Comparing Approaches to Create Time Series of LM

MD Touhidul Islam

February 2024

1 Introduction

People's opinions, emotions, and sentiments are not static, but dynamic and fluctuating over time. Knowing how these variables change is vital for both the government and business sectors. Businesses can use changing consumer trends to adjust their strategies and products. The government can use them to measure consumer confidence, approval ratings, and public opinions on policies. This can help them prevent potential losses of support by avoiding unpopular policies and designing policies that can address people's needs. However, surveys, which are the traditional methods of tracking these trends, are both time-consuming and expensive. In addition, in the context of rapid changes in public opinion, the information obtained from surveys may be outdated or irrelevant by the time it is available.

The proliferation of text-based social media platforms, such as Twitter, has provided a unique opportunity to tap into the collective thoughts and opinions of millions of individuals across a diverse range of topics. This vast repository of publicly available data offers the potential to infer population attitudes, akin to traditional public opinion polling methods. Having access to that information before your rival can provide a significant competitive edge. It can also enable the government to avoid unpopular policies that can permanently damage their reputation.

But the main question is can LM accurately predict the changes in people's emotions? If it can, what is the best approach to doing that? However, a key challenge with LMs is that they reflect only the dominant emotions present in the data they were trained on, which typically represents a specific point in time. To overcome this, the proposed approach involves training multiple LMs, each on data from different periods, so that each model captures the emotion of its specific point in time. By comparing the outputs of these time-specific models, we aim to track changes in emotions and opinions over time.

We used multiple approaches to create and train these time series language models, comparing their effectiveness in capturing the evolving sentiments of the population. Through this comparative analysis, we seek to identify not

only how we can use LM to track the change in opinion, but also the optimal approach to doing that.

2 Research Gap and Research Goal

2.1 Research Gap

In recent years, LMs have accumulated a vast amount of knowledge. Especially, a large amount of user-generated data from social networks has given us the opportunity to further train the model, which could lead to various development opportunities. Researchers have used these user-generated data to observe people's dominant opinions, sentiments, and emotions. However, previous opinion mining research using user-generated data has mainly focused on static sentiment analysis [9, 19, 6]. This means that it can be analyzed up to a point. However, people's opinion changes every day, and detecting that change can give us very useful information. So, the main question is can LM be used to accurately predict the changes in people's emotions? If it can, what is the best approach to doing that?

Researchers have mainly used dictionary-based approaches [9, 7] and/or sentiment analysis-based approaches [9, 14] to track the change in opinion over multiple points in time. However, sentiment analysis base methods, as highlighted in previous research [9, 7], require annotated training data, which can be costly and time-consuming to obtain. Furthermore, annotated data are reliant on subjective personal judgments, often constrained to predefined emotion sets, and exhibit limited adaptability to diverse contexts and scenarios. On the other hand, dictionary-based approaches face challenges related to the subjective nature of word meanings, the unawareness of the context, and limited coverage of dictionaries, leading to ambiguity in emotion detection. So, there is a need for an alternative that can overcome some of the limitations.

Using the vast amount of textual data and contextual understanding embedded in LMs, the research seeks to determine whether these models can accurately identify emotions and track their changing patterns over time. LM base method can solve most of the limitations that sentiment-based or dictionary-based methods face. Firstly, training with the mask language model does not require annotated data, which means that it does not face the same problem that sentiment-based methods face. Also, LM is aware of the context, which means it can infer the meaning of a word based on the context it is being used. It also does not have a problem with limited coverage as it uses the entire dataset to infer the meaning.

So, the main research question is

- Can we track changes in people's opinions using LMs trained on user-generated text, which can be used as an alternative to the survey?
- If we can, what is the best approach to do it?

2.2 Research Goal

Track changes in opinion over time using time series The main research goal is to use LMs to calculate a dominant emotion more accurately and calculate the changing patterns over time by training with data from different points in time.

Compare Approaches to Create Time Series using user-generated text The second goal of this thesis is to compare the different approaches to creating time series using LM that can predict dynamic variables over time.

Our goal is to create models that can track changes in public emotions over time by fine-tuning them on data from different periods. We will compare the models' predictions with self-reported emotions from surveys to evaluate their accuracy [4, 5]. By training models on data from specific time points, we aim to observe how opinions shift over time, helping us track changes in public sentiment.

3 Related Work

There have been a lot of studies analyzing emotion, opinion, and sentiment analysis in recent years. However, most of them are focused on sentiment analysis in a given time [9, 19, 6]. Few studies have dived into dynamic opinion detection.

3.1 Dynamics opinion detection:

Researchers have long tried to estimate, track, analyze, and detect people's emotional dynamics using vast amounts of data from the Internet. So far, the main methods used to calculate emotion dynamics are dictionary-based sentiment analysis approaches [9, 7] and sentiment analysis [9, 14].

