MASTER THESIS

Comparing Approaches to Create Time Series of LM

submitted by

MD TOUHIDUL ISLAM

Submitted to the

Chair for Data Science in the Economic and Social Sciences

within the

Faculty of Business Administration

at University of Mannheim

March 3, 2025

Marlene Lutz, M.Sc.

Advisor

Prof. Dr. Markus Strohmaier

Supervisor

Contents

Abstract *This thesis investigates the effectiveness of Masked Time-Aware Language Models (MTALMs) in detecting and tracking changes in public sentiment and emotions over time. Traditional sentiment analysis methods, including dictionary-based approaches and static deep learning models, often fail to capture the nuanced and evolving nature of public opinion. Similarly, survey-based sentiment tracking methods suffer from temporal lag, high costs, and limited sample sizes. Existing computational models typically treat sentiment as a static phenomenon, ignoring the temporal dynamics that influence emotional expression.*

To address these limitations, this research develops and evaluates a novel transformer-based approach that tries to track temporal changes. By training MTALMs on separate time-specific data, we aim to provide insights into how opinions and emotions shift in response to events and external influences. Through a comparative analysis of time-aware and static models, this study explores the potential of MTALMs for temporal sentiment tracking. The findings contribute to the growing field of temporal sentiment analysis, offering valuable tools for researchers, businesses, and policymakers seeking to understand and respond to the dynamic landscape of public opinion.

1

Introduction

Public opinion fundamentally shapes our decision-making processes across virtually all aspects of society. From selecting which products to purchase and political parties to support, to determining holiday destinations and investment opportunities, understanding collective sentiment provides crucial guidance for individuals and organizations alike [34]. However, our preferences regarding which smartphone to buy, which candidate to vote for, or which vacation spot to visit evolve continuously over time in response to new information, experiences, and external events [20].The true value of public opinion lies not merely in its static measurement but in its dynamic nature.

This temporal dimension reveals critical insights that would remain hidden in isolated snapshots, such as emerging trends, response patterns to specific triggers, and the effectiveness of various interventions. For instance, tracking the shift from initial enthusiasm about a product to growing disappointment can help companies identify quality issues before they become widespread problems [14].

In our daily lives, we regularly encounter metrics that quantify these opinion trajectories. The Consumer Confidence Index guides economic forecasts, political polling influences campaign strategies, approval ratings shaping policy decisions, Business Confidence Index informs investment choices, and product ratings direct consumer behavior. These temporal indicators serve as essential tools across diverse sectors, enabling more informed and adaptive decision-making [41]. The contemporary digital landscape, with its unprecedented wealth of user-generated content, now offers opportunities to track these opinion dynamics with greater precision, frequency, and scale than ever before, transforming how we understand and respond to changing public sentiment [35].

1.1 The Importance of Tracking Public Opinion

Public opinion serves as a vital signal that guides decision-making across diverse sectors of society. The ability to accurately track and analyze sentiment fluctuations provides stakeholders with

actionable intelligence that can shape strategies, policies, and research directions in increasingly dynamic environments.

Business Applications

For businesses, time series analysis has evolved from a peripheral monitoring tool to a central component of strategic decision-making. Organizations systematically track consumer sentiment to optimize product development, marketing strategies, and brand positioning in response to rapidly changing market preferences [26]. This proactive approach enables companies to identify emerging consumer needs and address potential issues before they significantly impact market performance. For instance, leading technology firms continuously analyze social media discourse surrounding their products, allowing them to prioritize features that resonate with users and quickly remediate concerns that might otherwise damage brand reputation [34].

Beyond product-specific applications, time series analysis has transformed financial markets through the development of algorithmic trading systems that incorporate social media signals. Research by Bollen, Mao, and Zeng [10] demonstrated that Twitter sentiment patterns could predict directional changes in the Dow Jones Industrial Average with remarkable accuracy, illustrating how temporal data can provide traders with meaningful competitive advantages.

Government and Policy Applications

In the public sector, time series analysis has become an essential component of responsive governance and effective policy implementation. Government agencies utilize sentiment tracking to gauge public reception of policy initiatives, monitor satisfaction with public services, and adapt communication strategies during periods of social or economic instability [41].

The application of time series analysis extends beyond immediate crisis response to long-term governance improvements. Public service agencies increasingly incorporate sentiment analysis into their quality management systems, using temporal data to identify deteriorating service areas and prioritize improvement initiatives [8]. In the security domain, intelligence agencies have developed sophisticated systems to analyze sentiment patterns in online discourse, helping identify radicalization processes and potential threats before they manifest in harmful actions [3].

Social Science and Media Applications

The social sciences have embraced temporal analysis as a methodological breakthrough that complements traditional research approaches. Sociologists and political scientists leverage temporal tracking to examine how public attitudes toward critical issues evolve in response to events, information campaigns, and social interactions [16]. This dynamic approach reveals causal mechanisms and attitude formation processes that remain invisible in static polling data, providing deeper insights into social phenomena. Electoral researchers, in particular, have demonstrated how senti-

ment trajectories can predict voting behavior more accurately than point-in-time polls by capturing momentum shifts and response patterns to campaign events [51].

1.1.1 Established Metrics

Several established metrics exemplify the institutionalization of sentiment tracking across sectors. [13]. These standardized metrics have demonstrated their value through decades of application, validating the fundamental importance of systematic sentiment monitoring.

- The Consumer Confidence Index (CCI) [13] provides valuable insights into consumer perceptions of current and future economic conditions. For instance, retailers might delay expansion plans when observing declining consumer confidence trends. Copy
- Political polling and approval ratings offer crucial feedback on public reception of candidates and policies. A presidential approval rating dropping below 40% might signal to lawmakers the need to reconsider certain policy directions.
- The Business Confidence Index serves as another vital tool, helping policymakers evaluate the impact of their decisions on the business community. When this index shows sustained decline, central banks might consider monetary policy adjustments to stimulate economic activity.

1.2 Limitations of Traditional Survey Methods

While traditional survey methodologies have been the primary means of gathering opinion data, they present significant limitations that affect their utility in today's fast-paced information environment.

According to industry estimates, the global annual spending on market research surveys is between $50-70 billion. This includes $10-20 billion spent on public sector surveys and $5-10 billion on internal corporate surveys, with total annual spending on surveys approximately $65-100 billion [43, 25]. These figures highlight the significant investment that organizations make to gather insights, yet the return on this investment is often compromised by inherent methodological limitations:

- **Time and Resource Intensity:** Conducting large-scale surveys is both time-consuming and expensive. A comprehensive national opinion poll might take weeks to design, implement, and analyze, by which time the results may no longer reflect current opinions [4].
- **Temporal Lag:** By the time traditional surveys are completed, the data often becomes outdated and fails to capture rapidly changing public sentiment. For example, during the

early stages of the COVID-19 pandemic, consumer attitudes about shopping habits changed weekly, rendering monthly surveys largely obsolete [6].

– **Sample Size Limitations:** Surveys typically rely on relatively small samples that may not accurately represent broader population sentiment. A political poll of 1,000 respondents might miss significant opinion shifts in specific demographic groups [27].

– **Various Forms of Bias:** Survey responses can be influenced by self-selection bias (where certain types of people are more likely to participate) and social desirability bias (where respondents provide answers they believe are more socially acceptable). For instance, pre-election polls have sometimes failed to accurately predict outcomes when respondents are reluctant to share their true voting intentions [44].

– **Static Nature:** Traditional surveys provide only snapshots rather than continuous measurements of public opinion. A quarterly customer satisfaction survey cannot capture the immediate impact of a product recall or viral negative publicity [28].

These limitations can result in outdated or incomplete information, potentially leading to misguided decisions. For example, a government agency relying on annual surveys to guide public health messaging might miss rapid shifts in public attitudes toward emerging health concerns.

1.3 The Promise of Social Media Data and Language Models

The convergence of social media platforms and advanced language models has opened unprecedented opportunities for sentiment analysis. Together, they offer potential to transform how we track and understand public opinion across various domains. The following paragraphs explore the key opportunities presented by each.

1.3.1 Opportunities from Social Media Data

The timeliness of social media data represents its most compelling advantage for sentiment analysis. Unlike traditional surveys that require weeks or months to design, distribute, collect, and analyze, social media content can be captured and processed in near real-time [21]. This immediacy enables organizations to detect emerging sentiment shifts as they happen, providing a decisive advantage in crisis management, product launches, or political campaigns where rapid response can determine success or failure.

The unprecedented scale of social media data constitutes another transformative opportunity. With platforms like Twitter processing approximately 500 million tweets daily, researchers gain access to sentiment expressions from millions of individuals across diverse demographics and

geographies [32]. This volume allows for more statistically robust analyses that can detect subtle trends and patterns invisible in smaller samples, while also enabling more granular segmentation by geographic, demographic, or behavioral factors.

The cost-effectiveness of social media analysis offers a third significant advantage over traditional opinion research methods. Conducting large-scale surveys or focus groups involves substantial expenses for participant recruitment, incentives, and administration. In contrast, social media data is largely publicly available, requiring primarily computational resources for collection and analysis [40]. This economic efficiency enables more continuous monitoring and broader coverage of topics than would be financially feasible with conventional methods.

1.3.2 Opportunities from Large Language Models

The contextual understanding capabilities of Large Language Models represent a fundamental advance in time series analysis. Unlike dictionary-based methods that consider words in isolation, LLMs process text holistically, recognizing how meaning shifts based on surrounding words, sentence structure, and broader discourse [18]. This contextual awareness enables LLMs to accurately interpret complex linguistic phenomena like sarcasm, irony, and implicit sentiment that traditional methods consistently misclassify, resulting in significantly more accurate time series analysis.

