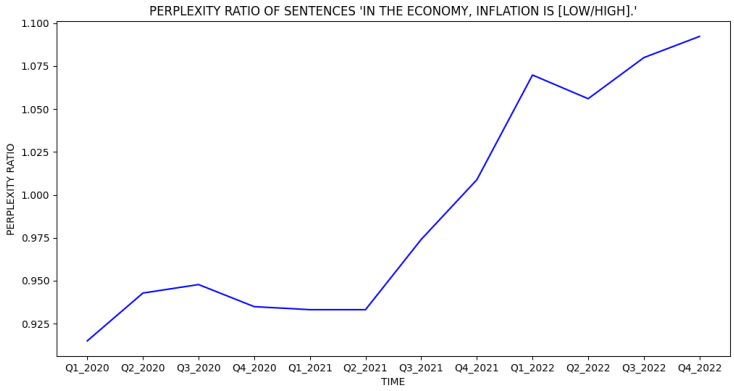
Despite the limitations mentioned, the results, as conveyed in Table 6.1, demonstrate that the GPT-2 model can acquire economic knowledge from the datasets under the specified conditions. Furthermore, the convergence of validation losses, together with the EA scores of 64% and 74%, underscore the model's potential applicability in performing more sophisticated economic analyses. The results serve as a promising indicator of the model's capabilities and provide a solid foundation for further exploration of domain-specific understanding. The findings attest to the potential adaptability of language models in realm of economics and suggests that the model may be employed for more nuanced economic inferences, as exemplified in the Reddit experiment.

	Fine-tuned model	Default model
Textbook dataset	64%	54%
Investopedia dataset	74%	58%

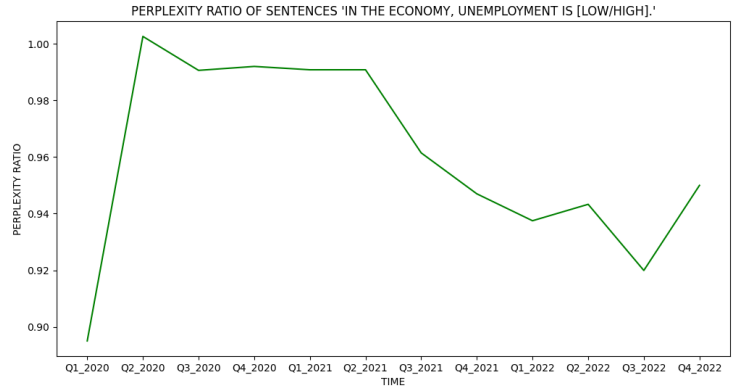
Table 6.1: Economic accuracies on the textbook and Investopedia datasets of the GPT-2 small model with and without fine-tuning. As can be seen, fine-tuning significantly increased the model's understanding of the domain-specific economic texts.



(a) : Perplexity ratio of sentences 'In the economy, interest rate is [low/high].'



(b) : Perplexity ratio of sentences 'In the economy, inflation is [low/high].'



(c) : Perplexity ratio of sentences 'In the economy, unemployment is [low/high].'

Figure 6.3: Perplexity ratios of contradicting sentences pertaining to interest rate, inflation, and unemployment obtained from GPT-2 small models fine-tuned on datasets from quarters of 2020, 2021 and 2022.

6.2 Reddit Experiment Results

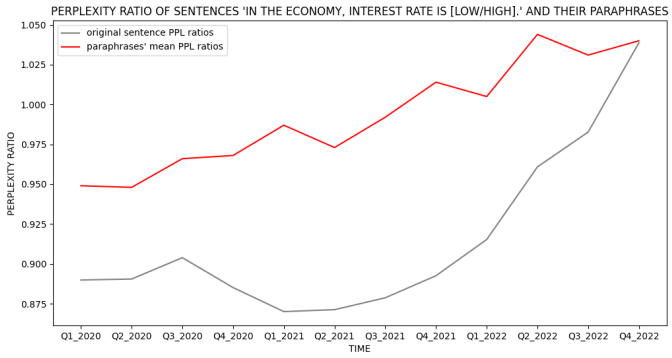
In Figure 6.3, one can observe the results of perplexity ratios of sentences in the format mentioned in Figure 5.1, which were tested on the model fine-tuned on Reddit datasets. As can be discerned from Figure 6.3a, up until the third quarter of 2020, the interest rate was deemed low rather than high. The curve shows an increase leading up to the third quarter of 2020, and then its progression becomes more convex, showing a clear trend that, gradually, the interest rate was perceived to be increasing. The progression of inflation is not too dissimilar, as seen in Figure 6.3b. Despite a similar trend, the shift of belief from deeming inflation relatively low to high began around the fourth quarter of 2021. It sharply rose till the first quarter of 2022, after which it slightly dropped yet continued on an upward trend suggesting the people on Reddit deem inflation to be high rather than low with increasing certainty.

The progression of unemployment depicted in Figure 6.3c had a very different trend. As conveyed by the graph, it rose sharply between the first two quarters of 2020, which coincided with the outbreak of the COVID-19 pandemic [137], and since then steadily assumed a downward trend, which culminated in the third quarter of 2022, when it started increasing. Regarding the semantic meaning of the perplexity ratio, the graph shows that the opinion of unemployment being relatively low rather than high persisted throughout the experiment. This is possibly a reason for concern when interpreting the results. In all three graphs, trends were measured for a particularly worded sentence. For each, several other sentences could be synonymous; however, this could lead to different perplexities and even trends. Before drawing any decisive conclusions, a more thorough approach consisting of testing perplexities on paraphrases specified in Figure 5.2 could lead to results of greater external validity.

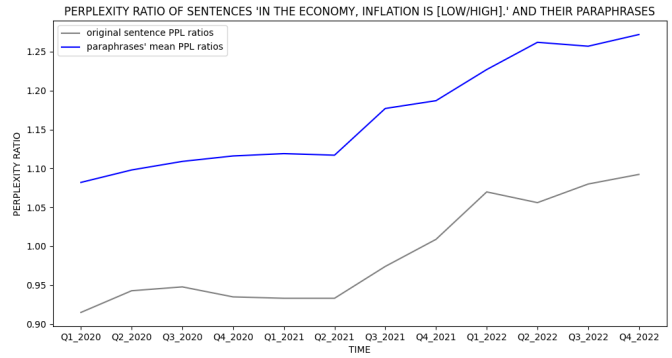
6.2.1 Effect of Paraphrases on Inference and Comparison to Actual Economic Variable Values Over Time

Upon comparing aggregate perplexity ratios of paraphrases, results suggest that trends of the ratios are preserved, although their positions relative to the 1.0 mark are not. Figure 6.4 displays each of the three variables' sentence perplexity ratios, analogous to Figure 6.3, with the addition of coloured curves representing the mean perplexity ratio of paraphrases for each sentence. Upon a cursory examination, it is clear that average perplexity ratios for paraphrases are consistently higher across all variables in Figure 6.4. This unexpected outcome suggests that the model might exhibit sensitivity to sentence reformulations, albeit minute ones, and it could be that the changes in perplexity ratios over time, rather than the absolute values of the ratios themselves, serve as a more meaningful indicator of shifts in opinions within Reddit comments.

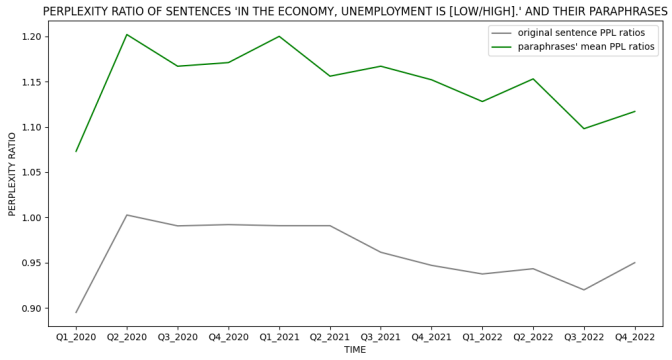
When focusing on overall trends, the results of original sentences and their



(a) : Perplexity ratio of sentences 'In the economy, **interest rate** is [low/high].' and means of their paraphrases' perplexity ratios.



(b) : Perplexity ratio of sentences 'In the economy, **inflation** is [low/high].' and means of their paraphrases' perplexity ratios.



(c) : Perplexity ratio of sentences 'In the economy, **unemployment** is [low/high].' and means of their paraphrases' perplexity ratios.

Figure 6.4: Perplexity ratios of sentences and perplexity ratios of means of respective paraphrases ('In the economy we can see that [economic variable] is [low/high].', 'When it comes to the economy, [economic variable] is [low/high].', and 'With respect to the economy, the [economic variable] is quite [low/high].') perplexities for statements about interest rate, inflation, and unemployment. While paraphrasing preserved trend patterns, it caused vertical shifts in the curves. This suggests that focus should be placed on overall trends rather than specific ratio values, as these may vary with paraphrasing.

paraphrases yield congruent results. However, inference can change when focusing on the perplexity ratio to determine which sentence is more likely. For example, in Figure 6.4c, the overall conclusion from the original sentence would be that unemployment is low. In contrast, the opposite would be drawn based on the averaged values of paraphrases.

The results bring up the question of how to ascertain whether the original sentences or the aggregate values align with the opinions expressed across the Reddit datasets, given the disparity between the perplexity ratios. The results underscore the necessity for an improved understanding of this observation, potentially incorporating the findings into a supervised method, such as logistic regression of economic variables, to enhance external validity and better anchor them to the variables' actual values.