Dictionary-based sentiment analysis approaches: The dictionary-based approach uses word co-occurrence to infer the semantic orientation of a word[15]. It Calculate the frequency of words around the target word to track the shift in sentiment[12, 20]. Each word is assigned a score from -5 to -1 for terms with a negative sentiment or from 1 to 5 for terms with a positive sentiment. Finally, the scores are summed to get the final score [10]. There has been research that has used similar research, which has used a dictionary-based technique to calculate dominant emotions like sadness, fear, and joy[9]. But it also faces its own set of challenges. Firstly, words' meanings can be subjective and depend upon context and situation, leading to ambiguity. Secondly, dictionary base methods suffer from limited coverage, while they only look at words within a specific window size (a window size of 2 would mean that the algorithm looks at two words before and two words after the target word) and may not include all relevant terms or expressions pertinent to the context being analyzed.

Additionally, It ignores words with non-standard spelling or spelling mistakes. Finally, it gives the same weight to every word. But in real life, some words are more important than others. LMs can understand context over an entire sentence, paragraph, or even document rather than being limited to a fixed window of words, which can also solve the problem with limited coverage[8]. It can also consider words with non-standard spelling or spelling mistakes using context.

Deep learning-based sentiment analysis techniques: In this method, Firstly the sentence containing the target word is selected. Then, deep learning-based sentiment analysis tools are used to calculate the sentiment of tweets. Each tweet can have three possible sentiments ("positive", "neutral", and "negative"). The model predicts the score of each tweet individually. Finally, the sentiment score of each tweet is summed up to calculate the overall sentiment of that point. This paper used a state-of-the-art supervised tool based on deep learning (German Sentiment) to analyze sentiment[17]. Sentiment analysis techniques [9, 14, 17] have several limitations, such as sometimes a question cannot be expressed in a single word. An example is the "opinion of the British population on tax hike in property". Here, only one word like property or tax can't capture the entire question. Another problem is it only calculates positive or negative sentiment around a word, but sometimes we need more than positive or negative like happy, sad, angry, apathetic, etc.

Therefore, alternative methods are needed that can overcome these challenges and provide more accurate and comprehensive insights into human emotions from user-generated text, such as social networks.

3.2 Time-aware language models:

The model trained on data from one point in time may not perform optimally at another point in time. Model training from the data of a specific period is able to capture emotion on that point. Recently, there has been an effort to build a language model that can perform optimally at a given time [13]. They employed state-of-the-art transformer-based models such as BERT and Roberta and fine-tuned them on weekly batches of tweets. The results showed that the time-aware models outperformed the base model on various natural language understanding tasks and were able to generate more relevant and diverse texts for different periods. The thesis aims to use the same methods by creating 200 time-specific language models and detecting how they predict a dominant emotion or opinion. There have also been approaches to using machine learning to monitor the evolution of factual knowledge on Wikipedia over time [11]. But it just showed how the perplexity has improved over the pre-train model. However, there is no study that attempts to track the emotional dynamic using social media data and see if it is comparable to real real-world data.

4 Data

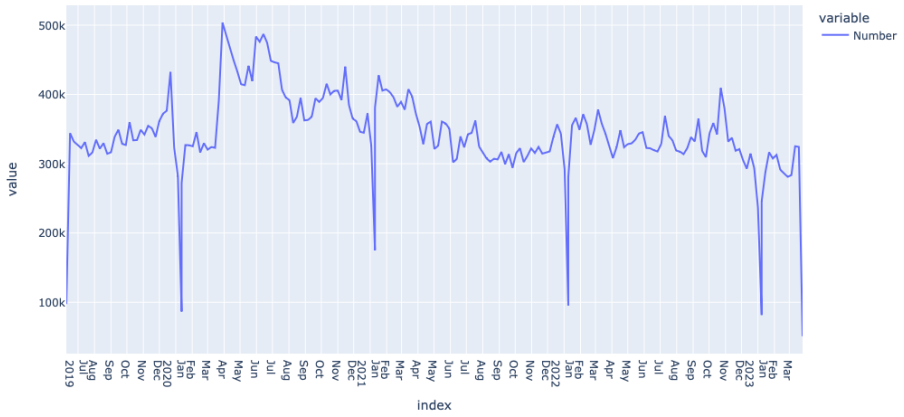


Figure 1: Number of tweets collected for training in a specific week plotted in a timeline

We use a data set containing tweets posted by a representative sample of UK Twitter users from the beginning of 2019 to March 2023. Organizations and users with too high or low activity levels were filtered out and the user sample is roughly gender balanced. The complete Twitter history of the users was extracted and split by week. The total number of tweets collected is around 70 million, with an average of 350 thousand weekly tweets. Survey data collected from YouGov [4] have been collected between the same period. The survey result was used to compare it with the model's predicted results, which monitor the emotions [1] and life satisfaction [2] of the British population every week. The total number of tweets used to train the model weekly is shown below.