The adaptability of LLMs to linguistic evolution provides another crucial advantage for contemporary time series analysis. Social media language evolves rapidly, with new slang, abbreviations, and semantic shifts emerging continuously. While dictionary-based approaches require manual updates to capture these changes, LLMs can more readily adapt to evolving language patterns through their statistical learning capabilities and transfer learning potential [33]. This adaptability ensures more consistent performance across the dynamic linguistic landscape of social media.

The multi-dimensional emotional analysis capabilities of advanced language models offer perhaps their most transformative potential. Traditional time series analysis typically reduces complex emotional expressions to simple positive/negative polarity. In contrast, LLMs can detect and quantify multiple emotional dimensions simultaneously—distinguishing between similar emotions like disappointment and anger, or recognizing mixed emotional states [2]. This nuanced understanding enables more sophisticated tracking of emotional responses to events, products, or policies, revealing insights that binary sentiment classification would miss entirely.

1.3.3 The Integration Challenge and Opportunity

Despite these advantages, conventional language models face a fundamental limitation: they typically reflect the dominant emotions and linguistic patterns present in their training data, which is often temporally constrained. This limitation highlights the need for specialized approaches like Masked Time-Aware Language Models (MTALMs) [35]. By integrating the temporal dimension into language models and applying them to the rich stream of social media data, researchers can develop unprecedented capabilities for tracking the dynamic nature of public opinion—transforming

how organizations understand and respond to evolving sentiment across various domains of societal importance.

2

Research Gap and Research Goal

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 [20, 46, 48]. 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?

2.1 Limitations of Existing Approaches

Tracking changes in public opinion over time has traditionally relied on two main methods: **dictionary-based approaches** and **sentiment analysis-based approaches**. While these methods have been widely used, they come with significant limitations that restrict their effectiveness in capturing evolving language trends and complex emotional expressions.

2.1.1 Dictionary-Based Approaches

Dictionary-based methods classify emotions or sentiments by using predefined word lists or lexicons [**beasley2016inferring**, 20]. These lexicons assign fixed sentiment values to words, which are then used to analyze large text corpora. However, these approaches have several limitations:

– **Contextual Ambiguity:** The meaning of a word can change depending on its context, which dictionary-based methods fail to capture. For example, the word *sick* can indicate illness in a medical context ("I am feeling sick today.") but can also be used as slang for something impressive ("That trick was sick!"). Similarly, the word *cold* could refer to temperature,

emotion (e.g., "a cold-hearted person"), or even an illness, making dictionary-based methods prone to misclassification.

– **Limited Vocabulary Coverage:** Dictionaries and lexicons are static and do not evolve with the rapid changes in language. New words, slang, and cultural shifts in word usage often go undetected. For instance, internet slang like *lit*, *vibe*, or *ghosting* carries emotional connotations that traditional sentiment lexicons may not recognize. As a result, dictionary-based approaches struggle to accurately capture the full spectrum of sentiment in modern digital communication, particularly in social media contexts.

– **Inability to Capture Nuanced Emotions:** Emotion is not binary (positive vs. negative) but exists on a spectrum with multiple dimensions such as joy, sadness, frustration, and anticipation. Dictionary-based methods often assign fixed sentiment scores to words without considering how emotions interact. For example, the word *bittersweet* conveys both happiness and sadness simultaneously, yet a dictionary approach may classify it as either positive or negative, missing the dual emotional nature of the term.

2.1.2 Sentiment Analysis-Based Approaches

Sentiment analysis-based methods typically use machine learning models trained on annotated datasets to classify text as positive, negative, or neutral [20, 38]. While these models improve upon dictionary-based approaches, they also have inherent limitations:

– **Dependence on Labeled Training Data:** Most machine learning models require large amounts of manually labeled data to learn sentiment classifications. However, obtaining high-quality annotated datasets is costly, time-consuming, and subject to human biases. For example, two annotators may interpret the same sentence differently based on their perspectives—one may label "I'm fine" as neutral, while another may detect sarcasm and label it as negative.

– **Simplistic Categorization of Emotions:** Many sentiment analysis models reduce emotions to basic categories such as positive, negative, or neutral, failing to capture complex and subtle variations in sentiment. For instance, emotions like *frustration* and *anger* may both be classified as negative, even though frustration often carries a sense of helplessness while anger implies a stronger reaction. Similarly, *contentment* and *enthusiasm* may both be classified as positive, but they differ in intensity and emotional impact.

– **Limited Adaptability to Different Target Scenarios:** RoBERTa models trained on specific emotion categories (such as happy, sad, loneliness, apathetic, angry) are fundamentally constrained to those predefined emotions. This rigidity presents a major obstacle when need to analyze different emotional dimensions. For example, a model trained to detect the

five basic emotions cannot be easily repurposed to analyze more nuanced states like "anticipation" or "contentment" without extensive retraining. This lack of flexibility becomes particularly problematic in longitudinal studies where research questions may evolve over time, or in cross-cultural contexts where emotional expressions vary significantly.

2.2 The Need for a Better Approach

Given the shortcomings of dictionary-based and sentiment analysis-based methods, there is a strong need for an alternative approach that can address these issues effectively. An ideal approach should:

- **Capture the Contextual Meaning of Words:** Rather than relying on static dictionaries, the method should understand how words derive meaning from the surrounding text. For example, in the sentence "That exam was a nightmare," the word *nightmare* should be recognized as a metaphor for a stressful experience rather than being taken literally.
- **Reduce Dependence on Annotated Data:** Instead of requiring manually labeled datasets, which are expensive to produce and prone to bias, the method should be capable of learning from large amounts of text data without extensive human intervention.
- **Track Changes in Sentiment Over Time:** Unlike snapshot-based sentiment models, the approach should allow for continuous tracking of sentiment evolution, identifying trends and shifts in emotional expressions over different time periods.

2.3 Leveraging Language Models (LMs)

This research explores whether **Masked Time-Aware Language Models (MTALMs)** can overcome these limitations and provide a more effective approach to sentiment tracking. Unlike traditional methods:

- **No Need for Annotated Data:** MTALMs leverage self-supervised learning techniques, predicting missing words in large text corpora rather than requiring human-labeled datasets. This removes a major bottleneck in sentiment research.
- **Context-Aware Understanding:** These models do not rely on fixed word lists; instead, they infer meaning based on surrounding words. For example, MTALMs can recognize that "cold" in "cold winter" refers to temperature, while "cold attitude" refers to emotional distance.
- **Adaptability to Evolving Language:** Because MTALMs are continuously trained on new data, they can detect emerging language trends. If a new slang term like *simp* (used to

describe someone overly eager in romantic interest) gains popularity, an MTALM will recognize and incorporate its meaning over time.

– **Ability to Track Emotional Shifts Over Time:** By training models separately on different time periods (e.g., tweets from 2019 vs. 2023), we can analyze how sentiment changes over time. This helps in studying how events—such as the COVID-19 pandemic or political shifts—affect public emotions.

2.4 Bridging the Methodological Gap

Public sentiment plays a critical role in shaping **markets, politics, and decision-making**. However, traditional sentiment analysis methods fail to reflect the **dynamic nature of public opinion**. By leveraging Masked Time-Aware Language Models, this research aims to provide a novel methodology that accurately tracks, analyzes, and predicts emotional trends over time. This approach bridges the gap between static sentiment analysis and the ever-changing linguistic landscape, offering a more robust and context-aware solution for understanding public sentiment.

> In summary, this thesis addresses a fundamental gap in sentiment analysis: despite abundant data and powerful language models, we lack a systematic approach to track how public sentiment evolves over time. Current methods excel at analyzing sentiment at single points in time, but fail to capture how opinions change and respond to events across different periods. This limitation significantly constrains our ability to understand the dynamic nature of public opinion and to predict emerging sentiment trends before they become obvious through traditional means.

2.5 Research Questions

This investigation is guided by the following primary research question:

> Can MTALM detect changes in people's opinions over time?

To systematically address this inquiry, the research is structured around three specific research questions:

1. **Q1: Can MTALM detect changes in people's opinions over time?** This question seeks to determine whether MTALM, with its time-aware capabilities, can effectively identify and measure shifts in individuals' or groups' opinions as they evolve. It focuses on evaluating the model's ability to recognize temporal patterns in textual data that reflect opinion dynamics.
2. **Q2: If MTALM can detect these changes, what is the most effective approach to achieve this?** Building on the first question, this inquiry explores the optimal strategies

for leveraging MTALM to track opinion changes. It specifically examines two key considerations:

a) *Q2.1: Do we need to adjust the temporal shift output?* This sub-question investigates whether we need to shift the result to adjust the time lag in train process.

b) *Q2.2: Which similarity matrices should we use?* This sub-question addresses the selection of appropriate similarity measures or matrices to give best results.

3. **Q3: Can MTALM outperform existing methods for tracking changes in emotions and opinions?**

   This question evaluates the comparative effectiveness of MTALM against traditional approaches. It is further divided into two specific sub-questions:

   a) *Q3.1: Can MTALM outperform dictionary-based methods?* This sub-question assesses whether MTALM's advanced language modeling capabilities surpass the limitations of rule-based, dictionary-dependent techniques in capturing nuanced opinion and emotion shifts.

   b) *Q3.2: Can MTALM outperform sentiment analysis-based methods?* This sub-question examines whether MTALM can provide superior performance compared to conventional sentiment analysis techniques.

By addressing these research questions, this study aims to advance both the theoretical understanding of temporal sentiment dynamics and the practical tools available for tracking these patterns, contributing to the broader field of computational social science while offering actionable intelligence for decision-makers navigating increasingly dynamic information environments.

