Analyzing U.S. Congressional Tweets with OpenAI GPT-3

Gokul Prasad and Annie Fan

UC San Diego
DSC180B
Suraj Rampure & Molly Roberts

Abstract. The China Data Lab at UC San Diego has been investigating the sentiment of Congress towards China as expressed in Tweets posted by Congress members. Currently, the sentiment analysis is done manually and the lab seeks to automate this process. Therefore, this project aims to employ machine learning models to classify: (1) whether the Tweets directly relate to China and (2) if relevant, what sentiment the Tweets belong to. Initially, this project found that classifying Tweets with a Bernoulli Naive Bayes model and scoring Tweets with a Ridge or Random Forest Regressor provided the best results. However, the second portion of this project looks to use the GPT-3 Large Language Model as a substantially better alternative in its ability to understand and classify text.

Keywords— Large Language Models, GPT-3, China, Sentiment Analysis, Classification

1. Introduction

The growth in China's socio-political and economic power has established it as a significant challenge to the United States' position as the dominant global superpower. Throughout recent history, the two countries have been allies and rivals, often both at the same time. As a result, the country has become a highly debated topic among U.S. politicians, with both Democrats and Republicans using discussions of China to mobilize their voting base. The use of social media platforms, such as Twitter, provides politicians with a global audience beyond just their constituents. Observing the discourse of American politicians on China on these platforms can therefore offer valuable insights into current political trends and inform future policy making.

1.1. Literature Review

Past work in this field has been conducted by the University of California San Diego's China Data Lab who created the dataset used in this paper as part of their exploration into "how members of Congress tweet about China" [2]. Their work was concerned with a variety of features, like measuring Congressional sentiment towards China, where tweets were originating from, and what aspects of addressing China does Congress have bipartisanship over. The work of the China Data Lab was conducted with one notable limitation: much of the building of the data was done with manual labor done to tag the tweet's relevance towards China and assign it a sentiment score on a scale of 1 – 5. This exacerbated the amount of time the task should take and kept the dataset from growing in real time as more Tweets are posted, which is a problem the China Data Lab wants to address for future analyses. This paper will address the same two tasks as before – finding relevancy and scoring sentiment – using trained and optimized machine learning models that reshape text as easily-digestible vectors of data that can be used in a variety of language processing techniques. The models will perform at higher speeds and accuracy than completing by hand.

Plenty of past work has explored the intersection of social media, politics, and machine learning. Sentiment analysis of social media, and Twitter in specific, has been a facet of machine learning exploration for a long while. State-of-the-art natural language processing models can accurately classify text at unbelievable speeds, and projects like OpenAI GPT-3 or HuggingFace can generate or summarize text with uncanny human qualities. With respect to sentiment analysis for Twitter in specific, Sarlan, Nadam, and Basri argue that applying machine learning techniques and solutions is "more suitable for Twitter" than other styles of sentiment analysis and opinion mining in text data [3]. Furthermore, the language of social media can vary quite heavily based on who sends a message and in what context the message was sent, so focusing on vectorizing these Tweets can sharply cut down on the time needed to classify. Twitter itself has a long history with politics, ever since the Obama campaign used it to generate funds in 2008 to great success. Some use it to boost their electoral chances, others use it to attack or promote certain views, but most importantly, Twitter allows all politicians "to discuss their political agenda... for free" [1]. Whether it is the President or a local judge, whether the person Tweeting has raised millions in fundraising or none at all, the ability to Tweet allows them to reach a broad audience with their viewpoints. Social platforms provide an unbelievable spread of information and audience to users.