5 Experimental setup

In the first step, a pre-trained LM like Bert will be taken as a base model. A pre-trained language model like BART has been trained on large amounts of text data before being fine-tuned for specific tasks [8]. Next, the model will be retrain with every 200 weekly data separately using that week's tweets as it has been depicted in the figure: 2. The mask language model will be used to train the models so that we do not need any annotated datasets. As we are using 70 million tweets, annotating individually is almost impossible. Since these models are trained on separate and distinct datasets, they are likely to reflect the dominant emotions of the respective periods [13]. In addition, they may be capable of capturing changes in emotions at various time points. That is the possibility that we want to check. The idea is to query the LMs with the

same questions that are asked in the case of human participants and to see how the result of the model compares with the real-world results. As surveys from different points of time are used to track the changes in people's opinions from one time to another, we want to track how the answer of the model changes over time as they have been re-trained on a dataset from that time. Third, do the changes in model prediction resemble the survey results?

The thesis aims to test whether it can accurately capture the intensity of different emotions and predict changes in emotional trends over time. For instance, we seek to determine whether the model can forecast increases or decreases in emotions such as confidence, happiness, or sadness over a specified period. This evaluation will provide insights into the model's ability to not only recognize and quantify emotions but also predict their changes over time.

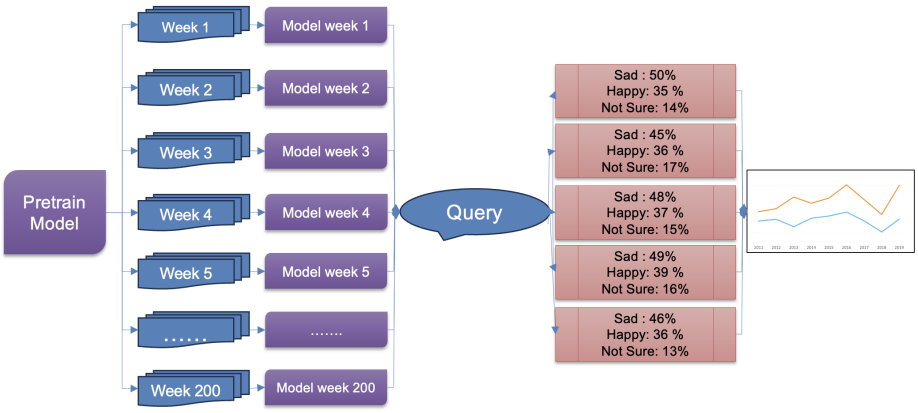


Figure 2: Process of Separate fine-tuning

The results of the model will be compared with the survey result at the time each model has been trained. For instance, if the question is "Broadly speaking, which of the following best describes your mood and/or how you have felt in the past week? Please select all those apply <mask>.", the probability of feeling sad might be 25% while feeling happy could be 20%. By comparing these model-generated results with survey data, we aim to track if the models can show similar trends or results. This approach allows for assessing the effectiveness of the models in capturing mood trends over time and their ability to align with real-world survey findings.

To construct a series of models from time-stamped data, we partition the data set into 200 subsets, each subset representing one week. This approach mirrors the methodologies used in previous research, such as an existing study [11], which used language LM to monitor the evolution of factual knowledge on Wikipedia over time. Similarly, our goal is to use LMs to track the evolution of public opinion or sentiment over time.

5.0.1 Different training Methods:

Each approach offers unique insight into the performance and effectiveness of the model in capturing temporal trends and patterns within the data.

Separate fine-tuning: Fine-tune the LM on each week's data separately, resulting in 200 distinct models as shown in the figure: 2. Each model is a version of the pre-trained model that was fine-tuned solely on the data of one selected week.

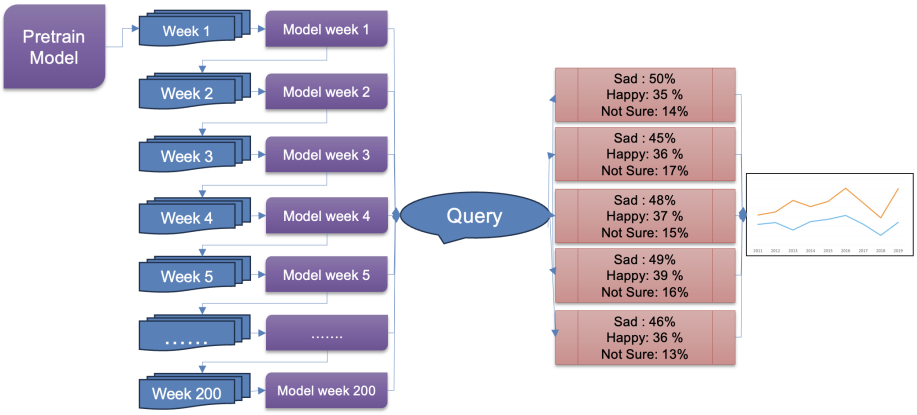


Figure 3: Process of cumulative fine-tuning

Cumulative fine-tuning: Fine-tune the LM on each week's data in a cumulative way, similar to Loureiro et al [13]. In this approach, instead of training the model from scratch, we will take the previously trained model and retrain it again as shown in the figure: 3. It will be interesting to see if the model gets better over time in predicting because of the additional data or if it will struggle to adjust to according to current opinion changes.