2.6 Research Goal

The primary goal of this research is to develop and evaluate Masked Time-Aware Language Models (MTALMs) that can accurately capture and predict temporal shifts in public sentiment across diverse domains. By training multiple language models on data from different time periods, this research aims to create a framework that enables more precise tracking of how opinions and emotions evolve over time.

Specifically, this research seeks to accomplish four interconnected objectives:

1. **Methodological Framework Development:** Create a comprehensive framework for training and deploying time-specific language models optimized for sentiment analysis, incorporating temporal information directly into model architecture.

2. **Architectural Optimization:** Identify the most effective model architectures and training strategies for temporal sentiment analysis through systematic comparative evaluation.

3. **Empirical Validation:** Rigorously validate the predictive accuracy of MTALMs against established sentiment measurement tools, including traditional surveys and existing computational methods.

4. **Application Demonstration:** Showcase practical applications of temporal sentiment analysis across business intelligence, policy assessment, and political analysis to demonstrate real-world utility.

The ability to accurately predict temporal shifts in public sentiment before they become apparent in traditional metrics offers substantial strategic advantages across multiple sectors—from financial markets and public health to corporate strategy and political campaigning.

3

Related Work

3.1 Introduction

The proliferation of user-generated content on the internet has provided researchers with an unprecedented opportunity to observe and analyze people's opinions, sentiments, and emotions on a large scale. This wealth of data has been instrumental in understanding collective attitudes and emotional states across various domains [12, 23]. However, the majority of opinion mining research using user-generated data has primarily focused on static sentiment analysis, which provides a snapshot of sentiment at a particular point in time [20, 5].

While static sentiment analysis offers valuable insights, it fails to capture the dynamic nature of human emotions and opinions, which can fluctuate rapidly in response to events, interactions, and personal experiences [41]. Recognizing this limitation, there is a growing need for methods that can detect and analyze these temporal changes in sentiment, as such information can provide more nuanced and actionable insights. This represents a significant gap in the literature, as understanding the dynamic nature of sentiment and emotion is crucial for gaining a more comprehensive understanding of human behavior and the factors that shape it [15]. Recent advancements in dynamic sentiment analysis aim to address this gap. The two main approaches that have been employed to calculate emotion dynamics are:

1. Dictionary-based sentiment analysis

2. Language model-based sentiment analysis

This chapter provides a comprehensive review of these approaches, their limitations, and the emerging research on time-aware language models that offer potential solutions to the challenges of dynamic sentiment analysis.

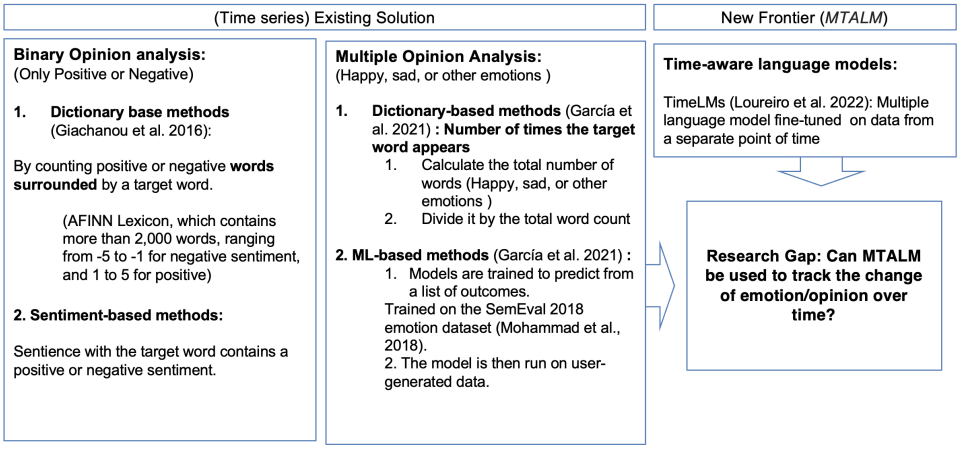

Figure 3.1: *Overview of related work and research gap*

3.2 Dictionary-Based Approaches

Dictionary-based sentiment analysis relies on predefined lexicons or dictionaries to determine the emotional tone of textual data. These approaches have been widely used due to their simplicity and interpretability, but they also face significant limitations in capturing the nuanced and context-dependent nature of human emotions.

3.2.1 Binary Dictionary-Based Methods (Positive or negative)

Dictionary-based sentiment analysis leverages predefined lexicons to determine the sentiment of textual data. One widely used lexicon is the AFINN lexicon [39], which contains over 2,000 words assigned scores ranging from -5 (strongly negative) to $+5$ (strongly positive). The overall sentiment of a text (e.g., a tweet) is computed by summing the scores of its sentiment-bearing words.

The process can be broken down into the following detailed steps:

1. **Temporal Aggregation:** Instead of examining individual tweets separately, this methodology groups tweets into weekly cohorts. Each week's collected tweets, represented as (denoted as $T1, T2, T3, \ldots, Tn$). This approach allow us to get results for each different weeks, and see how it is changing over time.

2. **Target Word Selection:**
   Identify a key term central to your analysis. For example, if you are analyzing political sentiment, you might choose the name "Merkel" as the target word. This selection is crucial because it defines the focus of the sentiment analysis , ensuring that subsequent steps are contextually relevant.

3. **Contextual Word Extraction:**
   Collect all tweets or textual data that include the target word. For each occurrence, extract a

window of words surrounding the target word to capture its context.
Example: Consider the tweet: "Merkel's performance was amazing during the debate." If a window of three words on each side is used, the extracted context might be: "performance was amazing during the debate." This context helps to accurately capture the sentiment related to the target word.

4. **Sentiment Scoring:**
For every word extracted in the previous step, assign a sentiment score using the AFINN lexicon. Each word in the lexicon has an associated score (e.g., "amazing" might score $+3$, while "terrible" might score -2).
The overall sentiment score for the extracted context is computed as:

$$\text{Sentiment Score} = \sum_{i=1}^{n} \text{score}(w_i)$$

where w_i represents each word in the extracted context. This summation produces a single sentiment value that reflects the polarity of the text segment.

5. **Trend Analysis with Moving Average:**
Finally, to smooth out short-term fluctuations and highlight longer-term trends, apply a moving average to the aggregated sentiment scores.
For instance, using a 7-day moving average, the smoothed sentiment score at day t can be calculated as:

$$\text{MA}_t = \frac{1}{7} \sum_{i=t-6}^{t} \text{Score}_i$$

This technique helps to reduce noise and reveal significant shifts in public opinion over time [21].

This method has been widely used to track public sentiment in various domains, including political discourse, market analysis, and crisis management [22, 7]. It can also be extended to analyze a range of emotions such as joy, fear, and sadness, thereby providing a more comprehensive understanding of the underlying sentiment dynamics.

3.2.2 Non-Binary Dictionary-Based Methods

Numerous studies have analyzed emotion, opinion, and sentiment analysis in recent years. However, most of them are focused on binary sentiment like position or negative [20, 46, 29]. Few studies have delved into multiple opinion detection [42]. One notable non-binary approach is the Linguistic Inquiry and Word Count (LIWC) framework utilized by García et al. [20]. This method offers a structured approach to transforming user-generated Twitter content into meaningful time series data that can reveal temporal patterns in sentiment and emotional expression.

LIWC-Based Tweet Processing Methodology

Dictionary-based methods such as LIWC involve analyzing text by comparing it against predefined lexical dictionaries [20, 42]. These dictionaries categorize words into thematic or emotional groups. For example, the LIWC dictionary includes categories like "positive emotion," "negative emotion," "anxiety," "anger," and "sadness." The text is processed by counting the occurrences of words belonging to these predefined categories, and the results are then used to infer the overall sentiment or emotional tone of the text.

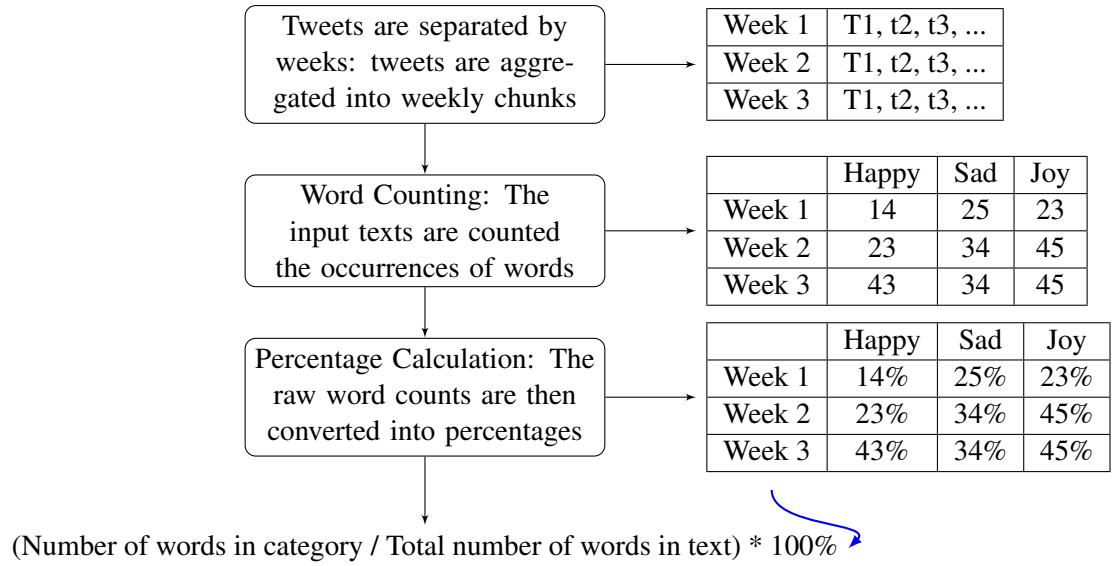

Figure 3.2: *LIWC (Linguistic Inquiry and Word Count) Process*

The LIWC approach employs a three-stage process for analyzing Twitter content over time:

1. **Temporal Aggregation:** Rather than analyzing individual tweets in isolation, the methodology aggregates tweets into weekly collection of tweets. Each week's collected tweets (denoted as $T1, T2, T3, \ldots, Tn$) form a coherent analytical unit. This approach enables us to obtain results for individual weeks and track how they evolve over time.