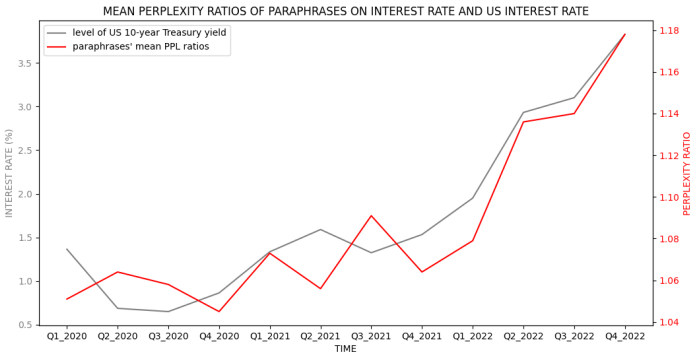
It could also be argued that the perplexity ratios of sentences are closely spaced, suggesting that training the model on larger or additional datasets may lead to more decisive results. As most perplexity ratios traverse the 1.0 threshold, it is crucial to determine the extent to which this could be attributed to a change in opinion across the datasets or potential noise. On the one hand, many of the curves' patterns are explainable through external shifts. For instance, the COVID-19 pandemic caused increased concern about job loss, as possibly evidenced by the spike in Figure 6.4c between the first two quarters of 2020. Similarly, further testing of prompts revealed that the model exhibits a clear shift in opinions regarding the war in Ukraine before and after its commencement.

On the upside, the preservation of trends between the original sentences' ratios and that of their paraphrases suggests the model's capacity to generalise and accurately identify relationships between economic concepts and set the tone of the succeeding experimental sections, which focus solely on trends. Although paraphrasing introduced an upwards shift in the perplexity ratios, the change in trends was minimal. The metric could thus still be effectively used as a relative measure, akin to stock market indices and many other economic indicators.

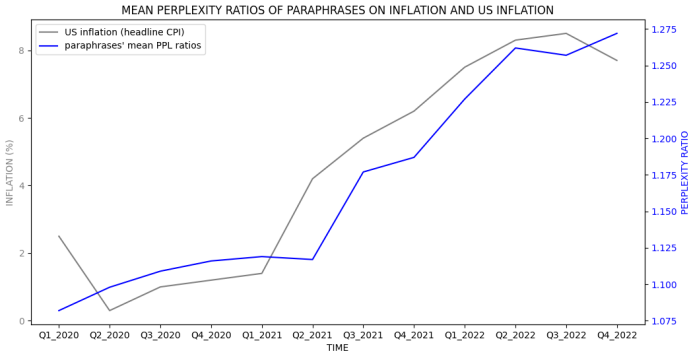
Despite the paraphrasing raising a potential question as to whether the perplexity ratio does represent whether the model truly believes one sentence to be more likely than another, thereby warranting further tests to explore this topic further, the maintenance of the trend in each ratio and the explainability of shifts through external factors affecting people's opinions indicate promise in this approach.

Comparison to Economic Variables

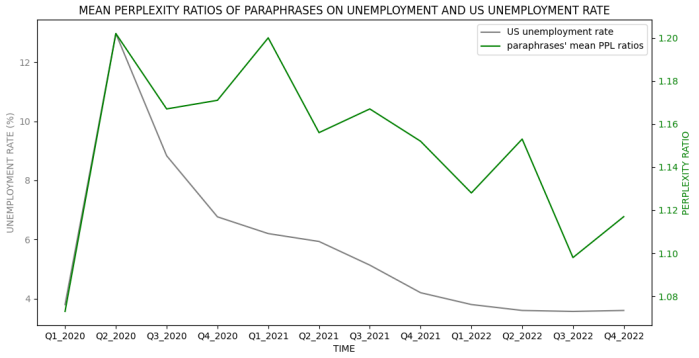
Arguably, even though certainty about the binary opinion on whether Reddit comments indicate that economic variables are high or low is less apparent due to the results in Figure 6.4, the trends observed are of greater significance



(a) : Perplexity ratio of paraphrases compared to levels of **interest rate** (% change in the 10-year Treasury Bills value) over the course of 2020, 2021 and 2022.



(b) : Perplexity ratio of paraphrases compared to levels of **inflation** (yearly headline CPI change) over the course of 2020, 2021 and 2022.



(c) : Perplexity ratio of paraphrases compared to levels of **unemployment** rate over the course of 2020, 2021 and 2022.

Figure 6.5: Mean perplexity ratios of paraphrases ('In the economy we can see that [economic variable] is [low/high].', 'When it comes to the economy, [economic variable] is [low/high].', and 'With respect to the economy, the [economic variable] is quite [low/high].') compared to actual values of interest rate, inflation and unemployment rate.

to the experiments within this thesis and exhibit notable similarities to actual metrics of the economic variables under analysis.

In Figure 6.5, the perplexity ratios of paraphrases are juxtaposed with the actual values of the metrics they mention. Throughout the plots, the trends of aggregate perplexity ratios of paraphrases and economic indicators essentially correspond. In Figure 6.5a, the ratio exhibits greater volatility, but the curves share nearly identical starting and ending points over the three-year time frame. As shown in Figure 6.5b, in the case of inflation, both paraphrases about inflation and the measure of inflation share an overall upward trend, which in theory could be in part attributable to COVID and war in Ukraine induced supply chain issues, and are closely intertwined. In Figure 6.5c, an almost identical response of the unemployment rate and perplexity ratio to the hypothesised change in unemployment triggered by the COVID-19 outbreak can be observed. Following this, both curves display a steady downward trend, with perplexity ratios revealing a more volatile progression, which could indicate uncertainty impacting opinions expressed on Reddit.

Many economic theories and models, such as Prospect theory [67], argue that people are susceptible to adverse shocks, like the one between the first two quarters of 2020. The theory posits that humans are more sensitive to negative stimuli than positive ones of equal magnitude, which may account for the slower decline in perplexity ratios.

Having compared the model's result with real-world metric values and established a close fit in terms of trends, the final step involves assessing predictions of economic indicators compared to survey methods.

6.3 Predictions of Economic Indicators

6.3.1 Short-Term Prediction Results

Following the plans outlined in Section 5.3.3, this part of the Reddit experiment compares short-term predictions to those of SPF, focusing on same-quarter predictions made by experts in the field. Figure 6.6 illustrates that the predictions from the SPF closely align with the actual levels of 10-year Treasury yield (interest rate) overall. However, perplexity ratios in the last three quarters of 2022 seem to convey a closer fit. The graph could be divided into three sections corresponding to the 2020, 2021 and 2022 results. In 2020, the perplexity ratio produced seemingly contradictory predictions compared to the actual results and the SPF. In 2021, the perplexity ratio displayed somewhat volatile results, but followed a trend similar to the other two curves. In 2022, however, the perplexity ratio arguably demonstrated a closer fit to the actual Treasury yield values than the predictions by professionals. Overall, a clear correlation exists between the models' results and real-world values

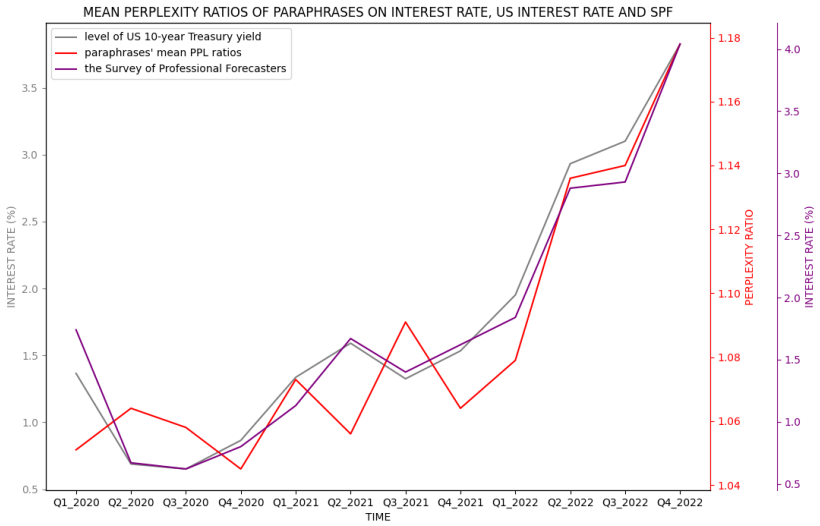


Figure 6.6: Short-term prediction of **interest rate**: actual interest rate values, perplexity ratios of paraphrases ('In the economy we can see that interest rate is [low/high].', 'When it comes to the economy, interest rate is [low/high].', and 'With respect to the economy, the interest rate is quite [low/high].') and SPF same-quarter predictions showing increasing and similar trends over quarters of 2020, 2021 and 2022.

measuring interest rates, even though the SPF's predictions are generally closer [2].

As depicted in Figure 6.7, US inflation initially decreased between the first two quarters of 2020 and then demonstrated an upward, nearly concave trend until the end of 2022. While the SPF results accurately grasped the initial drop, the perplexity ratio seemingly did not. On the other hand, from the second quarter of 2020 onward, the perplexity ratio exemplified a very similar, if not tighter, fit to the inflation values than the results by the professionals, as values extracted from the SPF showed more irregularities in dissonance with US inflation data, such as when their values spiked around the third quarter of 2021. Overall, all three curves show similar trends.