Training model with larger time window: The third approach involves training the model with larger monthly data rather than weekly data. Like, train the model with last month's data. This approach can reduce the amount of volatility and give more data to accurately predict the result. Instead of fine-tuning only one week, we could use a sliding window approach or mix the data of the week under observation with data from the preceding week(s). This could potentially make the trends smoother.

6 Evaluation

The survey data will be used as a gold standard for testing and will be compared with the prediction of the model. The survey data are collected from YouGov

[4]. We primarily took the survey, which monitors the emotions [1] and life satisfaction [2] of the British population every week. Firstly, the results generated by the trained model will be juxtaposed against the YouGov findings to gauge the model's performance. It can show us whether the model can predict the intensity or the changing pattern in the result of that survey. Then we will evaluate performance using three distinct methods: distance-based evaluation, correlation-based evaluation [18], and permutation testing [3].

6.1 Permutation Test

A permutation test (also called rerandomization test or shuffle test) is used to determine the statistical significance of a model by computing a test statistic on the dataset and then for many random permutations of those data[3]. The permutation test uses software to shuffle data before computing values (for example, mean and median differences) and to compare the results after shuffling to the original data. Repeat the process thousands of times to generate a proportion similar to the p-value you get in a t-test. It will determine whether the observed difference between the means of the gold standard and the model predictions is large enough to reject, at some significance level, the null hypothesis H that both the gold standard and the model prediction belong to the same distribution.

Hypothesis 0 (Test hypothesis) *The distributions of the two groups areidentical. (Gold standard = Model Prediction)*

Hypothesis 1 *The distributions of the two groups are not identical. (Gold standard ≠ Model Prediction*

6.2 Correlation-based evaluation

Correlation analysis will be conducted between the survey results and model predictions to assess any potential correlations between the variables [16, 21]. The correlation coefficient is a measure of how much two series vary together. A correlation of one indicates a perfect linear relationship with no deviations, while high correlations signify that models are good at predicting change.

$$y_t = b + a \sum_{j=0}^{K-1} x_{t-j} + \xi_t \tag{1}$$

If one result shows a higher correlation, Euclidean distance, or a positive Permutation Test result, could it just be by chance? To figure this out, we need to look beyond just that one result. For example, instead of only checking how the emotion "Happy" matches with the survey, we will also compare other emotions like "Sad," "Energetic," "Frustrated," and so on. In this way, we can see if the pattern holds true across different emotions, not just one. While it is plausible that one answer may exhibit a correlation by chance, it is improbable

that this would occur consistently across all answers. If the model can consistently reproduce similar correlations not only for the first answer but also for all answers, we can conclude that it is not happening by chance. Therefore, by analyzing all the answers from the model, we can better determine whether the observed correlations are statistically significant or merely random occurrences.

7 Initial experiments

In the initial experiment, we tested the Masked Language Model (MLM) using a separate fine-tuning technique. In a Masked Language Model, certain tokens in the input sequence are masked, and the model is tasked with predicting the most likely token to fill in those masked positions. The model is trained to predict the missing words in a sentence based on the context provided by the surrounding words. We train 200 models using a separate Fine-Tuning technique, where each model is trained from scratch every time. This means that there is no influence from the data of the previous week on the training of the current week's model.

TwHIN-BERT models have been selected to be fine-tuned on weekly data [22]. TwHIN-BERT has been selected over others because it is a language model trained on 7 billion Tweets from more than 100 distinct languages. It differs from prior pre-trained language models as it is trained with not only text-based self-supervision (e.g., MLM) but also with a social objective based on the rich social engagements within a Twitter Heterogeneous Information Network (TwHIN). This approach involves the creation of a TwHIN, which unifies various user engagement logs of various types. Subsequently, the TwHIN data are subjected to scalable embedding and approximate nearest-neighbor search techniques, allowing the exploration of hundreds of billions of engagement records to identify socially similar pairs of tweets [22]. This comprehensive training process allows TwHIN-BERT to capture both linguistic and social nuances, thereby enhancing its ability to understand and generate meaningful content in the context of Twitter interactions.

7.1 Training of models

Each model has been trained following recommendations provided by Hugging Face. Firstly, the existing model is loaded and fine-tuned on a specific downstream task or data set. We started training using a fresh TwHIN-BERT model every week and trained it exclusively on the data from that specific week, for each of the 200 models. This approach ensures that there is no influence from the previous week's data on the training of the current week's model. All 200 models have been trained using the same hyperparameter to maintain consistency.

The perplexity of the model after training has been compared with the perplexity of the model before training 4. The results show an overall improvement of 10-12 points compared to the model before training. This improvement suggests that the training process has enhanced the model's ability to accurately

Train	80%
Test	20%
Learning Rate	5e-5
Epoch	10
Chunks Size	128
masked	15%

Table 1: Retraining parameters

predict, leading to better overall performance.