2. **Lexicon-Based Word Counting:** Once tweets are aggregated by week, the text undergoes processing through the LIWC dictionary. This specialized lexicon categorizes words into psychologically meaningful categories, including emotional dimensions such as "Happy," "Sad," "Joy," and "Scared." The system counts occurrences of words belonging to each category, producing raw frequency counts for each emotional dimension within each weekly period.

3. **Proportional Representation:** To normalize the data and enable meaningful comparison across weeks with potentially different volumes of content, raw word counts are converted to percentages. This calculation follows the formula:

$$\text{Category Percentage} = \frac{\text{Number of words in category}}{\text{Total number of words in text}} \times 100\% \tag{3.1}$$

This normalization ensures that the relative prominence of each emotional category is comparable across time periods regardless of variations in the total volume of tweets [41].

As demonstrated in empirical studies, this approach reveals notable patterns in emotional expression across time. For instance, research has shown that the "Happy" category might show consistent upward or downward trends that correlate with external events, while other emotional categories might exhibit different patterns. Such time-series data provide valuable insights into temporal shifts in collective emotional expression that might be correlated with external events or other factors of interest [20, 45].

Limitations of Dictionary-Based Approaches

While dictionary-based approaches enable researchers to track the evolution of emotions and opinions over time, they face several important limitations that have motivated the development of more advanced techniques:

- **Lexical Bias:** The creation of sentiment dictionaries can be subjective, with the personal biases of the lexicon developers influencing how words are categorized as positive or negative [38]. For instance, the word "volatile" may be seen as negative in financial contexts but positive when describing an energetic person.
- **Lack of Semantic Understanding:** Traditional dictionaries lack the sophisticated language comprehension of modern models like BERT or GPT-3 [18, 11]. These advanced models can better grasp the nuanced meanings of words and phrases based on the context. For example, if a tweet says, "The new policy is a disaster," a dictionary-based approach might only flag "disaster" as negative, missing potential sarcasm or hyperbole.
- **Context Dependency:** The sentiment of a word can change dramatically depending on the overall context of the sentence or passage [47]. For example, "painful" can have a negative meaning when referring to physical discomfort but a positive one when describing emotional growth. If someone tweets, "That concert was painfully good," a dictionary-based method might misinterpret the "painful" component as negative, failing to recognize the intended positive connotation.
- **Inability to Adapt to Language Evolution:** Fixed dictionaries struggle to keep up with rapid changes in language, such as the emergence of new terminology or shifts in word meanings over time [33]. For instance, terms like "woke" or "cancel culture" have evolved significantly in meaning and sentiment over the past few years, and a static dictionary would struggle to capture these changes.
- **Limited Coverage:** Dictionaries tend to focus on standard, formal language and may miss slang, dialects, or other non-standard forms of expression [31]. If someone uses slang like "This new track is fire," a dictionary-based system might fail to recognize the positive sentiment.

These limitations highlight the need for more sophisticated approaches that can better capture the nuanced, context-dependent, and evolving nature of human emotional expression in text.

3.3 Deep Learning-Based Sentiment Analysis Techniques

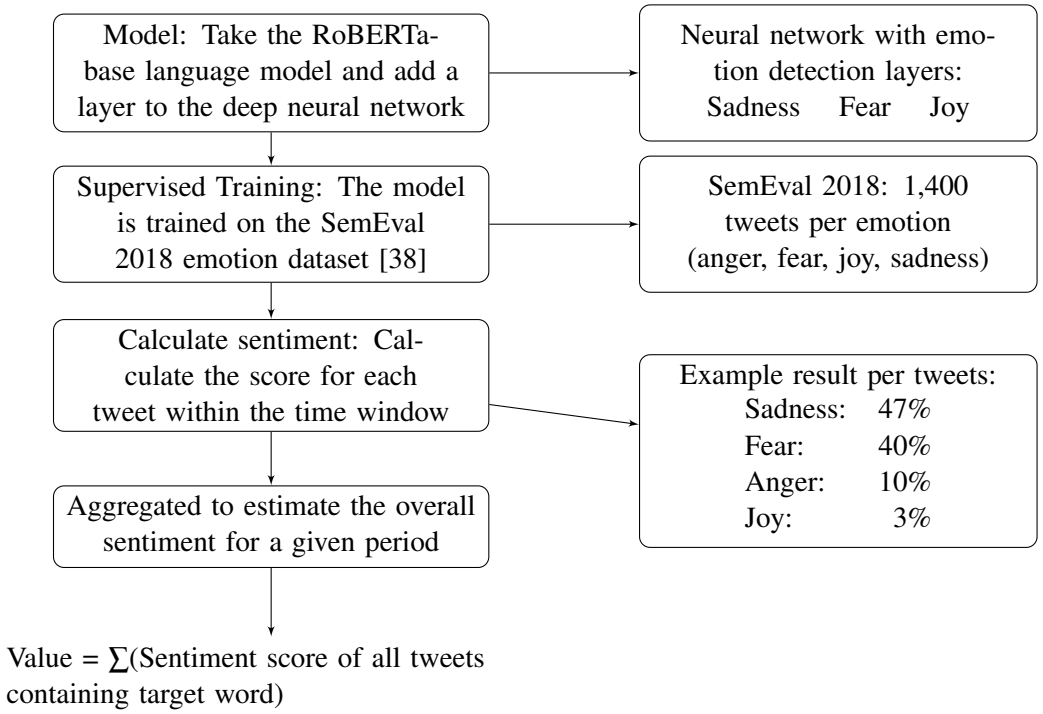

Figure 3.3: *Emotion macroscope - RoBERTa [20]*

3.3.1 RoBERTa Emotion Macroscope

The RoBERTa Emotion Macroscope [20] offers a supervised approach to emotion detection in social media text, as illustrated in Figure 3.3. This method follows a structured process:

1. **Pretrained Model:** The approach begins with the RoBERTa-base language model, a robust transformer architecture pre-trained on large text corpora. A specialized layer is added to the deep neural network specifically designed for emotion detection, enabling the model to identify emotional content in tweets.

2. **Supervised Training:** Unlike our MTALM approach, this model is explicitly trained on labeled emotion data from the SemEval 2018 dataset [38]. This dataset contains approx-

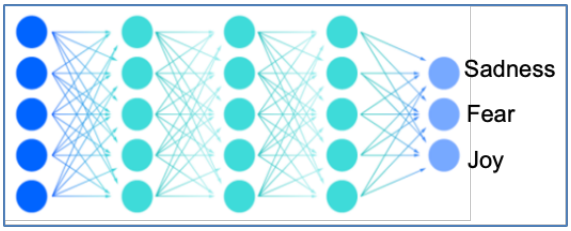

Figure 3.4: *RoBERTa Model Architecture for Emotion Classification*

imately 1,400 tweets per emotion category (anger, fear, joy, and sadness), providing the model with clear examples of how each emotion is expressed in social media text.

3. **Sentiment Calculation:** Once trained, the model calculates sentiment scores for each tweet within a specified time window. Each tweet is processed through the neural network, which assigns probability scores across the emotion categories based on the text content.

4. **Aggregation:** Individual tweet scores are then aggregated to estimate the overall sentiment for a given time period. This aggregation provides a comprehensive view of the emotional landscape during specific timeframes, allowing for temporal analysis of public sentiment.

The final value is calculated as the sum of sentiment scores for all tweets containing target words of interest. This approach enables researchers to track emotional trends over time and identify shifts in public sentiment related to specific topics or events.

As shown in the example period analysis in Figure 3.3, this method can produce detailed emotional profiles for each time period, such as the distribution showing 47% sadness, 40% fear, 10% anger, and 3% joy during a particular interval. These profiles can then be compared across different time periods to identify significant emotional shifts.

Deep learning-based approaches, particularly those utilizing transformer architectures, have emerged as powerful alternatives to dictionary-based methods for sentiment analysis. These approaches can better capture the contextual and nuanced nature of emotional expression in text.

Advantages of the RoBERTa Approach

The RoBERTa-based Emotion Macroscope offers several key advantages over traditional methods:

– **Contextual Understanding:** Unlike lexicon-based approaches that consider words in isolation, transformer models like RoBERTa account for the full linguistic context, capturing nuanced expressions of emotion that may be missed by simpler methods [34].

– **Multi-dimensional Analysis:** The model simultaneously detects multiple emotions, recognizing that a single text may express a complex mixture of emotional states with varying intensities (e.g., a text might exhibit 47% sadness, 40% fear, 10% anger, and 3% joy) [38].

- **Improved Accuracy:** Deep learning approaches generally demonstrate superior performance in emotion detection compared to dictionary-based methods, particularly for detecting implicit or subtle expressions of emotion [2].

- **Transfer Learning Efficiency:** By building upon a pre-trained language model, the approach leverages transfer learning to achieve high performance even with relatively modest amounts of emotion-labeled training data [17].