1.2. Data

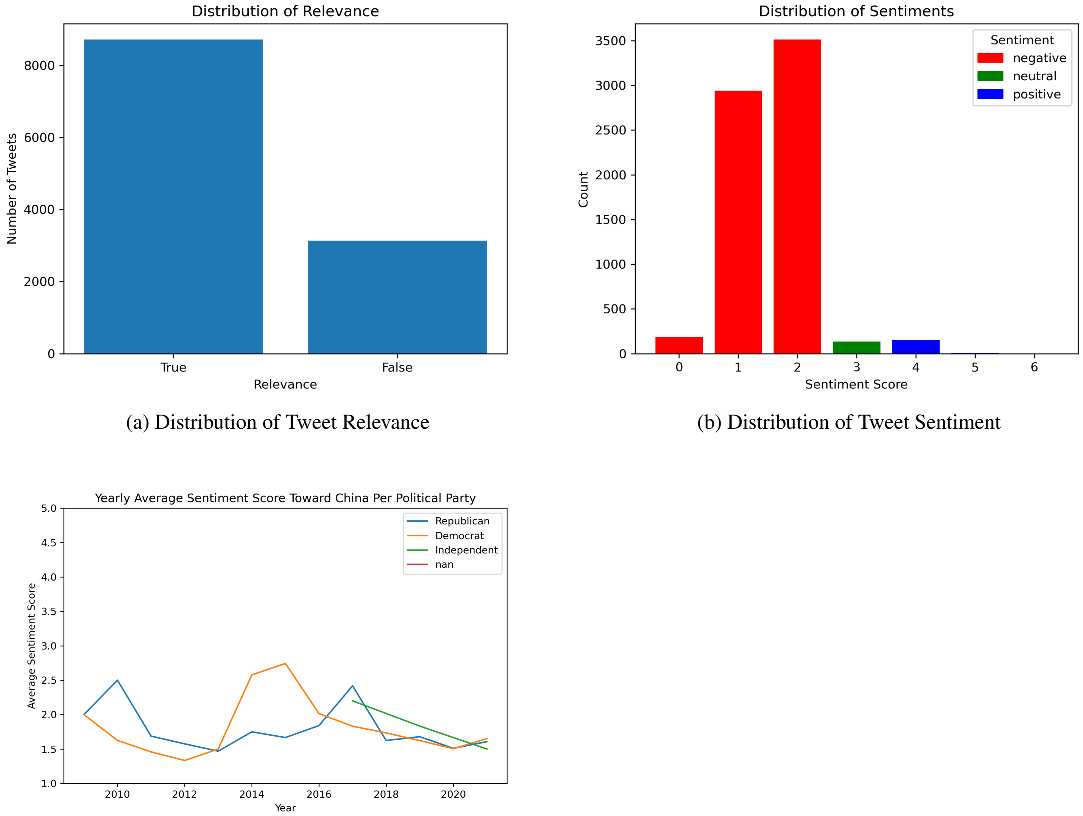
Our team used the data collected by The China Data Lab from Twitter's API. The dataset contains Congress Tweets about China, Canada, and Iran. Each row of the dataset contains the text of a Tweet, the category that the Tweet falls into, the Tweet's sentiment score, the politician of the Tweet, the politician's information, and the id of the Tweet. Tweets about Iran and Canada, which are sampled from all the collected Tweets for benchmark and comparison purposes, are approximately 20% of the dataset. The team mainly worked with Congressional Tweets about China in this project.

2. Method

Preparation for the application of an LLM for both classification and sentiment analysis required deep consideration of which model would suit this project best, as well as deep dives into prompt engineering.

2.1. Exploratory Data Analysis

Prior to applying an LLM, our team first needed to understand the distribution of our data. The dataset is highly imbalanced in both Tweet relevance and sentiments. Figure 1(a) shows that there are 8718 Tweets that are relevant to China while only 3130 Tweets are irrelevant. In Figure 1(b), the majority of Tweets in the dataset have negative sentiments toward China and there is a very small number of neutral and positive Tweets about China. Figure 1(c) shows the yearly trend of sentiment scores toward China. Every year, the average score of China-related Tweets from all Congress members (black line) is between 1 and 3, which is in the negative sentiment range. Republicans (blue line) and Democrats (orange line) take turns having higher sentiment scores than the other party. For example, Democrats on average had higher sentiment scores toward China than Republicans did from 2014 to 2016, while Republicans had higher sentiment scores from 2010 to 2012 and in 2017. After 2018, the two parties had similar levels of sentiments every year.