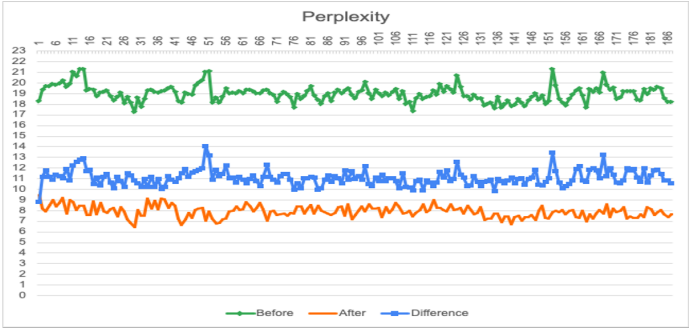


Figure 4: Pre-train and Post-train perplexity of 200 weekly models

7.2 Performing queries

The question that was initially asked to test the model was "Broadly speaking, which of the following best describes your mood and/or how you have felt in the past week? Please select all that apply <mask>.". The questions have been selected from the YouGov site[1] so that the model result can be compared to the actual survey result. The list of choices, which was given as an option at YouGov[1] survey has passed as targets. The targets for the question were ("Happy", "Sad", "Energetic", "Apathetic", "Inspired", "Frustrated", "Optimistic", "Stressed", "Content", "Bored", "Lonely", "Scared", "Other").

When we passed the token, we encountered the problem that the specified target token "Stressed" does not exist in the model vocabulary. Consequently, it was replaced with "Stress". This issue arises because the tokenizer converts tokens containing "Stressed" into two separate tokens: "Stress" and "ed". To address this, initially, we attempted to solve the problem by using multi-token replacements, such as placing two <mask>tokens side by side. After receiving the results, we took same-ranking tokens for both masks and concatenated them. However, this approach produced concatenated words like "glori bad," lacking meaningful interpretations. The result was 0- very good, 1- really ing, 2- glori

bad. Therefore, we opted for a single-token solution, acknowledging that models are trained to predict specific tokens and that "Stress" adequately represents both "Stress" and "Stressed" in the context of the vocabulary.

7.3 Preliminary Results

7.4 Initial Results

Firstly, p-value results from the permutation testing are below the critical value 0.05>9.999e-06. This means that the null hypothesis can be rejected by a huge margin. The p-value for all sentiments is almost zero. It indicate our model have failed to replicate the survay result. In Table 2, we computed the average values for the Euclidean distance, correlation, and p-value in all results. The average Euclidean distance is approximately 3.484, which reaffirms a substantial deviation from the survey data. The correlation coefficients are also generally low, with values below 0.1, except for the emotion "Sad". This indicates model are struggling to get similar results compared to the gold standard. Finally, examining the graphs 5b and 6a, a similar disparity between the survey results can be found. The predictions of the model become evident. The model prediction for the target option approaches zero, indicating a substantial deviation from the survey results.

7.4.1 Min Max rescaling

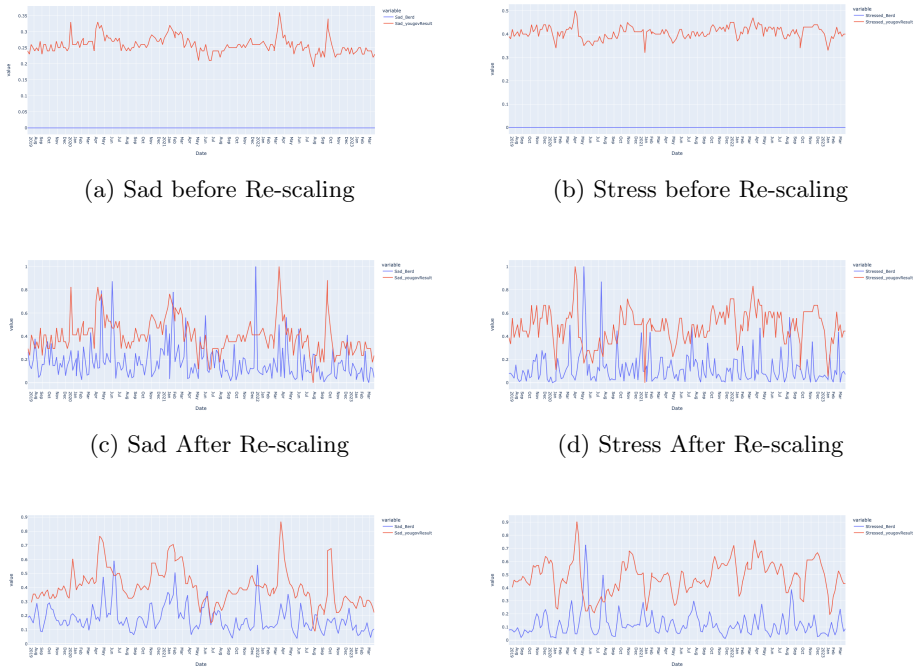
If you sum the result of all options in the survey in a certain week, it will be 1. But, if you do the same with model prediction, it will be less than one. This happens because both of them use different scaling methods. In the case of model prediction, the sum of all tokens is 1. As we are interested in only those tokens in the survey, the sum will always be less than 1. That could be a probable reason for the huge difference between the survey result and the prediction of the model. Normalization offers a solution by adjusting the input features to a consistent range, often [0, 1], ensuring uniformity and comparability of results. Among normalization techniques, min-max scaling stands out as the simplest, involving the re-scaling of feature ranges to fit within [0, 1]. We normalized the value of an individual target (e.g., an emotion such as sadness) using the lowest and highest values of all other targets in a range [0, 1].