Limitations of the RoBERTa-base and Dictionary-based Approaches

The RoBERTa-based Emotion Macroscope offers several key advantages but still faces significant limitations that necessitate a new approach:

- **Limited Adaptability to Different Target Scenarios:** RoBERTa models trained on specific emotion categories (such as happy, sad, loneliness, apathetic, angry) are fundamentally constrained to those predefined emotions. This rigidity presents a major obstacle when need to analyze different emotional dimensions. For example, a model trained to detect the five basic emotions cannot be easily repurposed to analyze more nuanced states like "anticipation" or "contentment" without extensive retraining. This lack of flexibility becomes particularly problematic in longitudinal studies where research questions may evolve over time, or in cross-cultural contexts where emotional expressions vary significantly. Dictionary-based approaches face similar constraints, as they rely on predefined lexicons that may not capture the full spectrum of emotional expression across different contexts.

- **Dependency on Supervised Annotated Data:** Both RoBERTa and traditional dictionary approaches require substantial amounts of labeled data, which presents several challenges. For RoBERTa models, high-quality emotion annotation is labor-intensive, expensive, and often subjective, leading to potential inconsistencies across annotators. The annotation process typically requires domain experts and multiple rounds of validation to achieve acceptable inter-annotator agreement. This requirement becomes particularly burdensome when analyzing specialized domains or less-resourced languages where annotated emotional data is scarce.

- **Dependency on quality Annotated Data:** The quality of the model is directly dependent on the quality and representativeness of the training data. If the annotated data have a biases or wrongly labeled, the resulting model will inherit these biases and blind spots. Dictionary approaches similarly depend on manually curated emotion lexicons, which may not keep pace with evolving language use, especially in informal contexts like social media.

These limitations highlight the need for more flexible, adaptable approaches to emotion detection that can accommodate evolving research questions, require less supervised data, and better capture the contextual and temporal dimensions of emotional expression.

3.4 Need for Time-Aware Language Models

Given the limitations of both dictionary-based and standard deep learning-based sentiment analysis techniques, there is a need for more advanced methods that can capture the dynamic shifts in public sentiment and emotions over time. This leads to exploring time-aware language models that can effectively track and analyze these changes [35, 33].

3.4.1 Transformer Architecture and Contextual Understanding

The Transformer architecture, introduced in the landmark 2017 paper "Attention Is All You Need" by [49], revolutionized natural language processing by eliminating the need for recurrence and convolutions that were previously considered essential for sequence modeling. Instead, it relies entirely on attention mechanisms to draw global dependencies between input and output.

Key Components

1. **Encoder-Decoder Structure:** The Transformer follows an encoder-decoder architecture common in sequence-to-sequence models. The encoder maps an input sequence to a continuous representation, which the decoder then converts into an output sequence. Both encoder and decoder are composed of stacks of identical layers (typically 6 in the original paper).

2. **Self-Attention:** The heart of the Transformer is the self-attention mechanism, which allows the model to weigh the importance of different words in a sentence when representing each word. For example, in the sentence "The animal didn't cross the street because it was too tired," self-attention helps the model understand that "it" refers to "the animal" by creating stronger connections between these words.

3. **Multi-Head Attention:** Rather than performing a single attention function, the Transformer uses multiple attention "heads" in parallel. Each head can focus on different aspects of the input, allowing the model to jointly attend to information from different representation subspaces. For instance, one attention head might focus on syntactic relationships while another captures semantic similarities.

4. **Positional Encoding:** Since the Transformer contains no recurrence or convolution, it has no inherent understanding of word order. To inject information about the position of words in the sequence, the model adds "positional encodings" to the input embeddings. These encodings use sine and cosine functions of different frequencies, allowing the model to learn to attend to relative positions.

5. **Feed-Forward Networks:** Each layer in both the encoder and decoder contains a fully connected feed-forward network applied to each position separately and identically. This consists of two linear transformations with a ReLU activation in between, allowing the model to process the attention outputs further.

6. **Residual Connections and Layer Normalization:** To facilitate training of deep networks, the Transformer employs residual connections around each sub-layer, followed by layer normalization. This helps combat the vanishing gradient problem and stabilizes training.

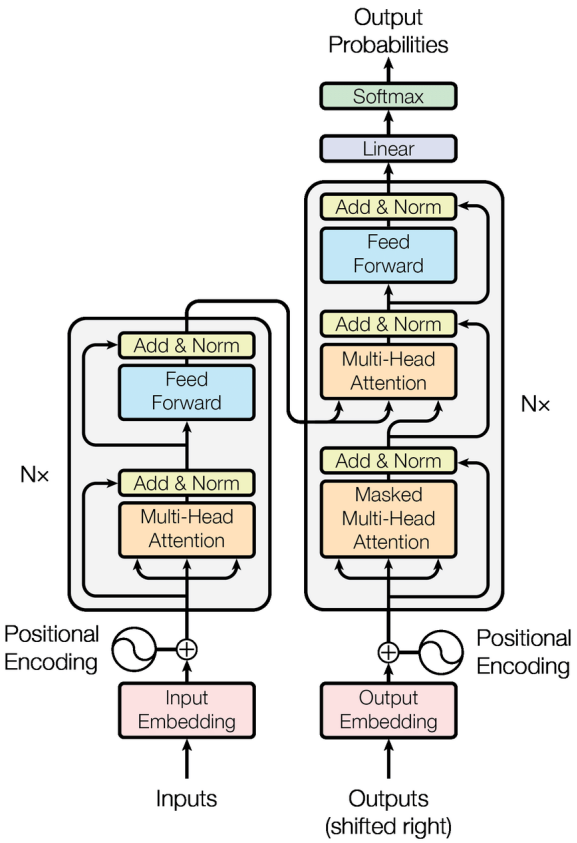

Figure 3.5: *Transformer Architecture [49]*

3.4.2 How It Works

1. **Input Processing:** Words are first converted to embeddings (vector representations) and combined with positional encodings to retain sequence order information.

2. **Encoder:** The input passes through multiple identical encoder layers. Each layer has two sub-layers:
   - A multi-head self-attention mechanism where each word attends to all words in the input sequence
   - A position-wise feed-forward network that processes the attention outputs

3. **Decoder:** The output sequence passes through multiple identical decoder layers. Each layer has three sub-layers:

- A masked multi-head self-attention mechanism that prevents attending to future positions
- A multi-head attention over the encoder output, allowing the decoder to focus on relevant parts of the input sequence
- A position-wise feed-forward network

4. **Output Generation:** The decoder's output is transformed through a linear layer and softmax function to produce probabilities for each word in the vocabulary.

3.4.3 Time-Aware Language Models

Recent research has focused on incorporating temporal information into transformer-based language models, enabling them to capture the evolving nature of language and sentiment over time [35, 33]. Since opinions and sentiments are constantly shifting, information that was accurate a few days ago may no longer be valid. For example, at the beginning of 2021, ChatGPT would not have recognized COVID-19 because its training data only included information up to a certain period. This highlights how a model's predictions depend on the timeframe of the data it was trained on. By training models on data from different time periods, they can better reflect the knowledge and sentiment of that specific moment.

Techniques for Incorporating Time Information

Several techniques can be used to incorporate time information into language models, enabling them to detect dynamic shifts in public sentiment and emotions:

1. **Time-Specific Training:** Train separate language models on data from different time periods. This allows each model to learn the unique characteristics and sentiment expressions prevalent during its corresponding time frame [35]. By fine-tuning these models on data from specific time periods, we can create time-aware language models that are better equipped to detect dynamic shifts in public sentiment and emotions.

   The paper *"Time-Aware Language Models as Temporal Knowledge Bases"* explores how language models (LMs) can be improved to better capture **time-sensitive information [19]**. Traditional LMs, such as BERT or GPT, are trained on large datasets but do not explicitly consider how knowledge and language change over time. As a result, they struggle with **temporal reasoning**—for example, predicting events or understanding that facts may evolve.

   How Time-Aware LMs Work

   - Train separate language models on data from different time periods.
   - Instead of treating knowledge as static, these models recognize that certain facts are tied to specific time periods.

Time-aware language models improve upon traditional LMs by **considering when knowledge is relevant**. This makes them better at handling dynamic topics such as news, politics, and scientific discoveries, where facts can change frequently.

2. **Temporal Embeddings:** Add temporal embeddings to the input sequence to represent the time period of each word or document. This allows the model to directly incorporate time as a feature in its analysis [33]. For instance, a temporal embedding e_t might be concatenated with the word embedding e_w for each token, resulting in a combined embedding $e_{combined} = [e_w; e_t]$.

3. **Time-Gated Mechanisms:** Use time-gated mechanisms to modulate the flow of information in the language model based on the time period, allowing the model to prioritize information from relevant time frames [36]. These mechanisms can be formulated as:

$$g_t = \sigma(W_g[h_t; e_t] + b_g)$$

Where:

- g_t is the gate value for time t.
- σ is the sigmoid function.
- W_g and b_g are the weights and biases for the gate.
- h_t is the hidden state at time t.
- e_t is the temporal embedding at time t.

3.4.4 Research Gap and Contribution

Despite the advancements in Transformer models and time-aware language models, there remains a significant research gap in the literature:

- RoBERTa base models perform relatively well but have limitation like lack of flexibility, reliance on labeled data, and sensitivity to data quality. Since these models are trained on specific emotions like happy, sad, or angry, they cannot easily detect new or more complex emotions like anticipation or contentment without retraining.