(a) Distribution of Tweet Relevance

(b) Distribution of Tweet Sentiment

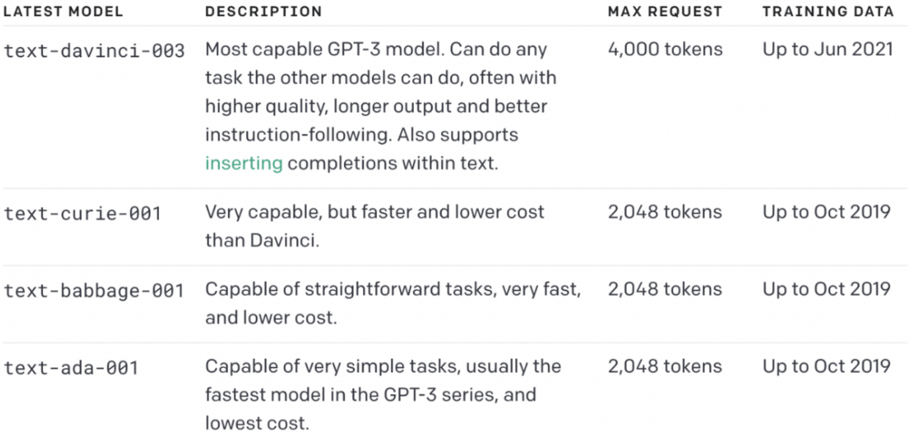
(c) Sentiment Yearly Trend

Fig. 1: EDA Plots

2.2. Model Selection

The incredible advancements in publicly-accessible language models meant that selecting an LLM to use was an incredibly important first step in this process. While GPT-3 reigns supreme and was eventually the final choice, there were other capable options that required careful consideration. Most prominently, there is the BLOOM model, or BigScience Large Open-science Open-access Multilingual Language Model, developed by Hugging Face Inc., based on existing GPT-2 architecture and deployed as a direct competitor to GPT-3. However, accessing and building a pathway from our dataset to model input was significantly easier with GPT-3, and given the models' similar performances across generation, comprehension, and output quality, GPT-3 was chosen as the LLM for this project.

Within GPT-3, however, there are several sub-model options: Ada, Babbage, Curie, or Davinci. While all four options perform well with text comprehension and classification, Davinci was trained on more recent data, as well as being significantly more capable of understanding specific directions. Given that both the classification and sentiment analysis tasks would require very specific outputs, Davinci was selected as the model of choice for both tasks.



LATEST MODEL	DESCRIPTION	MAX REQUEST	TRAINING DATA
`text-davinci-003`	Most capable GPT-3 model. Can do any task the other models can do, often with higher quality, longer output and better instruction-following. Also supports inserting completions within text.	4,000 tokens	Up to Jun 2021
`text-curie-001`	Very capable, but faster and lower cost than Davinci.	2,048 tokens	Up to Oct 2019
`text-babbage-001`	Capable of straightforward tasks, very fast, and lower cost.	2,048 tokens	Up to Oct 2019
`text-ada-001`	Capable of very simple tasks, usually the fastest model in the GPT-3 series, and lowest cost.	2,048 tokens	Up to Oct 2019

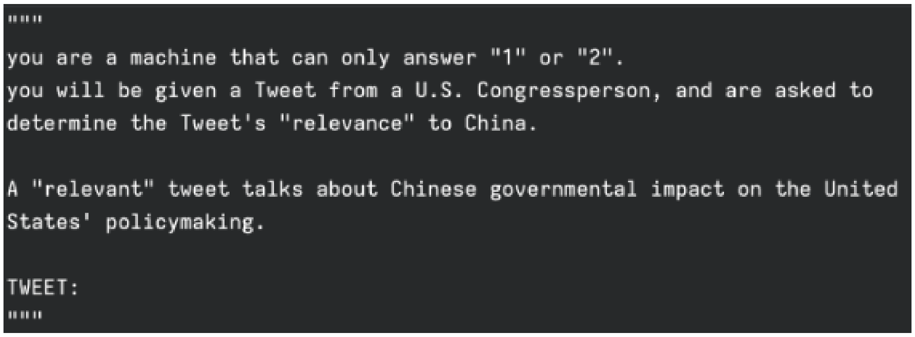
Fig. 2: Model descriptions from OpenAI documentation