Regarding unemployment, Figure 6.8 reveals that the short-term predictions from the SPF closely mirrored the unemployment rate trends. All three variables captured a significant spike between the first two quarters of 2020. While the perplexity ratio exhibited a slow downward trend with numerous

[2]Correlation measures were purposefully not included within the body of the thesis, as due to the small sample size, their statistical significance is diminished. That being said, the Pearson correlation between perplexity ratios and metric values overall was over 0.85 for short-term estimates, which is given the ranges from -1 to 1 indicative of a very strong correlation. However, the results could be very different with a larger dataset, hence the need for caution when interpreting results.

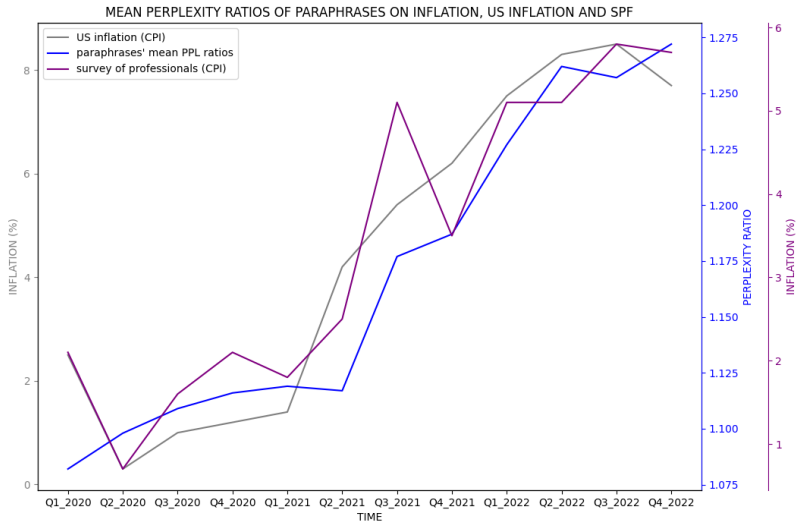


Figure 6.7: Short-term prediction of **inflation**: actual inflation (change in CPI) values, perplexity ratios of paraphrases ('In the economy we can see that inflation is [low/high].', 'When it comes to the economy, inflation is [low/high].', and 'With respect to the economy, inflation is quite [low/high].') and SPF same-quarter predictions of core CPI percentual change showing increasing and rather tight trends over quarters of 2020, 2021 and 2022, especially between results obtained from the fine-tuned GPT-2 and actual inflation.

spikes, the SPF demonstrated a near-perfect alignment of trends.

In summary, there is a closer fit between the SPF and the actual values of inflation, unemployment and interest rates than with the perplexity ratios for these three variables overall. However, two factors may improve the outlook of perplexity ratio predictions.

First, unlike the SPF, predictions based on comments from Reddit reflect peoples' opinions. Not necessarily their explicit forecasts and predictions. Consequently, the perplexity ratios and the SPF model different variables. Perplexity ratios are not intended to accurately mirror economic variables, unlike the results from the SPF. Therefore, It is expected and welcome to see discrepancies that can be analysed further.

Second, while perplexity ratios refer to terms such as 'inflation' and 'interest rate', the survey results predicted headline CPI and 10-year Treasury yield as proxies for the two variables. This difference may explain the closer fit of predictions, as inflation and interest rate are umbrella terms that can be approximated with various indices and variables. For example, CPI could be replaced by the PCE (Personal Consumption Expenditures Price) index, and both could be seasonally adjusted. Interest rate measures could also utilise bonds with different maturities, potentially altering the results.

Figure 6.8: Short-term prediction of **unemployment**: actual unemployment values, perplexity ratios of paraphrases ('In the economy we can see that unemployment is [low/high].', 'When it comes to the economy, unemployment is [low/high].', and 'With respect to the economy, unemployment is quite [low/high].') and the SPF same-quarter predictions of unemployment showing the same abrupt increase in trends between the first two quarters of 2020 and then a gradual decrease. Even though trends of perplexity ratios also decreased overall, it was at a much more volatile rate, suggesting the SPF exemplified much closer alignment to actual unemployment than the fine-tuned model.

Given these adjustments and the focus on overall trends, the predictions from the SPF display closer fits to actual values of the variables in question, but only by a slim margin. Nevertheless, given the adjustments mentioned above, this outcome is surprisingly positive.

6.3.2 Long-Term Prediction Results

The final stage of experiments investigates how predictions from models trained on Reddit compared to the SPF and the Survey of Consumers in predicting interest rate, inflation and unemployment one year ahead[3].

Upon observing Figure 6.9, it becomes clear that the Survey of Consumers' results diverge from the other predictions in terms of trends, beginning from the second quarter of 2020. While the other predictions and the Treasury yield increase, the Survey of Consumers steadily declines. Comparing perplexity

[3]On a technical note, to improve understanding, grey curves in graphs within this subsection 6.3.2 representing US interest rate, inflation and unemployment, have been shifted one year back so that they can be compared to their predictions more effectively. That is why when any of the non-grey curves shows a predicted value of one of the economic variables, the grey curve conveys the actual variable's value the curves predicted in the future.

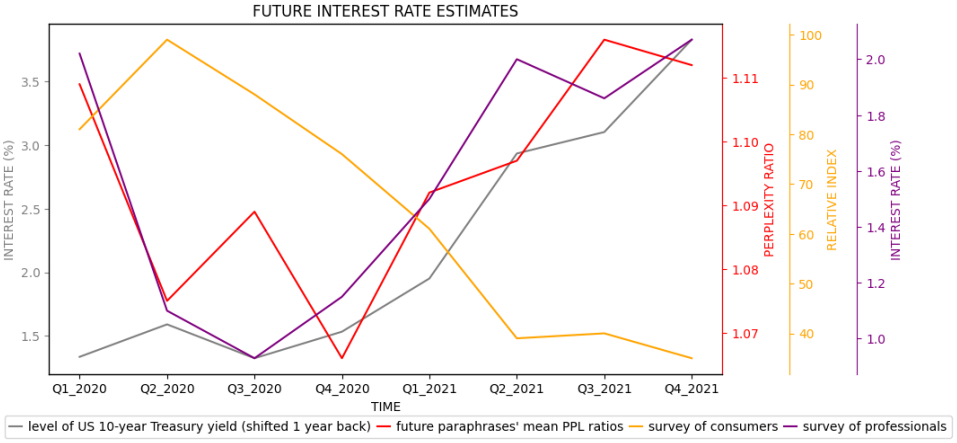


Figure 6.9: Long-term predictions (1 year in the future) of **interest rate**: Comparison of actual interest rate values, perplexity ratios from paraphrases ('In the economy we can see that the interest rate will be [low/high].', 'When it comes to the economy, the interest rate will be [low/high].', and 'With respect to the economy, the interest rate will be quite [low/high].'), along with predictions from the SPF and the Survey of Consumers. As shown in the graph, the SPF's predictions more closely capture future interest rate trends compared to the perplexity ratios, and significantly better than the Survey of Consumers. These results suggest that fine-tuned language models may extract public opinions more effectively than survey-based methods aimed at the public.

ratios and the SPF reveal a more nuanced relationship. Between the first two quarters of 2020, both significantly decreased, in contrast to the actual values they were used to predict. Subsequently, from the second quarter of 2020 to the first quarter of 2021, the SPF exhibited a closer fit to the Treasury yield in terms of trend, albeit with exaggerated quarter-to-quarter changes. Consequently, the more conservative progression of ratios aligned more closely with the Treasury yield.

Overall, there is less predictive power in all three estimates than in the short term, which is unsurprising. However, whilst the Survey of Consumers fails to grasp the overall tendencies of Treasury rate yields, the perplexity ratio and SPF results exhibit some degree of alignment with the trends.

Unlike with interest rate, in the case of inflation, as exemplified by the graph in Figure 6.10, the Survey of Consumers show similar trends to those of the actual inflation one year in future, as the trends of the Survey of Consumers are closer to trends of inflation than the SPF or the perplexity ratios. An overall increasing tendency, with a slight decrease between the first two quarters of 2020 for perplexity ratios and the SPF, describes all three predictions. Whilst the SPF's curve showed moderate fluctuations, similar to that of perplexity ratios, the Survey of Consumers curve in orange had a steady upward progression, especially from 2021 onwards. Unfortunately, the perplexity ratio predictions fail to capture the actual inflation rate as

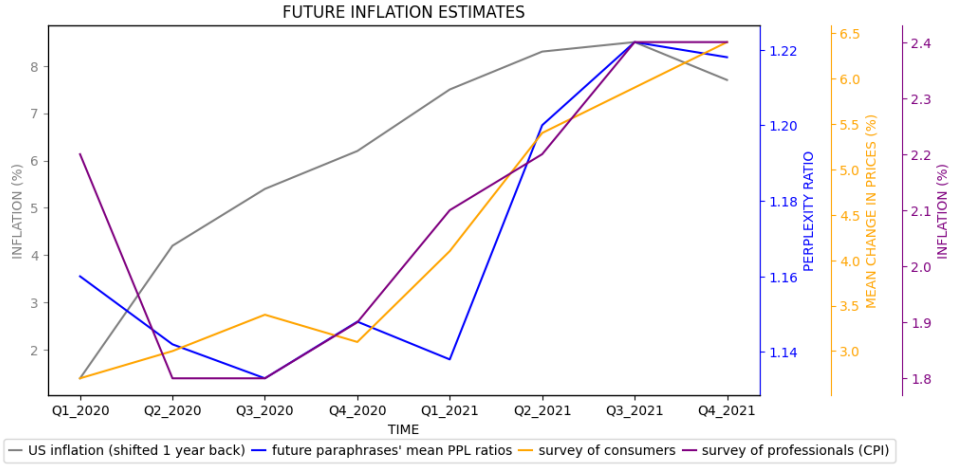


Figure 6.10: Long-term predictions (1 year in the future) of **inflation**: Comparison of actual inflation values one year in the future, perplexity ratios of paraphrased sentences ('In the economy we can see that inflation will be [low/high].', 'When it comes to the economy, inflation will be [low/high].', and 'With respect to the economy, inflation will be quite [low/high].'), predictions from the SPF, and predictions from the Survey of Consumers. As the graph suggests, most of these predictions failed to accurately forecast future trends in the inflation rate, with the SPF providing the closest predictions.

closely as in the short run. However, it could be argued that the results not only captured the overall increasing trend but also potentially matched in performance results by the SPF.