- Another challenge is that these models need a large amount of labeled data, which is expensive and time-consuming to create. For example, labeling emotions in text requires experts and multiple rounds of checking to ensure accuracy. If the training data has mistakes or biases, the model will learn them too, leading to incorrect results. Similarly, dictionary-based methods rely on fixed word lists, which may not capture how language evolves, especially in social media or informal conversations. These issues show the need for emotion detection models that can adapt to new situations, require less labeled data, and better understand emotions in different contexts.

– **No prior work has comprehensively explored the use of time-aware language models, specifically Masked Time-Aware Language Models (MTALMs), for quantitatively tracking and analyzing the dynamic changes in public sentiment and emotions over time.**

– Existing studies have primarily focused on static sentiment analysis or have not fully leveraged the potential of time-aware language models to capture the nuanced and evolving nature of public opinion [35, 33]. While existing solutions can analyze sentiment up to a point, this research seeks to overcome those limits by using MTALMs to better detect changes in people's opinions that can provide valuable information for various applications [41].

4

Experimental Setup

This chapter explains the experimental approach used to evaluate Masked Time-Aware Language Models (MTALMs) in tracking and analyzing changes in public sentiment over time.

First, we will use a pre-trained language model (LM) like BERTweet[18] as our base model. BERT has already been trained on a large amount of text data, so it understands language in a general way. We will then retrain this model every week using data from that specific week. Model will be trained on 200 weeks' worth of tweets, using Masked Language Model (MLM) approach. This means that instead of using labeled data (which would be too difficult because we have 70 million tweets), the model learns by predicting missing words in sentences.

Since each model is trained on data from a specific time period, we expect that these models will reflect the emotions common during those times. By retraining the model weekly, we hope to track how emotions change over time. For example, if there was an increase in feelings of happiness or sadness in a certain week, the model should reflect that change in its predictions.

The idea is to ask the language models the same questions that human participants in surveys are asked, and then compare the model's answers with actual survey results. These surveys are used to track how people's emotions or opinions change over time. By comparing the models' predictions to these surveys, we can assess how well the models capture emotional changes.

In summary, we want to test if the model can track changes in emotions over time, just as surveys track changes in people's opinions, and evaluate whether these changes align with real-world survey data.

4.1 Experimental Design

The experimental design of this study comprises three primary stages, as illustrated in Figure 4.1.

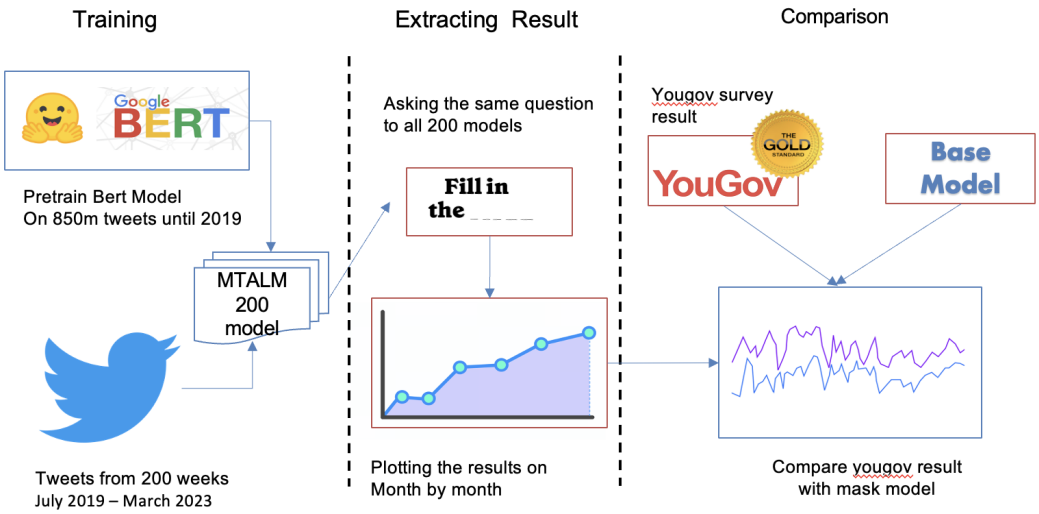

Figure 4.1: *Experimental Pipeline Overview*

Training Stage

1. A BERTweet (Bidirectional Encoder Representations from Transformers) model is pre-trained on a large corpus of Twitter data, consisting of approximately 850 million tweets up to the year 2019.

2. Subsequently, 200 specialized models, referred to as MTALMs (Masked Temporal Aware Language Models), are developed. Each MTALM is trained separately on weekly data spanning from 2019 to March 2023, allowing for a granular, temporal understanding of sentiment dynamics.

Extracting Results Stage

1. All 200 MTALMs are queried with identical prompts, designed to elicit sentiment-laden predictions.

2. The sentiment predictions from each MTALM are analyzed to discern the expressed sentiment for that particular prompt.

3. Results are plotted to create a temporal line graph, visualizing sentiment changes over time.

Comparison Stage

1. The predictions generated by the MTALMs are compared against established survey data from YouGov, a reputable source of public sentiment data.

2. This comparison serves to validate the accuracy of the MTALM predictions against real-world survey data.

3. Additionally, output will compare against existing research to see if it can outperform current approaches.

The experimental design allows for a systematic assessment of MTALMs' capability to capture temporal dynamics of sentiment over a 200 weeks period (2019-2023). By comparing model predictions against traditional survey methods, the experiment evaluates the reliability and validity of MTALMs in tracking and understanding temporal sentiment changes. This approach helps to identify any potential biases in the model compared to human-reported sentiment changes.

4.2 Training

4.2.1 Data Collection

The effectiveness of Masked Time-Aware Language Models (MTALMs) in tracking temporal sentiment changes relies fundamentally on access to appropriate data sources. This research requires two distinct types of data: a comprehensive social media dataset to train the models, and a gold standard reference sentiment measurements for validation.

Twitter Dataset for Model training

The Twitter dataset serves as the primary source for training our time-specific language models. The Twitter dataset was collected and processed as follows:

- **Data Extraction:** The complete Twitter history of UK users was extracted and divided into weekly segments.
- **Time Coverage:** The dataset spans from January 2019 to March 2023, encompassing over four years of social media activity.
- **Activity Filtering:** High and low activity periods were identified and filtered out to mitigate the impact of anomalous platform behaviors.
- **Temporal Segmentation:** The data was meticulously split into 200 distinct weekly intervals, creating a fine-grained temporal resolution for analysis.
- **Data Volume:** The complete dataset comprised approximately 70 million tweets, with an average of 350 thousand tweets per week.
- **Standardization:** To ensure consistent training conditions across all time periods, each weekly dataset was standardized to contain 200,000 tweets through undersampling and oversampling techniques. This standardization prevents training inequities that might confound temporal sentiment analysis by ensuring models from different periods receive comparable training exposure.

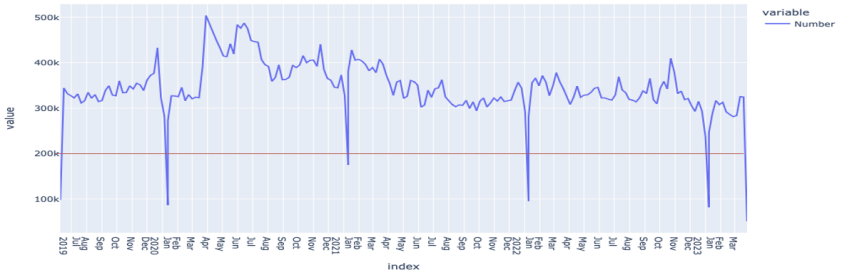

Figure 4.2: *Weekly Twitter data volume showing the standardization threshold at 200,000 tweets per week. The blue line represents the actual tweet count before standardization, which typically ranged between 300,000-400,000 tweets with occasional significant deviations.*

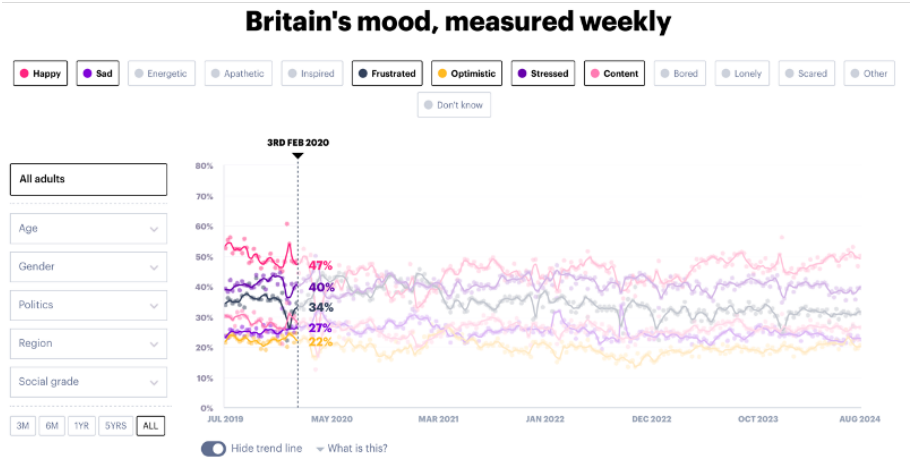

Figure 4.3: *Yougov survey about Britain's mood*

Reference Survey Data: Gold Standard

The YouGov survey data will be use to validate the MTALM predictions. The survey data was collected from YouGov [54] covering the same time period. These weekly surveys tracked emotions [1] and life satisfaction of the British population [53], as shown in Figure 4.3.

The YouGov surveys employed a consistent methodology throughout the study period, with approximately 2,000 UK adults sampled per week. Respondents directly reported their emotional states across multiple dimensions, providing a reliable benchmark against which to evaluate the MTALM. The survey data was temporally aligned with the corresponding Twitter data segments.