2.3. Tweet Relevance Classification

Thanks to the simplicity of accessing GPT-3 via a simple API call and the model's underlying ability to understand unconventional text, like Tweets, there was no cleaning required of the Tweets like there would be in the usage of a standard supervised machine learning model. As such, much of the work done was creating and improving the prompts used to generate outputs from GPT-3 by modifying them with concepts like:

1. One-shot Learning: Providing one example per class for the model to learn from
2. Few-shot Learning: Providing multiple examples per class for the model to learn from

The initial prompt passed for classification was very simplistic and did not include any in-context learning, as a baseline evaluator of GPT's performance:



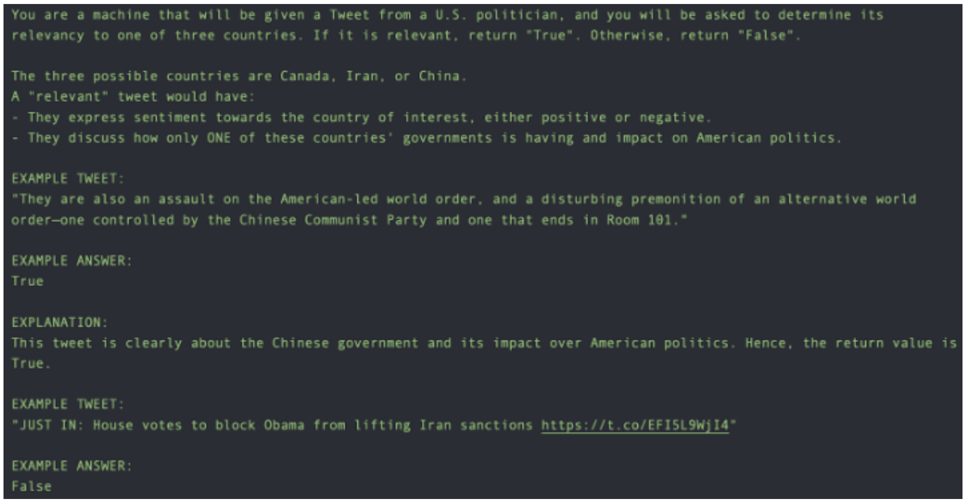
```
"""
you are a machine that can only answer "1" or "2".
you will be given a Tweet from a U.S. Congressperson, and are asked to
determine the Tweet's "relevance" to China.

A "relevant" tweet talks about Chinese governmental impact on the United
States' policymaking.

TWEET:
"""
```

Fig. 3: Initial Prompt

After running through the evaluation process, the prompt was then improved to include in-context learning to show GPT-3 what a correct response was, as well as the reasoning behind the answer, by including the following examples:



```
You are a machine that will be given a Tweet from a U.S. politician, and you will be asked to determine its
relevancy to one of three countries. If it is relevant, return "True". Otherwise, return "False".

The three possible countries are Canada, Iran, or China.
A "relevant" tweet would have:
- They express sentiment towards the country of interest, either positive or negative.
- They discuss how only ONE of these countries' governments is having and impact on American politics.

EXAMPLE TWEET:
"They are also an assault on the American-led world order, and a disturbing premonition of an alternative world
order—one controlled by the Chinese Communist Party and one that ends in Room 101."

EXAMPLE ANSWER:
True

EXPLANATION:
This tweet is clearly about the Chinese government and its impact over American politics. Hence, the return value is
True.

EXAMPLE TWEET:
"JUST IN: House votes to block Obama from lifting Iran sanctions https://t.co/EFI5L9WjI4"

EXAMPLE ANSWER:
False
```

Fig. 4: Refined Prompt

Here, the model is given an expanded purview to digesting Tweets about Iran and Canada beyond just China; although the overall report is focused solely on Chinese-centric Tweets, providing more examples to learn from will only improve the model's performance and understanding.

These prompts were applied to repeated samples of 100 Tweets from the dataset, the answers were cleaned to remove any extra information, and then compared to the man-made answers. For this particular task, the chosen metrics was accuracy, and confusion matrices were generated to visualize "pain points" for the model; these were common areas of disagreements between the human and the model.