In terms of unemployment predictions in the last graph of this subsection shown in Figure 6.11, unemployment displays a consistent decline in its trend throughout 2021 and 2022. Both the SPF and perplexity ratios incorrectly predicted an increase in unemployment between their predictions in the first and second quarter of 2020, and the Survey of Consumers followed suit, albeit with a more moderate prognosis. From the second quarter of 2020 predictions onward, the SPF forecasts seemingly captured the trend of unemployment, as it closely followed the trend of actual unemployment rate levels. However, the Survey of Consumers and the ratio of perplexities did not demonstrate the same level of accuracy, as their predictions' trends largely failed to reflect the actual values of the unemployment forecasted one year ahead. However, what is very interesting to observe is the similarity between predictions of the Survey of Consumers and perplexity ratios from the fourth quarter of 2020 onwards. The graduate increase in trends which peaked around the second quarter of 2021 and then steadily decreased in both of their cases, could be suggestive of a shift in opinions by consumers and certainly warrants further economic analysis assessing potential causes of this.

The expected distortion of values one year ahead is further exacerbated by external economic shocks, such as the COVID-19 pandemic, the war in

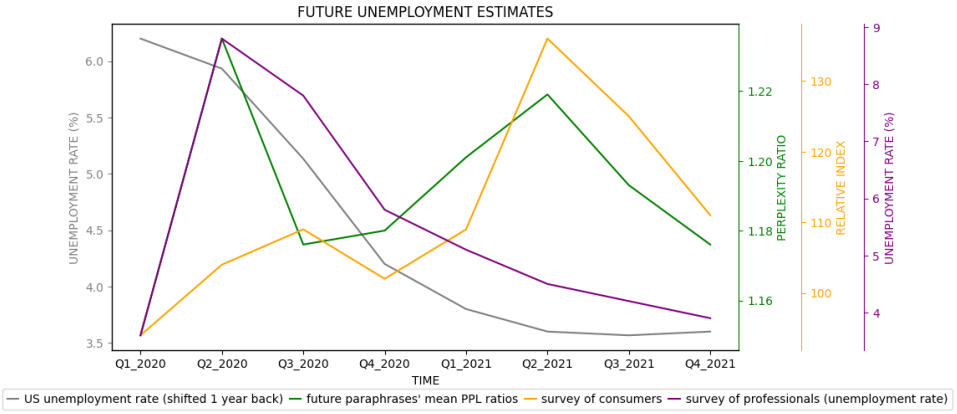


Figure 6.11: Long-term predictions (1 year in the future) of **unemployment**: actual unemployment values one year in the future, perplexity ratios of paraphrases ('In the economy we can see that unemployment will be [low/high].', 'When it comes to the economy, unemployment will be [low/high].', and 'With respect to the economy, the unemployment will be quite [low/high].'), SPF's four quarters ahead predictions and corresponding predictions by the Survey of Consumers.

Ukraine, and the ensuing food shortages, which may influence both economic indicators and their public perception in a multitude of ways.

As with the short-term predictions, the SPF unsurprisingly came up on top in terms of predictions of economic variables. This success can be attributed, in part, to the fact that it predicted specific values of particular variables (i.e. CPI, 10-year Treasury yield) rather than general terms such as changes in price and inflation, as was the case for the Survey of Consumers and perplexity ratios of sentences created.

The Survey of Consumers values in their year-ahead estimates, for the most part, did not adequately capture future economic trends. The intrinsic value of the survey, however, lies in the representation of the public's opinions on said variables, which are expected to be skewed by several factors, such as biases, lack of information, or misunderstanding.

The results derived from the perplexity ratios provided intermediate outcomes, falling primarily between predictions of both surveys. This finding is encouraging, as it indirectly supports the hypothesis that public knowledge possesses greater predictive power for economic indicators' values. However, survey results may fail to capture this knowledge accurately. On one side, some results, such as the first three quarters of 2020 unemployment predictions and the majority of inflation and interest rate predictions, show positively correlated results to those of the SPF. At the same time, in instances like 2021, unemployment rate predictions strongly aligned with trends by the Survey of Consumers, possibly suggesting an underlying shift in public opinion.

In conclusion, while the predictions from the SPF outperformed the other two sources of predictions in short and long-term forecasts, the perplexity ratios of sentences from models trained on Reddit datasets have demonstrated potential in accurately predicting economic indicators better than the Survey of Consumers. However, further research and provisions ranging from training on larger datasets spanning longer periods of time and considering incorporating alternative sources of datasets to enhance the predictive power of models could greatly improve the results obtained.

Chapter 7

Evaluation

This section evaluates the models' performance in extracting economic knowledge and opinions from datasets and using them in predicting economic indicators. Several aspects of the methodology and results are assessed, and improvements and avenues of further research are delineated.

7.1 Evidence of Potential

Results presented in the previous chapter show that despite falling short in their predictive power behind results of the SPF, the trends derived from the perplexity ratios of sentences illustrate the potential of the proposed methodology when applied to economic opinions and possibly other domains. In addition, the fact that results fall between the SPF and Survey of Consumer values regarding their proximity to actual metrics is encouraging. This lends indirect substantiation to one of the thesis' motivational notions that social media platforms harbour data of great expressiveness when compared to opinion polls and could sometimes be used as substitutes for polls and surveys. Concurrently, its bolsters the assumption that while predictions may not outperform those made in the SPF in the long run, they can offer an improved perspective on what people think about various facets of the economy. In addition, results of fine-tuning on economic texts leading to a scores of 64% and 74% of EA with respect to the textbook and Investopedia datasets respectively underline further potential in using fine-tuned GPT-2 models for domain-specific inference.

Nevertheless, the experimental findings have highlighted several limitations inherent to the methods employed.

7.2 Dataset Constraints and Statistical Inference

A significant impediment of the method used, bearing far-reaching externalities, is the scarcity of datasets. The inclusion of additional data points would allow the experiments to stretch across a more extensive time frame, possibly bolstering the strength of predictions and refining inference. Had a richer set of data points been available, statistical tests of correlation could have been employed to compare the trends of curves and yielded statistically significant results, thereby quantifying the degree of correlation between results obtained. Due to the scarcity of data and its volatility, tests of correlation were not used. In future research, however, this possibility would surely contribute towards improved analysis of results, as with larger datasets of good quality, the language model can generalise better.

7.3 Limitations of the Perplexity Ratio Metric

Perplexity ratio, one of the key statistics used in measuring the model's alignment with statements, has been instrumental throughout the experimental phase in describing what authors of posts language models were trained on thought. At the same time, it is crucial not to overstate its predictive power. As exemplified in section 6.2.1 of the Results chapter, the metric is sensitive to paraphrases of prompts, which can yield misleading outcomes by suggesting a more substantial alignment of the model with one statement over another and subsequently reversing the decision when faced with the statements' paraphrases. Further tests assessing the degree to which the metric's inconclusive results were caused either by possible nuanced errors when paraphrasing, the choice of sentences, or merely a manifestation of the model's ambivalence between the two prompts could help in improving inference when evaluating perplexities of opposing statements. At the same time, however, it proved to be an effective and across paraphrases consistent metric in capturing trends, which were vital for experiments used.

7.4 Benchmarks and Supervised Learning Methods

One of the most significant challenges encountered in this thesis, which is in part attributable to the innovative nature of the proposed method of extracting public opinions, is the absence of benchmarks and external metrics which could validate the obtained results. An attempt to anchor results was made when perplexity ratios were compared to the SPF and the Survey of Consumers. Nevertheless, posters on Reddit are not expected to hold the same opinions as professionals. Similarly, since one of the arguments used is that the methodology used could supersede survey methods, it is expected the results to be different from those obtained through surveys also. Lamentably, as already relayed in the literature review, there are no apparent applications of other methods that would either propose a different approach or provide

comparable results.

A potential solution is to draw inspiration from survey strategies. For example, whenever survey results are disputed, another survey is often conducted to either corroborate or disprove the results of the first one. Similarly, a possible benchmark for results obtained would be fine-tuning another model on a different dataset, possibly different enough to improve generalisability. In this case, potentially Twitter, Facebook or other social media data, with potentially different demographic composition of their userbase, thereby discussing potentially different aspects of economics, could provide greater variety of opinions under analysis. From these, experiments at hand could be replicated and their results compared to assess the robustness of inference.