4.2.2 Data Preprocessing

The collected data underwent several preprocessing steps to clean and prepare it for model training and analysis. These steps include:

1. **Concatenation:** All individual weekly tweet sequences were concatenated into a single, large corpus and split into shorter chunks, each with a block size of 128 words. This serves two purposes:
   - It ensures the individual input sequences are shorter than the maximum input length supported by the model, preventing truncation that could lead to loss of valuable information.
   - It ensures the chunks are short enough to fit within the available RAM of the training GPU, enabling efficient processing during training.

   By combining all sequences, the model can learn from the collective data, ensuring no potentially meaningful information is discarded due to input length limitations. This approach allows the model to capture broader context and patterns present across the entire dataset.

2. **Removing Noise:** Irrelevant information, such as URLs, hashtags, and mentions, was removed from the tweet text. This step focused the analysis on the tweets' core content.

3. **Tokenization:** Tweet text was tokenized into individual words or sub-words using techniques suitable for language models. By using an established auto tokenizer, the model could efficiently parse the input text into a sequence of tokens, ensuring access to a comprehensive lexical understanding and enabling more accurate predictions during language modeling tasks.

4. **Stemming/Lemmatization:** The tokens were stemmed or lemmatized to reduce them to their root form. Stemming involves removing suffixes from words, while lemmatization involves converting words to their base or dictionary form. This step helps reduce vocabulary size and improve the model's ability to generalize.

5. **Masking:** Finally, 15 percent of the words within each chunk were randomly masked. This technique helps the language model learn to predict missing words, which is a core aspect of language modeling. By concealing a subset of words, the model is forced to leverage surrounding context to accurately infer and reconstruct missing information. This process challenges the model to develop a deeper understanding of language patterns and semantic relationships, strengthening its ability to capture the nuanced dependencies that underlie natural language. Masking during training has been shown to yield significant improvements in the model's overall language understanding capabilities.

4.2.3 Model Training

Multiple Masked Time-Aware Language Models (MTALMs) were trained, each on data from a specific time period. This approach allows each model to capture the unique characteristics and sentiment expressions prevalent during its corresponding time frame. For example, if the study involves three time periods (T1, T2, and T3), three MTALMs will be trained: MTALM-T1,

MTALM-T2, and MTALM-T3. MTALM-T1 will be trained on the data from time period T1, MTALM-T2 on the data from time period T2, and MTALM-T3 on the data from time period T3.

Training Methodology

1. **Model Selection:** A Transformer-based architecture, specifically BERTweet [9], forms the foundation for the MTALMs. As cited in the paper, "Unlike dictionary-based sentiment analysis, which relies on predefined lexicons, LLMs can learn the nuanced, contextual meanings of words and how they convey different emotional states."[9] This selection is motivated by BERT's proficiency in capturing contextual information and semantic relationships. The model have been fine tuned on 850million tweets.

   We chose this pre-trained language model because it was trained on data from before 2019. This ensures the model's understanding is not biased towards or influenced by the current data. It also ensures that the results we have received from the model are solely influenced by our training, not the pre-training. By using an older model, we can better capture the dynamic evolution of language and sentiment over time, as the pre-trained model's knowledge will not be skewed toward the dominant emotions or patterns that have emerged more recently.

2. **Masked Language Modeling (MLM):** MTALMs were trained using a Masked Language Modeling (MLM) objective, involving the random masking of words within input texts and training the model to predict those masked words. This approach fosters contextual understanding, which is consistent with the principle that language models are aware of context and can infer the meaning of a word based on its usage context [18].

3. **Time-Specific Training Data:** Each MTALM was trained on preprocessed data from a distinct time period. This ensures that each model captures unique characteristics and sentiment patterns specific to its corresponding time frame. This approach allows each model to capture the unique characteristics of its corresponding time frame.
   - **2**00 distinct language models were trained, each on data from a specific weekly time period between 2019-2023.
   - **M**odels were saved at each of the 3 training epochs, resulting in 600 total model checkpoints for comprehensive analysis.

4.3 Extracting results

Ask the same question

We will ask identical questions to all 200 weekly models, as shown in Figure 4.1. For example, we might ask each model to complete the sentence "The current mood in the country is <mask>."

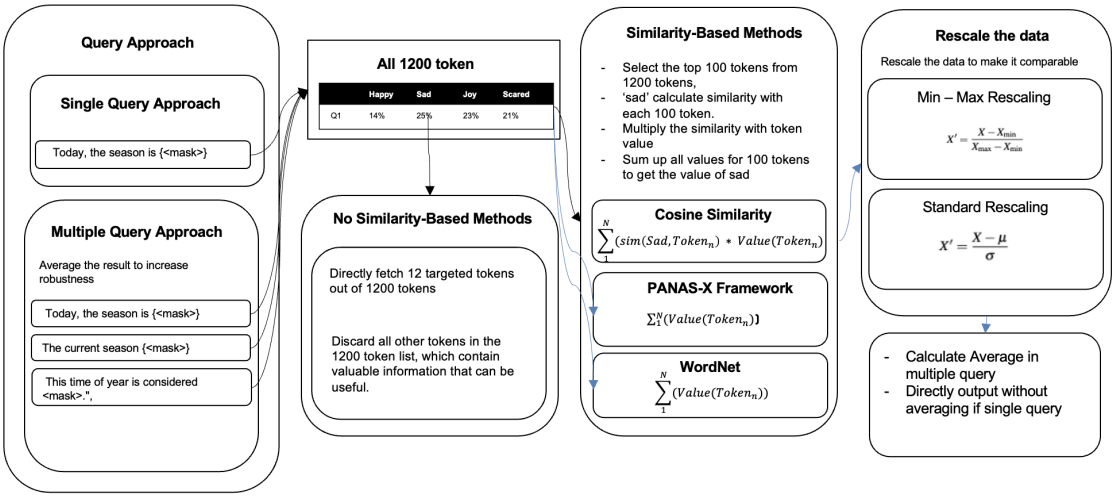

Figure 4.4: *Process of Extracting results*

This allows us to compare how models from different time periods predict different emotional words, revealing shifts in sentiment over time. When analyzing language model predictions, two main approaches can be used: the **Single Query Approach** and the **Multiple Query Approach**. Each approach has its own advantages and implications for the reliability of results. The trained MTALMs will be used to analyze sentiment from different time periods using the following process:

4.3.1 Query Approach

– **Single Query Approach** Using one question only and use the output as final results.

"The season we are in currently is `<mask>`."

– **Multiple Query Approach** uses different ways of asking the same question and then averages the results. Instead of relying on a single sentence, we might ask:

"Today, the season is `<mask>`."

"The current season is `<mask>`."

Since language models can give different responses depending on how a question is framed, using multiple variations helps smooth out inconsistencies. This makes the results more reliable and robust, as shown in Figure 4.4.

By using multiple queries, we reduce the risk of obtaining results that depend too much on specific wording. This approach also helps reveal patterns in the model's understanding and highlights its strengths and weaknesses in different contexts. In turn, it improves confidence in the model's ability to track changes in language and emotional expression over time.

4.3.2 Similarity-Based Methods

In traditional approaches, we might only count when a model predicts the exact word "happy" to measure positive sentiment. However, this approach misses similar emotions expressed through words like "joyful," "delighted," or "cheerful." Similarity-based methods solve this problem by considering semantic relationships between words.

For example, if we're tracking sadness over time, we shouldn't just count instances of the word "sad." Our model might predict "devastated" or "gloomy" - words that express the same fundamental emotion but would be ignored in a strict word-matching approach.

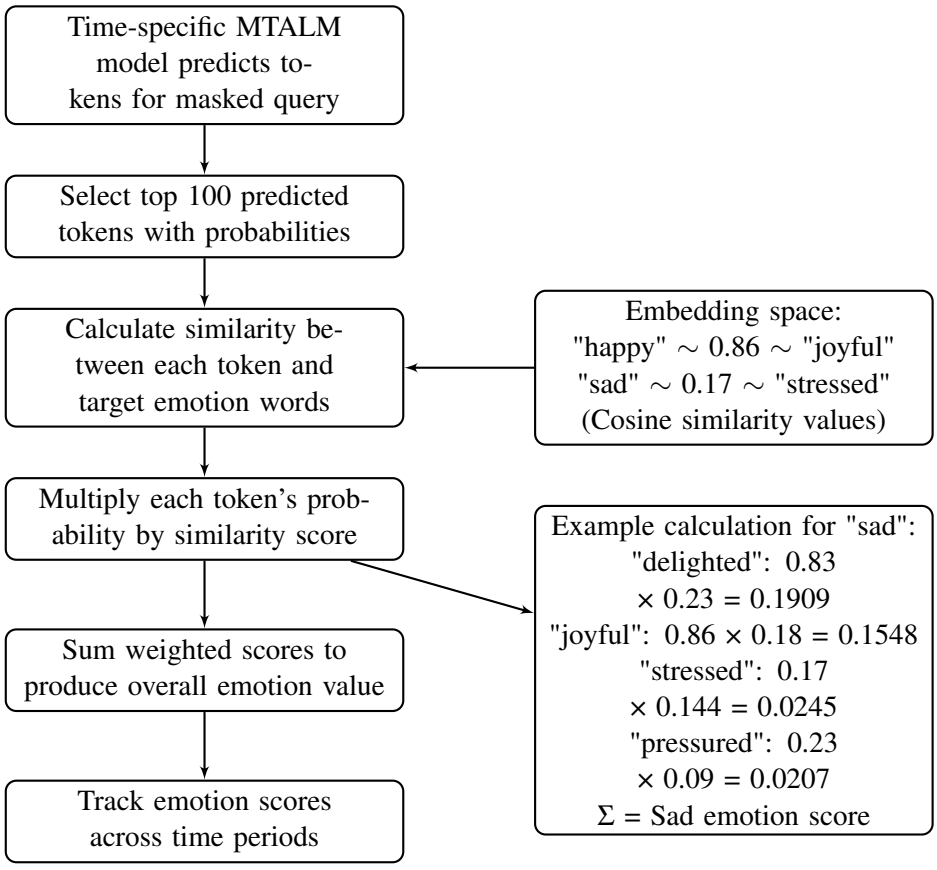

Figure 4.5: *Process flow for calculating emotion scores using similarity-based approach*

How Similarity-Based Methods Work

Our approach uses a weighted combination of related words:

1. We select the top predicted tokens from each model (typically the top 100)
2. For each target emotion (e.g., "sad"), we calculate its similarity to each predicted token using cosine similarity in the embedding space
3. We multiply each similarity score by the token's prediction probability

4. We sum these weighted scores to get an overall value for that emotion

As shown in Figure 4.5, when calculating sadness, words like "stressed" (similarity: 0.17) and "pressured" (similarity: 0.23) contribute to the score, but less than words like "joyful" which have negative similarity with "sad."