2.4. Tweet Sentiment Classification

Data Processing In the original dataset, sentiments of Tweets are evaluated on a five-point scale, with 1 being very negative, 3 being neutral, and 5 being very positive. To reduce the complexity of sentiment classification, we reduced sentiments into three categories: negative for Tweets with sentiment score 1 and 2, neutral for Tweets with a sentiment score of 3, and positive for Tweets with score 4 and 5. In addition, some Tweets from the original dataset have multiple sentiment scores or averaged scores because human coders had disagreements in evaluating the sentiment of these Tweets. These Tweets are excluded from the training process in this study in order to reduce ambiguity and complexity of classification tasks.

Running GPT-3 Similar to the relevance classification task, a prompt with the classification task specification, examples of how the task is expected to be done, and a Tweet to be evaluated is sent to GPT-3. The prompt looks like the following:

> "Determine the given tweet's sentiment toward China. Return either positive, neutral, or negative.
>
> Example 1: "The humanitarian, security and health threats personified in the coronavirus are being exacerbated by authoritarian socialist policies and the dishonestly of foreign aggressors and abusers, like China. https://t.co/CKAtP9qOUT coronavirus".
> Answer: negative.
> Reason: The tweet uses explicit negative sentiment towards China through the words such as "dishonest", "aggressor", and "abuser".
>
> Given these examples, value the following tweet:"

The actual prompt we used in training includes one example for each sentiment category in order to provide enough examples and instructions about our expected output for GPT-3. The model would output its evaluation of the input Tweet's sentiment toward China, following the format in the examples provided in the prompt. The team then removed information in the model's output by extracting the evaluated sentiment - either positive, neutral, or negative. The metric the team used to measure the model's performance is accuracy of the model's classification.

In order to compare the GPT-3 model with machine learning models the team previously experimented with, our team also trained a Random Forest Classifier on a balanced dataset - that is, we downsampled the number of negative Tweets to 135, which is the number of positive Tweets. The train-test split ratio used is 75% to 25%. We used default model parameters from Sklearn.

Data Sampling Again, the prompt is run on repeated samples of Tweets from the dataset in order to optimize our usage of GPT-3. For each trial, the team sampled 10-30 Tweets per sentiment category and ran the model on the sampled data Tweet by Tweet.

3. Result

3.1. Relevance

The chosen metric of evaluation for the model was accuracy, and further visualization of model output via a confusion matrix. In contrast to last quarter where the metrics included accuracy, precision, recall, and F-1 scores, the lack of imbalance issues in GPT-3 rendered calculating metrics beyond accuracy somewhat pointless since the model was not being 'trained' in the same sense as a supervised model. A detailed table of the model's performance by prompt can be seen below:

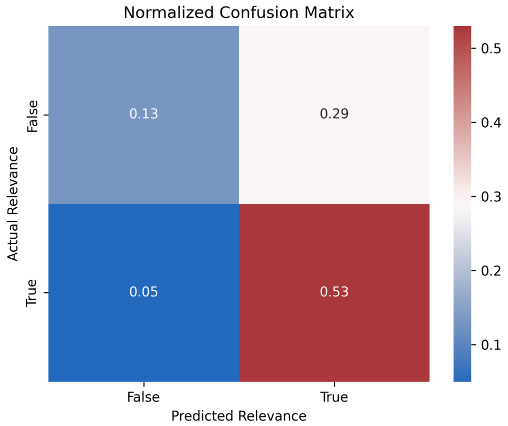


Fig. 5: Relevance Classification Confusion Matrix

Trial	Prompt Type	Peak Accuracy
1	Command only	43%
2	Command + 2 examples	62%
3	Command + 3 examples	65%
4	Command + 3 examples + 2 ‘pain-points’	73%
5	Command + 3 examples + 2 ‘pain-points’ + basic contextual information	75%