Additionally, further research could apply results obtained to supervised-learning methods to augment their predictive power. As part of future research, the findings will be applied to logistic regression or other supervised-learning methods to improve their predictive power. Moreover, ensemble methods could be employed across several platforms and domains to extract as much information from social media datasets as possible. This could facilitate the extraction of more specialised and nuanced or generalised public opinion information, which could be extended even into private sectors for deciphering consumer behaviour, market predictions, or even fintech applications.

7.5 Economic Metrics and Unification of Predicted Variables

A predicament that became more prominent in the later stages of experiments is the comparison of heterogeneous variables. Economics is a convoluted discipline wherein variables lack clearly defined and easily measurable values. As such, economic metrics often merely serve as approximations rather than precise measures of some of the variables of interest, such as inflation and interest rate. Concurrently, comment sections on social media platforms rarely discuss economic indicators in a scientifically exact manner. For future research, methods aiming to unify predicted variables across surveys and language model predictions could enhance the effectiveness of comparisons. Focusing on relatively easily measurable variables such as tariffs or taxes, which definition is more straightforward, or using multiple indicators at once (e.g., seasonally adjusted CPI, PCE and their core versions) could offer valuable insights into the model's sensitivity to changes in each. However, as already mentioned, this and many other statistical tests would benefit from more data to draw inferences from.

7.6 Explainability of External Shocks

Other ramifications of conducting economics research which should be addressed prior to further processing of results are potential external factors possibly altering the results obtained. Suppose a politician promises to lower taxes, a neighbouring country defaults or even a volcano erupts in a remote country. In that case, such external factors could influence people's opinions domestically, as cross-country external shocks can have far-reaching externalities exemplified by changes in peoples' opinions [3].

Consequently, a thorough analysis of potential external factors, such as the war in Ukraine influencing inflation expectations described in the results section, influencing data collected is necessary when examining results reflecting public opinions on economics. Once these factors are controlled or accounted for, deeper conclusions can be drawn about potential inconsistencies between predictions and actual values, as either due to external factors or even inherent biases and assumptions, public opinion can be skewed and offer seemingly inconsequential data.

Therefore, despite observing promising trends, a thorough analysis of potential sources of shifts in behaviour, including external factors and domestic political and economic climates, should be considered. If such deeper links with economic research could be established, they could be used to assess the model's results and the effectiveness of the perplexity ratio and economic accuracy used. Adding a layer of economic reasoning atop the results obtained could improve the results' explainability and the model's predictive power.

7.7 Technical Limitations

Some of the limitations of the results obtained are inherently technical. While the model employed is a state-of-the-art Transformer, its performance is contingent upon the quality of its hyperparameters. Despite the thorough search for optimal hyperparameters for each dataset using random and grid search, numerous other parameters and their permutations are yet to be tested when training the model, ranging from warmup, learning rate schedulers, mixed-precision adjustments, and many more. Although a lower validation loss does not guarantee better results in terms of other metrics, such as economic accuracy, improvements in regularisation techniques and hyperparameters could increase the generalisability of the model and, thereby, its predictive power.

Chapter 8

Conclusion

This thesis aimed to explore the potential relationships between public opinions and forecasts of economic variables, specifically unemployment, inflation and interest rate, and actual economic events and variables' trends, through a new approach using a language model fine-tuned on datasets from social media over different time periods.

One of the key contributions of this thesis were datasets curated for the task at hand, varying from expert-level textbooks to Reddit comments containing the three economic keywords spanning quarters from 2020 till 2022. Leveraging several sources of domain-specific data, a statistical model was developed equipped with economic knowledge, as shown by the economic accuracies of 64% and 74% on the textbook and Investopedia datasets respectively, lending itself to further uses in economics and other downstream tasks.

At the same time, when it comes to evaluation of datasets, however, Reddit is hardly an all-opinion-encompassing platform, as it may be a too biased dataset to draw inference with regards to public opinions from due to bots, the demographic composition of its users which could potentially detract from its representativeness and other possible factors creating noise within the data. That is why in future research, larger, more varied datasets from several social networks and over longer time frames could improve the robustness of the methods used, as the more relevant data is, the better the model's results will be and with data accounting for more time frames, statistical tests of correlations and supervised-learning methods to analyse results' external validity could be employed and lead to improved results.

Another notable contribution of the thesis are the economic accuracy and perplexity ratio metrics. Thanks to these, it was possible to automatically monitor the model's comprehension of economic texts as well as trends of opinions in Reddit over time. Due to the fact the perplexity ratio metric has reservations, especially given the discrepancy of ratios among paraphrases, to further solidify their value, they would benefit from validation ideally

deployed on the same datasets or similar demographic groups to strengthen their value for future applications.

Once datasets and metrics of evaluation were tested, they gave rise to a framework, as part of which the fine-tuned model was applied to Reddit data, which unveiled results suggesting that when tallied against traditional methods of economic variable prediction such as the Survey of Professional Forecasters (SPF) or the Survey of Consumers, it attained surprisingly good results, as trends of perplexity ratios, especially in the short run, overall mirrored those of the SPF and actual economic variables‘ values. Even though the model did not fully capture every trend, especially those in the long run, its results present possible causal links such as the one between the outbreak of the COVID-19 pandemic and unemployment forecasts given the mentioned prospect theory-based explanation. Furthermore, if a layer of additional economic analysis was applied to the results obtained, it could uncover further links between economic outcomes and opinions thereby providing additional impetus for this line of research and further substantiating the notion of causal relationships between economic variables and public opinions.

This thesis represents an important stride towards harnessing opinions through language models and their use in assessing public's predictions of economic variables. Despite the need for further refinement and validation of certain aspects of the work presented, the methodology within this thesis laid important groundwork for further applications in many other domains, and serves as a potential complement and perhaps even an alternative to surveys and polls in extracting opinions.

Appendix A

Bibliography

[1] Abi Adams-Prassl, Teodora Boneva, Marta Golin, and Christopher Rauh. Inequality in the impact of the coronavirus shock: Evidence from real time surveys. *Journal of Public economics*, 189:104245, 2020.

[2] Gati Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans. *arXiv preprint arXiv:2208.10264*, 2022.

[3] John Ahlquist, Mark Copelovitch, and Stefanie Walter. The political consequences of external economic shocks: evidence from poland. *American Journal of Political Science*, 64(4):904–920, 2020.

[4] Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3159–3166, 2019.

[5] Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90, 2015.

[6] Dolan Antenucci, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew D Shapiro. Using social media to measure labor market flows. Technical report, National Bureau of Economic Research, 2014.

[7] Anthropic. Introducing claude. `https://www.anthropic.com/index/introducing-claude`, 3 2023. Accessed: 2023-04-25.

[8] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

[9] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *arXiv preprint arXiv:2209.06899*, 2022.

[10] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 492–499, 2010.

[11] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[12] Reg Baker. Big data: A survey research perspective. *Total survey error in practice*, pages 47–69, 2017.

[13] World Bank. Individuals using the internet (% of population): United states. `https://data.worldbank.org/indicator/IT.NET.USER.ZS?end=2020&locations=US&start=2019`, 2023. Accessed: 2023-04-25.

[14] Pablo Barberá and Gonzalo Rivero. Understanding the political representativeness of twitter users. *Social Science Computer Review*, 33(6):712–729, 2015.

[15] Christine Basta, Marta R Costa-Jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783*, 2019.

[16] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In Munmun De Choudhury, Rumi Chunara, Aron Culotta, and Brooke Foucault Welles, editors, *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 830–839. AAAI Press, 2020.

[17] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? parrot. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[18] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[19] Adam J Berinsky. Measuring public opinion with surveys. *Annual review of political science*, 20:309–329, 2017.

[20] Francesco Bianchi and Cosmin Ilut. Monetary/fiscal policy mix and

agents' beliefs. *Review of economic Dynamics*, 26:113–139, 2017.

[21] Tiago Bianchi. Regional distribution of desktop traffic to reddit.com as of may 2022 by country, May 2022. Accessed: 2023-04-15.

[22] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[23] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

[24] Board of Governors of the Federal Reserve System (US). Interest rates and price indexes; 10-year treasury yield, level [bogz1fl073161113q]. `https://fred.stlouisfed.org/series/BOGZ1FL073161113Q`, n.d. [Online; accessed 2023-05-04]. Retrieved from FRED, Federal Reserve Bank of St. Louis.

[25] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[26] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[27] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[28] Filipe Campante, Federico Sturzenegger, and Andrés Velasco. *Advanced macroeconomics: an easy guide*. LSE Press, 2021.

[29] Christopher D Carroll. Macroeconomic expectations of households and professional forecasters. *the Quarterly Journal of economics*, 118(1):269–298, 2003.

[30] Pew Research Center. Internet/broadband fact sheet. `https://www.pewresearch.org/internet/fact-sheet/internet-broadband/`, 2021. Last updated: 2021-04-07, Accessed: 2023-04-25.

[31] Andrea Ceron, Luigi Curini, Stefano M Iacus, and Giuseppe Porro. Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to italy and france. *New media & society*, 16(2):340–358, 2014.

[32] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[33] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[34] Eric Chu, Jacob Andreas, Stephen Ansolabehere, and Deb Roy. Language models trained on media diets can predict public opinion. *arXiv preprint arXiv:2303.16779*, 2023.