$$rescaled = \frac{e_i - E_{min}}{E_{Max} - E_{min}} \tag{2}$$

where,

e_i = prediction from the target at week i
E_{min} = minimum prediction from all target predictions at week i
E_{max} = maximum prediction from all target predictions at week i

Following the 0-1 rescaling process, the results become considerably more comparable. In both Figures: 6c and Figure: 5d, it is apparent that the model



(a) Sad before Re-scaling

(b) Stress before Re-scaling

(c) Sad After Re-scaling

(d) Stress After Re-scaling

(e) Sad 3 weeks moving average

(f) Stress 3 weeks moving average

Figure 5: The result of two emotions sad and content before and after re-scaling compared against YouGov[2] survey results

tends to underestimate values compared to the YouGov[2] survey results. However, there is a noticeable alignment in the shifts in people's opinions between YouGov[2] and the model, suggesting some level of synchronicity in the observed changes.

Specifically, when examining the emotion "Sad" 6c the abrupt peaks in this emotion are also reflected in the model predictions, although not as prominently as in the survey data. Conversely, in the case of "Stress," apart from a single significant peak, the alignment with the survey data is less distinct, with the model's predictions displaying less clarity.

However, referring to Table: 2, the correlation values remain below 0.1, except for 'Sad', and the Euclidean distance is still quite high.

Examining the correlation in Tables: 2 and 2, we noted a slight increase, but it is not strong enough to indicate a clear connection between the variables. The Euclidean distance, indicating the difference between survey results and model predictions, remains considerable for both "Sad" and "Stress," signifying significant discrepancies.

Despite these limitations, the model demonstrates an ability to capture some trends and changes in public sentiment. Therefore, it is worthwhile to explore

		Sad	Content	others	Stress	Lonely	Inspired	Frustrated	Optimistic	Bored	Average
Euclidean Distance	Initial	3.615	3.631	0.964	5.671	2.462	1.305	5.026	2.838	3.325	3.484
	Re-scaled	4.077	6.640	6.228	5.806	5.644	6.230	6.984	6.086	4.436	5.792
	Rolling Average	3.527	6.377	5.747	5.450	5.174	5.911	6.688	5.684	3.844	5.378
	Multi-query	3.613	5.994	4.401	5.007	4.570	5.662	7.125	5.612	3.958	5.105
	Multi-query Mean	4.165	6.098	5.166	5.254						
Correlation	Initial	0.243	-0.018	-0.023	-0.089	0.123	0.045	0.163	-0.070	-0.038	
	Re-scaled	0.243	-0.018	-0.023	-0.089	0.123	0.045	0.163	-0.070	-0.038	
	Rolling Average	0.422	-0.113	-0.113	-0.155	0.250	0.078	0.379	-0.104	-0.033	
	Multi-query	-0.162	-0.043	0.046	-0.097	0.122	-0.047	0.179	-0.202	-0.030	
	Multi-query Mean	-0.046	-0.052	-0.012	-0.027	5.254					
P-Value	Initial	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	
	Re-scaled	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	
	Rolling Average	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	
	Multi-query	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	
	Multi-query Mean	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	9.999e-06	

Table 2: Comparison between Initial, Resealed, Rolling average and, Multi-query average

the utility of this method with newer models to track emotional fluctuations over time.

7.4.2 Rolling average

Despite training the model on a large number of tweets, considerable noise persists in the results. To mitigate this issue, we have applied a rolling average technique as a potential remedy. By recalculating the graph using a three-week rolling average, we've observed a noteworthy reduction in noise, leading to more comparable and stable results.

$$SMA = \frac{A_1 + A_2 + A_3 ..... A_n}{n} \tag{3}$$

where,

SAM = Simple Moving Average
A_n = The prediction on specific week n
n = Number of weeks

Figures 6e and 5f illustrate the outcomes after implementing the 3-week rolling average. The trend in changes in people's opinions becomes much clearer post-rolling average computation. However, it is important to note that the results vary significantly between different emotions. Although the emotion "sad" in Figure:6e exhibits greater similarity to the correlation survey results, the results for the emotion "stress" in Figure:5f are considerably divergent. This discrepancy underscores the lack of consistency in the results, making replication between different emotions more challenging.

Post-application of rolling averages at Table: 2, the average Euclidean distance decreases to 5.278 Furthermore, the correlations, predominantly below 0.5, fail to reach a significant threshold to indicate a strong relationship between the variables.