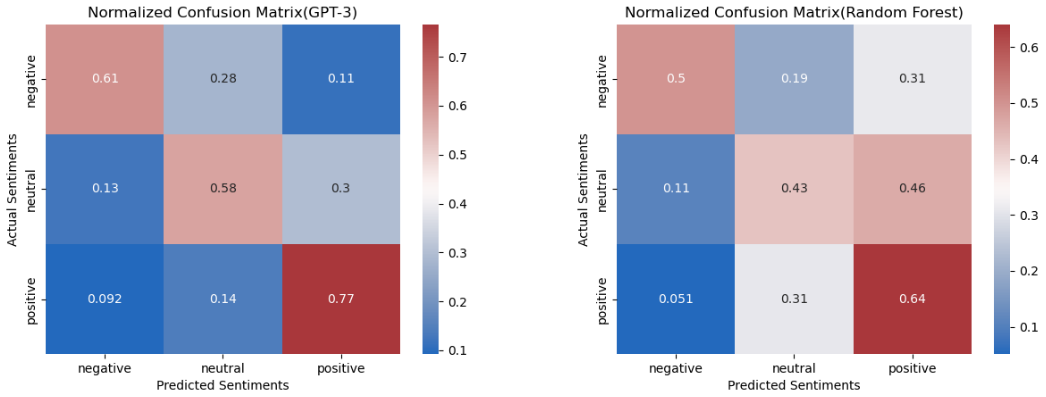
(Each prompt was run several fifteen times on samples of 100 Tweets at random. More information on the prompts can be found in the appendix.)

An example confusion matrix from Trial 4 can be seen below, demonstrating an accuracy of 68%. Even though the prompt was later tuned and accuracy later improved, the confusion matrix reveals a common issue to the GPT-3 model that persisted throughout all five trials: Tweets that were marked by the human encoder as irrelevant were often judged as relevant by the model.

In comparison with the results we had obtained from the first quarter, these results fall well short of expected outputs. While the GPT-3 model did not have to contend with the problems of balancing and class distribution that the supervised models, it still struggled with some Tweets based on their interpretation by the human encoders.

3.2. Sentiment

On each training trial on a sample of 100 Tweets, the GPT-3 model can achieve between 60% and 70% accuracy in classifying Tweet sentiments. According to the confusion matrix in Figure 6(a), the model performed best in classifying Tweets with positive sentiment. The model had the lowest accuracy in classifying neutral Tweets and struggled in distinguishing neutral Tweets with positive and negative Tweets. On the other hand, the Random Forest Classifier achieved approximately 55% accuracy. From the confusion matrix in Figure 6(b), Random Forest’s accuracy of classifying each sentiment class was all lower than the performance of the GPT-3 model.



(a) GPT-3 Result Confusion Matrix

(b) Random Forest Classifer Result Confusion Matrix

Fig. 6: Sentiment Classification Results

Sentiment Score Labeled by Human Coders	GPT-3 Accuracy
1 (very negative)	73%
2 (negative)	49%
3 (neutral)	58%
4 (positive)	78%
5 (very positive)	67%

4. Discussion

4.1. GPT-3 in Classifying Tweet Relevance

In comparison with the results from Quarter 1, the output of GPT-3 fell well short in the measured metrics. Specifically, while the supervised Naive Bayes classifier could consistently reach accuracies greater than 90%, the LLM only achieved a peak of 75%, and even then after extended prompt engineering. While this was disappointing, it is important to consider why these discrepancies are occurring, because LLMs like GPT-3 have proven to be far more capable than traditional machine learning models in a variety of usage cases beyond simple classification. Examining the consistent pain-point of a Tweet categorized as irrelevant while the model predicts relevance, it can be seen that GPT-3 is simply suffering from a natural human difference in understanding context and intent in text. While the supervised ML models are simply fitting themselves onto a dataset, the accuracy is high since it takes the provided labels as absolute truth, and so learns to embed and read the Tweets through that singular lens. However, GPT-3, being an LLM, actually understands the Tweet on its own, and much like the human encoders who created the dataset of Tweets, can have a different reading of some particular Tweet that causes it to output a different label. Class imbalances are of no concern since GPT-3 does not necessarily need to be trained, and can be taught to evaluate or consider information in its own context. As such, despite the poorer comparative performance, it seems clear that this initial experimentation gives weight to the argument that GPT-3 has a stronger future in this field; its malleability to any text can transform these problems from basic classification to deeper text comprehension and analysis on its own.