[35] Michael Collins. Natural language processing lecture notes. Course Notes for CS4705, 2017. Columbia University.

[36] Frederick G Conrad, Johann A Gagnon-Bartsch, Robyn A Ferg, Michael F Schober, Josh Pasek, and Elizabeth Hou. Social media as an alternative to surveys of opinions about the economy. *Social Science Computer Review*, 39(4):489–508, 2021.

[37] Dean Croushore et al. Evaluating inflation forecasts. Technical report, 1998.

[38] David Cummings, Haruki Oh, and Ningxuan Wang. Who needs polls? gauging public opinion from twitter data. *Unpublished manuscript*, 2010.

[39] Chester Curme, H Eugene Stanley, and Irena Vodenska. Coupled network approach to predictability of financial market returns and news sentiments. *International Journal of Theoretical and Applied Finance*, 18(07):1550043, 2015.

[40] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

[41] Fernando Diaz, Michael Gamon, Jake M Hofman, Emre Kıcıman, and David Rothschild. Online and social media data as an imperfect continuous panel survey. *PloS one*, 11(1):e0145406, 2016.

[42] Hugging Face. Gpt-2. Hugging Face Transformers Documentation, 2021. Accessed: 2023-04-25.

[43] Hugging Face. The hugging face course, 2022. `https://huggingface.co/course`, 2022. Accessed: 2023-04-25.

[44] Federal Reserve Bank of Philadelphia. Survey of professional forecasters. `https://www.philadelphiafed.`

org/surveys-and-data/real-time-data-research/survey-of-professional-forecasters, n.d. [Online; accessed 2023-04-25]. Retrieved from Federal Reserve Bank of Philadelphia.

[45] Ingrid E Fisher, Margaret R Garnsey, and Mark E Hughes. Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, 23(3):157–214, 2016.

[46] Jordi Galí. *Monetary policy, inflation, and the business cycle: an introduction to the new Keynesian framework and its applications*. Princeton University Press, 2015.

[47] Marta Garnelo and Murray Shanahan. Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. *Current Opinion in Behavioral Sciences*, 29:17–23, 2019.

[48] Daniel Gayo-Avello. No, you cannot predict elections with twitter. *IEEE Internet Computing*, 16(6):91–94, 2012.

[49] Daniel Gayo-Avello, Panagiotis Metaxas, and Eni Mustafaraj. Limits of electoral predictions using twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 490–493, 2011.

[50] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

[51] Steven A Greenlaw, David Shapiro, and Timothy Taylor. Principles of macroeconomics 2e (openstax). 2018.

[52] Emma Haddi, Xiaohui Liu, and Yong Shi. The role of text pre-processing in sentiment analysis. *Procedia computer science*, 17:26–32, 2013.

[53] Bruce E. Hansen. *Econometrics*. Princeton University Press, Princeton, NJ, 2022.

[54] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*, 2023.

[55] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[56] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus).

arXiv preprint arXiv:1606.08415, 2016.

[57] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[58] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[59] John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv preprint arXiv:2301.07543*, 2023.

[60] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[61] Sofia Hurtado, Poushali Ray, and Radu Marculescu. Bot detection in reddit political discussion. In *Proceedings of the fourth international workshop on social sensing*, pages 30–35, 2019.

[62] Investopedia. Investopedia: Sharper insight, better investing, 2023. Accessed: 2023-04-15.

[63] Shanto Iyengar, Mark D Peters, and Donald R Kinder. Experimental demonstrations of the "not-so-minimal" consequences of television news programs. *American political science review*, 76(4):848–858, 1982.

[64] Chris Janiszewski and Robert S Wyer Jr. Content and process priming: A review. *Journal of consumer psychology*, 24(1):96–118, 2014.

[65] Dan Jurafsky and James H. Martin. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J., 2009.

[66] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

[67] DANIEL Kai-Ineman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):363–391, 1979.

[68] Hannah Kim and Young-Seob Jeong. Sentiment classification using convolutional neural networks. *Applied Sciences*, 9(11):2347, 2019.

[69] Stephanie Kirchgaessner, Manisha Ganguly, David Pegg, Carole Cad-

walladr, and Jason Burke. Revealed: Disinformation team 'jorge' claim meddling elections, Feb 2023. Accessed: 2023-04-15.

[70] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. Opennmt: Open-source toolkit for neural machine translation. In *Proc. ACL*, 2017.

[71] Anton Korinek. Language models and cognitive automation for economic research. Technical report, National Bureau of Economic Research, 2023.

[72] Anis Koubaa. Gpt-4 vs. gpt-3.5: A concise showdown. 2023.

[73] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the international AAAI conference on web and social media*, volume 5, pages 538–541, 2011.

[74] Mathias Kraus and Stefan Feuerriegel. Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees. *Expert Systems with Applications*, 118:65–79, 2019.

[75] Jon A Krosnick, Charles M Judd, and Bernd Wittenbrink. The measurement of attitudes. 2005.

[76] Paul Krugman and Robin Wells. Macroeconomics, 4th, 2015.

[77] Robert M Kunst. Introduction to macroeconomics lecture notes. *Retrieved from Universitat Wien: https://homepage. univie. ac. at/robert. kunst/macro1. pdf*, 2006.

[78] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*, 2019.

[79] Vasileios Lampos and Nello Cristianini. Tracking the flu pandemic by monitoring the social web. In *2010 2nd international workshop on cognitive information processing*, pages 411–416. IEEE, 2010.

[80] Victor Lavrenko, Matt Schmill, Dawn Lawrie, Paul Ogilvie, David Jensen, and James Allan. Language models for financial news recommendation. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 389–396, 2000.

[81] Jieh-Sheng Lee and Jieh Hsiang. Patent claim generation by fine-tuning openai gpt-2. *World Patent Information*, 62:101983, 2020.

[82] Sergey Levine. Cs 182: Lecture 12: Part 3: Transformers. `https://www.youtube.com/watch?v=DepabjkETSA&t=640s`, 2021. CS W182

/ 282A at UC Berkeley. Designing, Visualizing and Understanding Deep Neural Networks. Lectures: M/W 5:30-7 p.m., via Zoom.

[83] Junlong Li, Yiheng Xu, Lei Cui, and Furu Wei. Markuplm: Pre-training of text and markup language for visually-rich document understanding. *arXiv preprint arXiv:2110.08518*, 2021.

[84] Lars Ljungqvist and Thomas J Sargent. Recursive macroeconomic theory second edition, 2004.

[85] Kiel Long, John Vines, Selina Sutton, Phillip Brooker, Tom Feltwell, Ben Kirman, Julie Barnett, and Shaun Lawson. " could you define that in bot terms"? requesting, creating and using bots on reddit. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3488–3500, 2017.

[86] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65, 2011.

[87] Bernhard Lutz, Nicolas Pröllochs, and Dirk Neumann. Sentence-level sentiment analysis of financial news using distributed text representations and multi-instance learning. *arXiv preprint arXiv:1901.00400*, 2018.

[88] Macedo Maia, André Freitas, and Siegfried Handschuh. Finsslx: a sentiment analysis model for the financial domain using text simplification. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pages 318–319. IEEE, 2018.

[89] N Gregory Mankiw. *Principles of macroeconomics*. Cengage learning, 2020.

[90] Christopher Manning, Ashish Vaswani, and Anna Huang. Stanford cs224n: Nlp with deep learning | winter 2019 | lecture 14 – transformers and self-attention. YouTube, 2019. Available: `https://www.youtube.com/watch?v=5vcj8kSwBCY` (Accessed: 2023-05-15).

[91] Johnnatan Messias, Pantelis Vikatos, and Fabrício Benevenuto. White, man, and highly followed: Gender and race inequalities in twitter. In *Proceedings of the international conference on web intelligence*, pages 266–274, 2017.

[92] Kostadin Mishev, Ana Gjorgjevikj, Irena Vodenska, Lubomir T Chitkushev, and Dimitar Trajanov. Evaluation of sentiment analysis in finance: from lexicons to transformers. *IEEE access*, 8:131662–131682, 2020.

[93] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measur-

ing stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

[94] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2010.

[95] Hadas Orgad and Yonatan Belinkov. Debiasing nlp models without demographic information. *arXiv preprint arXiv:2212.10563*, 2022.

[96] Organization for Economic Co-operation and Development. Consumer price index: All items for the united states [usacpiallminmei]. `https://fred.stlouisfed.org/series/USACPIALLMINMEI`, n.d. [Online; accessed 2023-05-03]. Retrieved from FRED, Federal Reserve Bank of St. Louis.

[97] Soyoung Park, Sharon Strover, Jaewon Choi, and MacKenzie Schnell. Mind games: A temporal sentiment analysis of the political messages of the internet research agency on facebook and twitter. *New Media & Society*, 25(3):463–484, 2023.

[98] Josh Pasek, H Yanna Yan, Frederick G Conrad, Frank Newport, and Stephanie Marken. The stability of economic correlations over time: identifying conditions under which survey tracking polls and twitter sentiment yield similar conclusions. *Public Opinion Quarterly*, 82(3):470–492, 2018.

[99] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

[100] Aidan Pine and Mark Turin. Language revitalization, 2017.

[101] Tracy Qian, Andy Xie, and Camille Bruckmann. Sensitivity analysis on transferred neural architectures of bert and gpt-2 for financial sentiment analysis. *arXiv preprint arXiv:2207.03037*, 2022.