Upon analysis of changes for both emotions (see Tables: 2 and 2), inconsistencies were observed. However, in both cases, there was a significant decrease in the Euclidean distance, from 4.077 to 1.6433 for "Sad" and from 5.644 to

4.152 for "Stressed." However, the distances remain substantial, highlighting significant differences between the survey results and the model predictions.

7.4.3 Average of multiple queries

The results continue to show a strong dependence on the query used. Even minor alterations in query wording can lead to notable changes in the results, raising concerns about the reproducibility of findings across different queries and scenarios. Addressing this issue is crucial to enhance the usability of the methods employed. We still need to tackle the problem with differences in results due to the difference in query. One potential solution involves running the same query multiple times with slight variations in wording and averaging the results. This approach aims to mitigate the impact of specific phrasing on the outcome, thus yielding a more consistent and reproducible result across different types of questions. In our testing, we give the model three queries. Those queries are "Broadly speaking, <mask>is the emotion that best describes your mood and/or how you have felt in the past week., "Broadly speaking, the emotion <mask>best describes your mood and/or how you have felt in the past week." and "Which of the following best describes your mood and/or how you have felt in the past week Please select all that apply <mask>". The result is then averaged to calculate the final result.

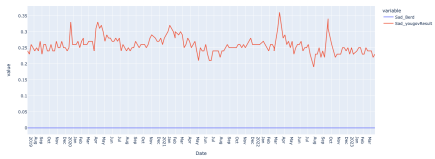
The expectation is that by averaging across multiple query formulations, any variance introduced by wording differences will be balanced out, leading to greater reliability and consistency in the final result. From the figure, it is evident that the results before and after implementing the multi-query average are largely similar. There are no significant changes observed in the outcomes. While the volatility has been slightly altered leading to a smoother graph following the multi-query averaging, the model still struggles to accurately replicate the survey results or precisely predict changes.

The average Euclidean distance is 5.105, respectively, but they remain relatively high. Additionally, the correlation remains very low. Upon examination of Tables 2 and 2, it is evident that none of the metrics, Euclidean distance, correlation shows a significant improvement.

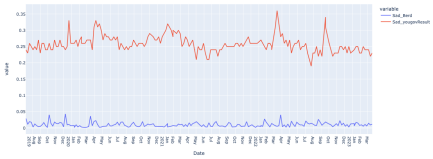
These findings suggest that, despite employing the approach of averaging across multiple queries, the model's performance remains insufficient in capturing the nuances of the data or accurately predicting fluctuations. However, it is still worth exploring whether the approach might yield better results with alternative training or querying methods.

8 Timeline

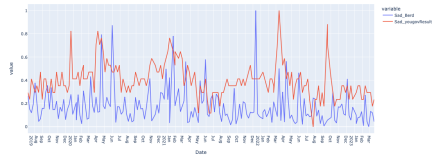
Exploring alternative training methods or quarrying techniques could indeed provide valuable insights into improving the performance of the model. Using cumulative training that incorporates older information over time could mitigate issues stemming from brief periods of data, potentially improving performance,



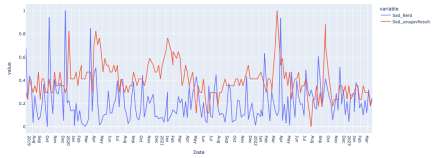
(a) Sad before Re-scaling



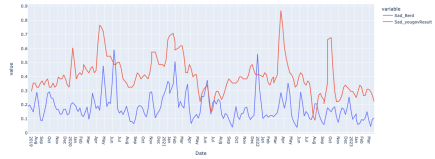
(b) Sad multi-query



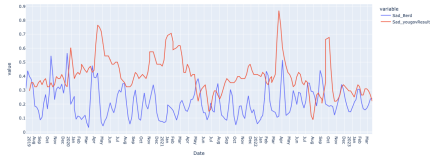
(c) Sad After Re-scaling



(d) Sad multi-query Re-scaled



(e) Sad 3 weeks moving average



(f) Sad multi-query moving average

Figure 6: The result of two emotions sad and content before and after multi-query compared against YouGov[2] survey results

and increasing correlation.

Things we want to test in the remainder of the thesis are.

1. Training methods:
   - Seperate Fine Tuning
   - Cumulative Fine Tuning
   - Training model with larger time windows
2. Multi-Query selection methods: Mean, median, mode, best query overall, best result from all a specific month.
3. Re-scaling methods.
   - Min-Max Scaling technique re-scales the range of features to scale the range in [0, 1] or [-1,1].
   - Standard scaling transforms the values of each feature to have a mean of 0 and a standard deviation of 1 by subtracting the mean and dividing by the standard deviation.

- Robust scaling re-scales the values of each feature to the range [0, 1] by subtracting the median and dividing by the interquartile range (IQR). This reduces the influence of outliers and extreme values, but can also reduce variance and information from the data.

The expected timeline for the thesis is shown below.