4.2. GPT-3 in Classifying Tweet Sentiment

The GPT-3 model yields better performance in classifying sentiment than the Random Forest Classifier. In order to minimize the bias caused by the imbalanced dataset and optimize the performance of Random Forest Classifier, we needed to balance the dataset by downsampling the majority class and excluding approximately 90% of the data from the training process. As a result, training machine learning classifiers such as Random Forest is costly given the limitation of our imbalanced dataset. Comparatively, GPT-3 is a more appropriate option for our data because it evaluates each Tweet independently and thus is not affected by the imbalanced data.

During the process of applying GPT-3 in classifying sentiment, GPT-3 showed good performance in classifying Tweets that expressed explicit sentiments toward China. However, GPT-3 often struggled with currently classifying Tweets that include implicit sentiment toward China and confused these Tweets with neutral Tweets. When evaluating Tweets with implicit sentiments and neutral Tweets, GPT-3 had significant disagreements with the evaluation from the human grader. For example, one of the misclassified Tweets with implicit sentiment is “Your GhostFleet read of the week is the new @CNASdc report on China’s pursuit of quantum dominance. This is a vital competition for technological superiority we cannot afford to lose.” The human coder evaluated the sentiment of this Tweet as negative because the Tweet is about the competition against China and thus it implies negative sentiment toward China. GPT-3’s output to this Tweet is neutral and its reasoning is that “The tweet expresses the importance of competition between China and the US for technological superiority, but does not explicitly express any opinion about China.” As a result, despite GPT-3’s capability in understanding text sentiments, it still has substantial differences with human interpretation.

4.3. Conclusion of Project

The application of GPT-3 to this problem suffered from the same problems that plagued the first quarter’s results; namely, that this entire problem relies heavily on subjectivity and human interpretation. Disagreements between the human and programmatic classifiers are likely to never be fully worked out – LLMs learn from human input, and so natural differences in interpretation or preexisting beliefs can be a major influence on the result. However, this process shows that there definitely is a place for LLM-usage in Tweet analysis, and that with more time, the result could be improved. The simplification to the process and the removal of imbalancing issues clearly made this a more powerful option than using supervised ML models.

4.4. The Future of Language Modeling

Large Language Models like GPT-3 are undoubtedly the future of language processing and modeling: one only needs to look how quickly ChatGPT was popularized and made accessible to the broader public, as well as the quick creation of competing models like Bard by Google or LLaMa by Meta.. However, it is important to note that these models are not a magic solution and do require significant subject matter expertise to achieve consistent high-quality outputs.

One of the primary challenges with LLMs is their ability to generate coherent and accurate outputs that align with the specific context and requirements of the task at hand. While these models have the ability to learn from vast amounts of data and generate impressive outputs, their effectiveness largely depends on how well they are fine-tuned to the specific task and the domain in which they are being applied. This is where subject matter expertise comes into play.