[102] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[103] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[104] Reddit. Reddit markdown. `https://www.reddit.com/wiki/markdown/`, n.d. [Online; accessed 2023-04-25]. Quotes and references

used as examples: Douglas Adams, Rumi, Marcus Aurelius, Miguel de Cervantes, Thomas Paine, eden ahbez, Aeschylus, Chuck Palahniuk, Sylvia Plath. Last revised by LanterneRougeOG - 4 years ago.

[105] The European Business Review. Statistics about reddit – one forum to rule them all. February 2023. Accessed: 2023-05-16.

[106] Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. A multilingual and multidomain study on dialog act recognition using character-level tokenization. *Information*, 10(3):94, 2019.

[107] John M Roberts. New keynesian economics and the phillips curve. *Journal of money, credit and banking*, 27(4):975–984, 1995.

[108] Margarita Rodríguez-Ibánez, Antonio Casánez-Ventura, Félix Castejón-Mateos, and Pedro-Manuel Cuenca-Jiménez. A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, page 119862, 2023.

[109] David Romer. Advanced macroeconomics fourth edition, 2011.

[110] David Rotman. How chatgpt could revolutionize the economy—and help decide what it looks like. `https://www.technologyreview.com/2023/03/25/1070275/chatgpt-revolutionize-economy-decide-what-looks-like/`, 3 2023. Accessed: 2023-04-25.

[111] r/pushshift. Removal request form - please put your removal requests here! `https://www.reddit.com/r/pushshift/comments/10yj803/removal_request_form_please_put_your_removal/`, 2010. [Online; accessed 2023-03-30].

[112] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023.

[113] Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424, 2021.

[114] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

[115] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[116] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

[117] Nakatani Shuyo. Language detection library for java, 2010.

[118] Tajinder Singh and Madhu Kumari. Role of text pre-processing in twitter sentiment analysis. *Procedia Computer Science*, 89:549–554, 2016.

[119] Nicholas S Souleles. Consumer sentiment: Its rationality and usefulness in forecasting expenditure-evidence from the michigan micro data, 2001.

[120] Nicholas S Souleles. Expectations, heterogeneous forecast errors, and consumption: Micro evidence from the michigan consumer sentiment surveys. *Journal of Money, Credit and Banking*, pages 39–72, 2004.

[121] Arthur Spirling. How openai's chatgpt could change science publishing forever. `https://www.nature.com/articles/d41586-023-01295-4`, 4 2023. Accessed: 2023-04-25.

[122] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

[123] TensorFlow. Subwords tokenizer, 2023. Accessed: 2023-04-25.

[124] Richard H Thaler. From homo economicus to homo sapiens. *Journal of economic perspectives*, 14(1):133–141, 2000.

[125] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[126] Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the international AAAI conference on web and social media*, volume 4, pages 178–185, 2010.

[127] Stephen J Turnovsky. Empirical evidence on the formation of price expectations. *Journal of the American Statistical Association*, 65(332):1441–1454, 1970.

[128] University of Michigan, Survey Research Center. Surveys of consumers. `http://www.sca.isr.umich.edu/`, 2023. Accessed: 2023-04-25.

[129] U.S. Bureau of Labor Statistics. Unemployment rate [unrate]. `https://fred.stlouisfed.org/series/UNRATE`, 2023. [Online; accessed 2023-05-03]. Retrieved from FRED, Federal Reserve Bank of St. Louis.

[130] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[131] Gang Wang, Tianyi Wang, Bolun Wang, Divya Sambasivan, Zengbin Zhang, Haitao Zheng, and Ben Y Zhao. Crowds on wall street: Extracting value from collaborative investing platforms. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 17–30, 2015.

[132] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[133] Stephen D Williamson and Anisha Sharma. *Macroeconomics*. Pearson New York, 2014.

[134] Stefan Wojcik and Adam Hughes. Sizing up twitter users. *PEW research center*, 24:1–23, 2019.

[135] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[136] Jeffrey M Wooldridge. *Introductory econometrics: A modern approach*. Cengage learning, 2015.

[137] World Health Organization. Covid-19. `https://www.who.int/europe/emergencies/situations/covid-19`, n.d. [Online; accessed 2023-04-25]. Retrieved from World Health Organization.

[138] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

[139] Frank Z Xing, Erik Cambria, and Roy E Welsch. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73, 2018.

[140] Ashima Yadav and Dinesh Kumar Vishwakarma. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385, 2020.

[141] Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120, 2020.

Appendix B

Samples of Sentences Used to Test Economic Accuracy

Each group of statements consists of a rephrased correct sentence, a rephrased incorrect sentence, and the statement extracted from the corpus supporting the true statement from the textbook and Investopedia datasets respectively.

B.1 Textbook Dataset Sentences

1. a. *Correct Statement:* A decrease in prices raises unemployment.

   b. *Incorrect Statement:* A decrease in prices lowers unemployment.

   c. *Supporting Statement:* Prior to a period of deflation, a liquidity trap occurs, which is a period where there is zero demand for investment in bonds, and people hoard cash because they anticipate a period of deflation or war. The Pigou effect proposes a mechanism to escape this trap. According to the theory, price levels and employment fall, and unemployment rises. As price levels decline, real balances increase, and by the Pigou effect, consumption is stimulated in the economy. The Pigou effect is also known as the "real balance effect."

2. a. *Correct Statement:* A higher ROE is better.

   b. *Incorrect Statement:* A higher ROE is worse.

   c. *Supporting Statement:* ROE is calculated by dividing a company's net income by total shareholders' equity. Although a higher ROE figure is generally a better ROE figure, investors should exercise caution when a very high ROE is a result of extremely high financial leverage. This is one reason why it is also important to consider a pharma company's debt and liquidity situation.

3. a. *Correct Statement:* Bonds are low-risk investments.

   b. *Incorrect Statement:* Bonds are high-risk investments.

   c. *Supporting Statement:* Bonds tend to be stable, lower-risk investments that provide the opportunity both for interest income and price appreciation. It is recommended that a diversified portfolio have some allocation to bonds, with more weight to bonds as one's time horizon shortens.

4. a. *Correct Statement:* Financial accounting is used to provide information to people outside the company.

   b. *Incorrect Statement:* Financial accounting is used to provide information to people inside the company.

   c. *Supporting Statement:* Managerial accounting is different from financial accounting in that financial accounting is centered on providing quarterly or yearly financial information to investors, shareholders, creditors, and others outside the organization. Conversely, managerial accounting is used internally to make efficiency improvements within the company.

B.2 Investopedia Dataset Sentences

1. a. *Correct Statement:* Both consumption and investment are procyclical.

   b. *Incorrect Statement:* Neither consumption nor investment are procyclical.

   c. *Supporting Statement:* As well, when the interest rate target falls, this causes investment and consumption to rise, so that investment, I, and consumption, C, are procyclical, as is true for the data.

2. a. *Correct Statement:* A drop in the world real interest rate increases domestic investment and has an ambiguous effect on domestic consumption.

   b. *Incorrect Statement:* A decrease in the world real interest rate increases domestic investment and has a negative effect on domestic consumption.

   c. *Supporting Statement:* Thus, a decrease in the world real interest rate in our model acts to increase domestic investment, has an ambiguous effect on domestic consumption, and reduces the current account surplus.

3. a. *Correct Statement:* Real interest rate increasing means domestic consumption expenditures are likely to fall.

   b. *Incorrect Statement:* Real interest rate increasing means domestic consumption expenditures are likely to rise.

   c. *Supporting Statement:* The increase in the real interest rate causes domestic consumption expenditures and investment expenditures to fall, though the increase in current income causes consumption to rise.

4. a. *Correct Statement:* The decrease in the real interest rate then increases the demand for consumption goods and investment goods, and so firms provide extra output given that in the short run prices remain constant.

   b. *Incorrect Statement:* The decrease in the real interest rate then increases the demand for consumption goods and investment goods, and so firms provide extra output given that in the long run prices remain fixed.

   c. *Supporting Statement:* The decrease in the real interest rate then increases the demand for consumption goods and investment goods, and so firms supply the extra output given that prices are fixed in the short run.

Appendix C

Supplementary Experiments

To better elucidate processes and results of experiments which informed decisions within the thesis body, and to further highlight areas where future research could help enhance the model's performance, this section of the Appendix shows results of supplementary experiments.

C.1 Examining the Model's Sensitivity to Changes in Time and Quantities of Variables

As conveyed in subsection 5.3.4 of the Experiments chapter, in order to make effective comparisons between the results obtained through the fine-tuning, survey results and actual economic variable values, it would have been ideal for making comparisons based on perplexities of sentences that would specify predictions both concerning the time period and quantity of variables under consideration so that time frames and overall semantic nature of sentences would match surveys and actual variables' values obtained better.

To uncover such results, an experiment testing the model's sensitivity to time and another focusing on sensitivity to numerical values were undertaken.

C.1.1 Testing the GPT-2 Model's Sensitivity to Time With Respect to Forecasts

In an attempt to consider the model's sensitivity to time, the pairs of opposing sentences which perplexity was subsequently evaluated through the perplexity ratio, were drafted based on templates displayed in Figure C.1.