Week	Training Model	Rescaling	Analyzing result	Final Report writing
1 - 4	Cumulative Model Training(CMT)			Introduction and Motivation
4 - 8		Min – Max Rescaling	Analyze result from Cumulative training	Related work and Research Gap
8 - 12	Training Models with different time window(TMDTW)	Standard Rescaling	Analyze result from TMDTW	Details explanation of experimental set up
12	**Mid Tarm Presentation**			
12 - 16		Robust Rescaling	Comparison between different approach	Companion between multiple approaches
16 - 20				Evaluation of test Result
20-22	**Final proofreading & Incorporating feedback**			
22	**Final Presentation**			

Figure 7: Expected Timeline

References

[1] Britain's mood, measured weekly. URL: https://yougov.co.uk/topics/politics/trackers/britains-mood-measured-weekly.

[2] Life satisfaction, measured weekly. URL: https://yougov.co.uk/topics/politics/trackers/life-satisfaction-measured-weekly.

[3] Permutation test - an overview | ScienceDirect topics. URL: https://www.sciencedirect.com/topics/mathematics/permutation-test#.

[4] YouGov | what the world thinks. URL: https://yougov.co.uk/.

[5] YouGov | what the world thinks. URL: https://yougov.co.uk/.

[6] Nirmal Varghese Babu and E. Grace Mary Kanaga. Sentiment analysis in social media data for depression detection using artificial intelligence: A review. 3(1):74. doi:10.1007/s42979-021-00958-1.

[7] Asaf Beasley, Winter Mason, and Eliot Smith. Inferring emotions and self-relevant domains in social media: Challenges and future directions. 2(3):238–247. Publisher: Educational Publishing Foundation. URL: http://www.redi-bw.de/db/ebsco.php/search.ebscohost.com/login.aspx%3fdirect%3dtrue%26db%3dpdh%26AN%3d2016-47442-004%26site%3dehost-live, doi:10.1037/tps0000086.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. URL: http://arxiv.org/abs/1810.04805, arXiv:1810.04805[cs].

[9] David Garcia, Max Pellert, Jana Lasser, and Hannah Metzler. Social media emotion macroscopes reflect emotional experiences in society at large. URL: http://arxiv.org/abs/2107.13236, arXiv:2107.13236[cs].

[10] Anastasia Giachanou and Fabio Crestani. Tracking sentiment by time series analysis. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, SIGIR '16, pages 1037–1040. Association for Computing Machinery. URL: https://dl.acm.org/doi/10.1145/2911451.2914702, doi:10.1145/2911451.2914702.

[11] Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. URL: http://arxiv.org/abs/2204.14211, arXiv:2204.14211[cs].

[12] Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey. URL: http://arxiv.org/abs/1806.03537, arXiv:1806.03537[cs], doi:10.48550/arXiv.1806.03537.

[13] Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. TimeLMs: Diachronic language models from twitter. URL: http://arxiv.org/abs/2202.03829, arXiv:2202.03829[cs].

[14] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. SemEval-2018 task 1: Affect in tweets. In Marianna Apidianaki, Saif M. Mohammad, Jonathan May, Ekaterina Shutova, Steven Bethard, and Marine Carpuat, editors, *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17. Association for Computational Linguistics. URL: https://aclanthology.org/S18-1001, doi:10.18653/v1/S18-1001.

[15] Le T. Nguyen, Pang Wu, William Chan, Wei Peng, and Ying Zhang. Predicting collective sentiment dynamics from time-series social media. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '12, pages 1–8. Association for Computing Machinery. URL: https://dl.acm.org/doi/10.1145/2346676.2346682, doi:10.1145/2346676.2346682.

[16] Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith. From tweets to polls: Linking text sentiment to public opinion time series. 4(1):122–129. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14031, doi:10.1609/icwsm.v4i1.14031.

[17] Max Pellert, Hannah Metzler, Michael Matzenberger, and David Garcia. Validating daily social media macroscopes of emotions. 12(1):11236. Number: 1 Publisher: Nature Publishing Group. URL: https://www.nature.com/articles/s41598-022-14579-y, doi:10.1038/s41598-022-14579-y.

[18] Joan Serrà and Josep Ll. Arcos. An empirical evaluation of similarity measures for time series classification. 67:305–314. URL: https://www.sciencedirect.com/science/article/pii/S0950705114001658, doi:10.1016/j.knosys.2014.04.035.

[19] Pawel Sobkowicz, Michael Kaschesky, and Guillaume Bouchard. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. 29(4):470–479. URL: https://www.sciencedirect.com/science/article/pii/S0740624X12000901, doi:10.1016/j.giq.2012.06.005.

[20] Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. 37:141–188. URL: http://arxiv.org/abs/1003.1141, arXiv:1003.1141[cs], doi:10.1613/jair.2934.

[21] Xiaoyue Wang, Abdullah Mueen, Hui Ding, Goce Trajcevski, Peter Scheuermann, and Eamonn Keogh. Experimental comparison of representation methods and distance measures for time series data. 26(2):275–309. doi:10.1007/s10618-012-0250-5.

[22] Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. TwHIN-BERT: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. URL: http://arxiv.org/abs/2209.07562, arXiv:2209.07562[cs], doi:10.48550/arXiv.2209.07562.