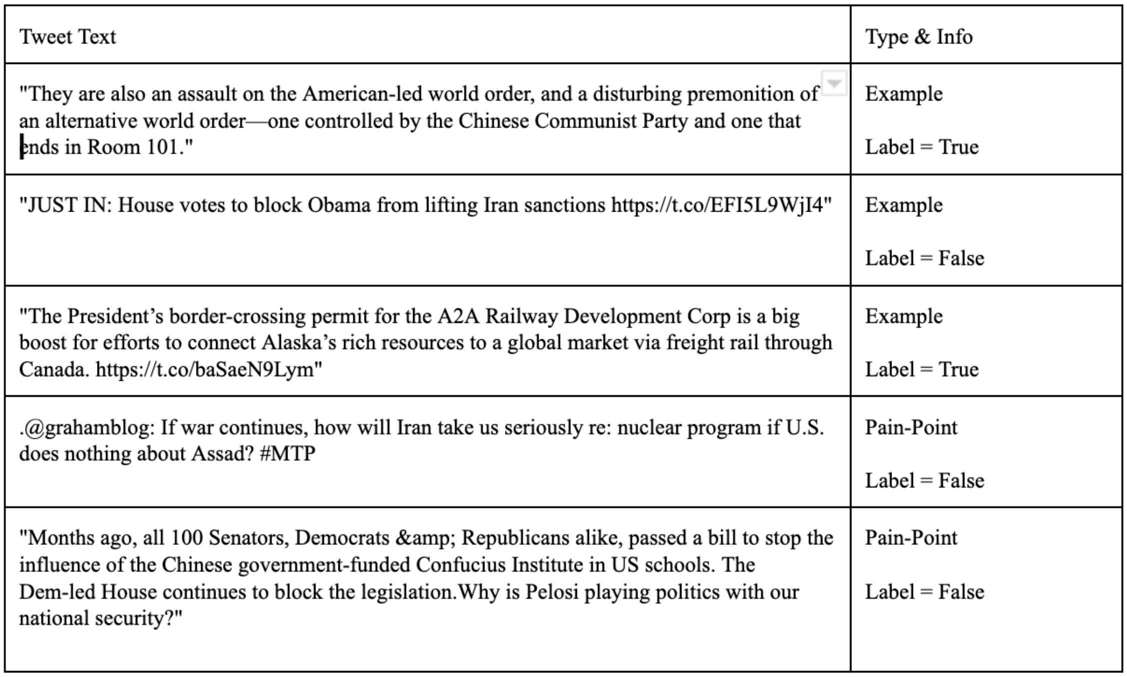
Experts in a given field possess a wealth of knowledge and understanding of the nuances and complexities of that domain. This expertise is essential for designing prompts that accurately convey the task requirements and constraints, as well as fine-tuning the model's parameters to generate relevant and meaningful outputs. Without this expertise, LLMs may generate outputs that are irrelevant, inaccurate, or even harmful in certain contexts.

Furthermore, subject matter expertise is also critical for evaluating the outputs generated by LLMs. Experts can provide feedback on the accuracy and relevance of the outputs and can identify errors or biases that may have been introduced by the model. This feedback is important for refining and improving the model and ensuring that it continues to produce high-quality outputs over time.

In addition to subject matter expertise, it is also important to consider the ethical implications of LLMs. As these models become more widely adopted and integrated into various applications, there is a risk that they may perpetuate biases and reinforce existing power structures if not carefully designed and monitored. Therefore, it is essential that experts work closely with developers and stakeholders to ensure that LLMs are developed and applied in an ethical and responsible manner.

5. Appendix I: Prompt Information

The example Tweets included:



Tweet Text	Type & Info
"They are also an assault on the American-led world order, and a disturbing premonition of an alternative world order—one controlled by the Chinese Communist Party and one that ends in Room 101."	Example Label = True
"JUST IN: House votes to block Obama from lifting Iran sanctions https://t.co/EFI5L9WjI4"	Example Label = False
"The President's border-crossing permit for the A2A Railway Development Corp is a big boost for efforts to connect Alaska's rich resources to a global market via freight rail through Canada. https://t.co/baSaeN9Lym"	Example Label = True
.@grahamblog: If war continues, how will Iran take us seriously re: nuclear program if U.S. does nothing about Assad? #MTP	Pain-Point Label = False
"Months ago, all 100 Senators, Democrats &amp; Republicans alike, passed a bill to stop the influence of the Chinese government-funded Confucius Institute in US schools. The Dem-led House continues to block the legislation.Why is Pelosi playing politics with our national security?"	Pain-Point Label = False

The basic contextual information was sourced from our dataset, and included information such as year of Tweet, political party associated, and term state of the Congressmember.

Bibliography

[1] Heather K. Evans, Victoria Cordova, and Savannah Sipole. Twitter style: An analysis of how house candidates used twitter in their 2012 campaigns: Ps: Political science amp; politics, Apr 2014.

[2] Lei Guang, Harris Doshay, Zeyu Li, Bailey Marsheck, Molly Roberts, and Young Yang. Part i: Who in the u.s. congress tweets about china?, May 2022.

[3] Aliza Sarlan, Chayanit Nadam, and Shuib Basri. Twitter sentiment analysis, Nov 2014.