As shown in Figure C.2, the trends of perplexity ratios of sentences pertaining to different time windows are very similar. They experience both peaks, such as in the second and fourth quarters of 2020 or the third quarter of 2021, for instance, and depressions, for example, as shown by values in the third quarter of 2020 or the second quarters of 2021 and 2022. Overall, this

'In a week, unemployment will [decrease/increase].'
'In a year, unemployment will [decrease/increase].'
'In a few years, unemployment will [decrease/increase].'
'In the future, unemployment will [decrease/increase].'
'Unemployment will [decrease/increase].'

Figure C.1: Templates for constructing sentences about unemployment increasing or decreasing in the future.

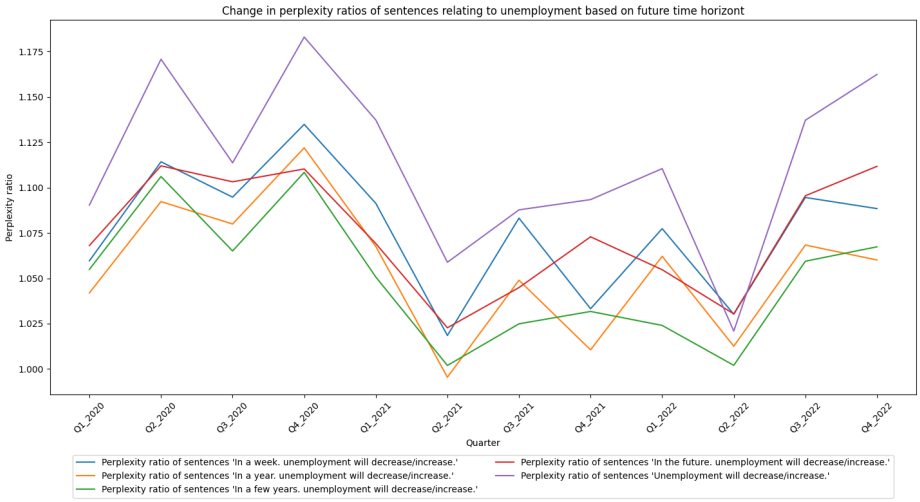


Figure C.2: Graph illustrating perplexity ratio values of sentences making predictions about unemployment over the course of several time frames. As conveyed by the graph, trends of perplexity values seem highly correlated, in spite of each referring to a different point in time, suggesting the model is not yet equipped to make robust estimates with respect to time.

leads to the conclusion that perhaps the models at hand are not sensitive enough to changes in time frames, as there is a clear lack of variability in the curves' trends. Interestingly, the perplexity ratios of sentences referring to non-deterministic time periods, i.e., sentences including the expressions 'in a year' and 'in the future', have a distinctly similar pattern in trends in comparison to that of sentences referring to a week's time and years. On top of these lies the purple curve showing the perplexity ratios of the sentences 'Unemployment will decrease.' and 'Unemployment will increase.', which in its trend stands out from the rest ever so slightly. It was posited that given it is the most abstract out of all of them, it may hold more value, as it is likely that an indeterministic reference about the future value of an economic variable such as unemployment would be more likely than specifically phrased prognoses about particular points in time.

The fact the model shows signs of insensitivity to time frames could be due to many reasons. For one, there could be a misstep somewhere along the fine-tuning pipeline. Second, this could be because of the sheer lack of comments referring to specific points in time and the value of the unemployment rate

with respect to them. Third, it could be because the model is too small to grasp this facet of the datasets. Either way, steps could be taken to distinguish which of the reasons could be true. For instance, datasets could be injected with artificial sentences specifying that at a certain point in time, the unemployment rate will be high. For example, 'In a week, unemployment will be extremely high.'. Then the same sentences' perplexity ratios as shown in Figure C.2 could be verified again for different numbers of injections (for example, for 10, 100, 1000 or more injections of the 'In a week, unemployment will be extremely high.') of the sentence. The model's sensitivity could thereby be verified, which could shed light on the reasons as to why the model was not perceptive enough to previously mentioned formulations.

Unfortunately, these findings happened at a later stage in the research, and more thorough checks of the reasons why the model lacked sensitivity to time horizons have yet to be undertaken. That is why, to offer some indication of the possible opinions about economic variables' trends in a year's time, the least specific sentence template time-wise: 'In the economy, [inflation/interest rate/unemployment] will be [low/high].' was chosen as the basis for paraphrases used in section 5.3.4 of the Results chapter. This is a tentative decision to see whether there is a possible implicit correlation between the general sentence structure chosen and values one year in the future.

C.1.2 Testing the GPT-2 Model's Sensitivity to Quantities With Respect to Forecasts

To assess whether the model is sensitive to specific quantities of economic variables was the purpose of the supplementary experiment within this part of the Appendix.

'Unemployment will increase by [less/more] than 1%.'
'Unemployment will increase by [less/more] than 2%.'
'Unemployment will increase by [less/more] than 3%.'
'Unemployment will increase by [less/more] than 0.5%.'
'Unemployment will increase by [less/more] than 5%.'

Figure C.3: Templates for constructing sentences about the degree of unemployment increase in the future.

In order to consider the model's sensitivity to numerical quantities of economic variables forecasted, the pairs of opposing sentences displayed in Figure C.3 were drafted, the perplexity of which was subsequently evaluated.

Future estimates would ideally compare real-world quantities and numerical predictions by the SPF to values obtained from the fine-tuned model. Unfortunately, however, results presented by Figure C.3 suggest the model's incapability of effectively capturing numerically precise predictions due to the correlated peaks of trend lines pertaining to various values of the unemploy-

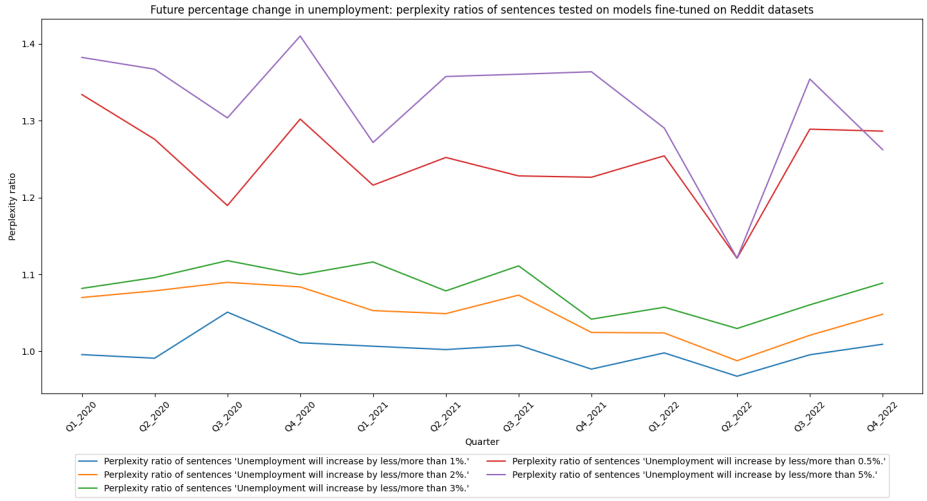


Figure C.4: Graph conveying perplexity ratios of sentences referring to various degrees of change in unemployment in the future. Despite varying numerical values, trends of curves follow very similar patters, hinting at the current model's incapability to effectively process numerical changes in economic variables.

ment rate. Even though yearly changes in unemployment over the majority of the 2020-2022 period did not fluctuate by more than 5%, perplexity ratios across the board seem to suggest extremely similar trends, and furthermore, the fact that sentences with the quantity 0.5% showed more similar perplexities and trends to those mentioning 5% than either to quantities such as 2-3%, is further evidence of either lack of relevant data suggesting posters on Reddit divulge their prognoses on specific numerical values of unemployment in the future, or the model's incapability of effectively processing numerical data. Either way, it underlines the model's results' inconsistencies.

When it comes to reasons leading to the lack of predictive power of the model both in this case and in the one concerning time sensitivity mentioned in section C.1.2 of the Appendix, it could be hypothesised that it is due to the lack of relevant data within the datasets, as manual inspection of datasets revealed, that for example within the dataset corresponding to the first quarter of 2020, there were less than 100 comments including words 'unemployment' and either 'percentage' or '%', and 'year', suggesting further datasets are needed for improved accuracy.

These results indicate that given the datasets it was fine-tuned on, the model is not yet equipped to handle specific numerical quantities, which is why sentences selected for testing within the Experiments and Results chapters refrained from specifying quantities by which economic metrics would have risen or dropped in value, and merely referred to whether they rose or decreased.

Given these findings, it is clear that further steps could be taken to enhance

the model's understanding of both time and quantity. Additional data, especially with specific time-based and numerical predictions, might improve the model's capabilities in these areas. Also, a more extensive or differently fine-tuned model might be better at picking up these nuances.

Appendix D

Glossary

Term	Meaning
BPE	Byte-Pair Encoding
CPI	Consumer Price Index
CNN	Convolutional Neural Network
EA	Economic Accuracy
FNN	Feedforward Neural Network
FRED	Federal Reserve Economic Data
GELU	Gaussian Error Linear Unit
GPT	Generative Pretrained Transformer
GPT-2	Generative Pretrained Transformer 2
GPT-3	Generative Pretrained Transformer 3
ICS	Index of Consumer Sentiment
LLM	Large Language Model
LSTM	Long Short-Term Memory
MLE	Maximum Likelihood Estimation
NER	Named Entity Recognition
NLP	Natural Language Processing
PCE	Personal Consumption Expenditures Price
PPL	Perplexity
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
SPF	Survey of Professional Forecasters
SVM	Support Vector Machine