The formula can be expressed as:

$$Sad = \sum_{n=1}^{N} (sim(Sad, Token_n) \times Value(Token_n)) \tag{4.1}$$

This method provides a more comprehensive understanding of sentiment by considering both the model's confidence in each prediction and how semantically similar each predicted word is to our target emotions. It helps us capture the full spectrum of emotional expression rather than relying on a small set of exact word matches. Aggregate the sentiment scores or emotion categories for each time period to obtain an overall measure of public sentiment during that time. Sentiment scores will be extracted from the MTALM outputs to quantify the emotional tone of texts. Methods for doing this may include using the model's predictions for masked words or fine-tuning for sentiment classification tasks. Sentiment scores will be aggregated for each time period to provide an overall measure of public sentiment during each specific interval. The distribution of sentiment scores can also be analyzed to gain insights into prevailing emotions.

1. Utilizing similarity-based methods can provide a more comprehensive and nuanced approach to analyzing the model's sentiment predictions. When we focus solely on a single token like "happy", we may be overlooking important contextual information and related emotional concepts that could be equally relevant. Words like "joy", "elation", and "cheerfulness" share similar semantic meanings and could offer valuable insights into the model's understanding of the emotional landscape.

2. Furthermore, the relative position of the predicted tokens can also be informative. Tokens that rank higher in the model's output are often more accurate representations of the underlying sentiment. If a related term like "joy" is predicted with a higher probability than "happy", it could indicate a more precise capture of the emotional state expressed in the text.

Cosine Similarity:

This metric measures the cosine of the angle between two vectors, providing a way to quantify the similarity between the predicted tokens and a reference set of emotionally relevant terms [24]. By comparing the cosine similarity between the model's output and a lexicon of emotion-related words, we can gain insights into how well the model is capturing the underlying sentiment.

PANAS-X:

The Positive and Negative Affect Schedule - Expanded Form is a well-established framework for assessing a broad range of emotional states [37]. By mapping the model's predicted tokens to the PANAS-X lexicon, we can evaluate how accurately the model is representing the different dimensions of emotion, such as positive affect, negative affect, and specific emotional states like fear, anger, and joy.

WordNet:

WordNet is a lexical database that groups English words into sets of synonyms, called synsets, and provides semantic relations between these synsets [50]. By leveraging WordNet, we can identify semantically similar words to the model's predicted tokens and expand the analysis beyond exact matches. This approach allows us to capture a more nuanced understanding of the model's ability to represent the contextual and associative aspects of sentiment and emotional expression.

By incorporating these three similarity-based methods, we can gain a deeper and more comprehensive understanding of the language models' performance in predicting and representing the dynamic changes in sentiment and emotional expression over time. This multi-faceted approach will enable us to identify the strengths, limitations, and nuances of the models' outputs, ultimately guiding the most effective strategies for leveraging these models to capture the temporal shifts in language use and emotional expression within the social media data.

4.3.3 Data Aggregation and Rescaling

To compare our model's predictions with YouGov survey data (Figure 4.3), we need to adjust them to the same scale because they use different measurement methods. Without rescaling, the numbers would not be directly comparable, making it difficult to evaluate how well our model aligns with survey results.

Our model works by distributing probabilities across 1,200 possible words, ensuring that their total sum is always 1. In contrast, the survey data only considers the top 12 words and adds up their scores. This difference in scale means that our model spreads its predictions over a much larger set of words, while the survey focuses on a small subset. To make a fair comparison, we rescale the data so that both sources are on the same level. This allows us to see more clearly whether our model's predictions match public sentiment as measured in the surveys.

- **Min-Max Scaling:** Min-Max Scaling is a normalization technique that rescales data to a fixed range, usually between 0 and 1. It ensures that the minimum value in the dataset maps to 0, the maximum maps to 1, and all other values are proportionally adjusted within this range. The formula for Min-Max Scaling is:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{4.2}$$

where:

- X is the original data value,
- X_{min} is the minimum value in the dataset,
- X_{max} is the maximum value in the dataset,
- X' is the scaled value in the range [0,1].

This method preserves the relationships between values but can be sensitive to outliers, as extreme values will directly influence the scaling.

– **Standardization:** Standardization, also known as Z-score normalization, transforms data to have a mean of 0 and a standard deviation of 1. This method is particularly useful when the data follows a Gaussian distribution and is beneficial for models that assume normally distributed input features. The formula for Standardization is:

$$X' = \frac{X - \mu}{\sigma} \tag{4.3}$$

where:

- X is the original data value,
- μ is the mean of the dataset,
- σ is the standard deviation of the dataset,
- X' is the standardized value.

Unlike Min-Max Scaling, Standardization is not affected by outliers because it does not depend on the absolute minimum and maximum values. Instead, it transforms the data into a standard normal distribution, making it useful for many machine learning models.

By applying these rescaling methods, the model predictions and survey data can be placed on a common scale, enabling more robust and meaningful comparisons. This will help evaluate the models' performance in capturing the dynamic shifts in sentiment and emotional expression over time, ultimately guiding the most effective approach for modeling these temporal changes in the social media data. After rescaling, we aggregate the data by week to create a timeline of sentiment evolution throughout the 200 weeka period, allowing us to observe how public emotions responded to major events and changed over time.

5

Evaluation

Evaluating the effectiveness of the Masked Time-Aware Language Models (MTALMs) in capturing dynamic shifts in public sentiment requires rigorous benchmarking against both established ground truth data and baseline sentiment analysis methods. The evaluation process is designed to measure how well MTALMs align with real-world sentiment trends and whether they provide improvements over traditional approaches.

Two primary aspects will be assessed:

- **Comparison to Gold Standard Data:** The sentiment trends predicted by MTALMs will be compared against survey-based ground truth data, such as the **YouGov** survey [52]. This will validate whether the model accurately captures shifts in public sentiment.
- **Comparison with existing work:** The performance of MTALMs will be benchmarked against traditional sentiment analysis techniques, including dictionary-based approaches and sentiment classification models (e.g., RoBERTa). This will measure the added value of time-awareness in sentiment tracking.

5.0.1 Comparison to Gold Standard Data

To validate the accuracy of MTALMs, their sentiment trends will be compared against ground truth data, specifically the **YouGov** survey [52]. This survey collects public opinion on various topics over time, making it a reliable benchmark.

- **Visual Comparison:** The sentiment trends predicted by the model will be plotted alongside the YouGov survey results to observe how closely they align.
- **Statistical Correlation:** The degree of correlation between the model's sentiment predictions and the YouGov data will be measured using correlation coefficients (e.g., Pearson correlation). A strong correlation would indicate that the model successfully captures real-world sentiment shifts.

For example, if the YouGov survey reports an increase in public anxiety during a major global crisis (e.g., a pandemic or financial downturn), we expect the MTALM-trained model to reflect a similar rise in negative sentiment during that period.

5.0.2 Comparison with existing Methods

To assess the advantage of MTALMs over traditional sentiment analysis approaches, their performance will be compared against three baseline methods:

- **Base Model Without Training:** The base BERTweet model (without time-aware retraining) will be tested to see if it already contains temporal sentiment information.
  - The model will be queried with a timestamp (e.g., *"What was the public sentiment about the economy in March 2020?"*).
  - If the base model produces results similar to the trained MTALM, it suggests that temporal information is already embedded in the original model and retraining might be unnecessary.
- **Dictionary-Based Methods:** A dictionary-based approach [20] will be used to extract sentiment trends from social media data.
  - This approach classifies sentiment based on predefined word lists. In simple terms, it counts the relative frequency of words belonging to a specific sentiment category.
  - For example, to measure sadness, the method counts the occurrences of words associated with sadness (e.g., *unhappy, sorrow, grief*) relative to the total word count. A higher relative frequency indicates a stronger presence of that sentiment.
  - The results from this approach will be compared to MTALM predictions to assess whether contextual learning in MTALMs provides a significant improvement over rigid word lists.
- **Sentiment-Based Methods:** A sentiment classifier, such as a fine-tuned RoBERTa model [20], will be used to generate time-series sentiment predictions.
  - This method calculates sentiment using a pre-trained language model (LM) like RoBERTa. The model is trained to classify tweets into specific emotions using sentiment analysis techniques.
  - The process involves:
    1. Labeling each tweet with an emotion (e.g., happiness, anger, sadness) using the RoBERTa model.
    2. Summing the predicted emotions across all tweets within each week.
    3. Comparing weekly emotion scores to track changes over time.

- The sentiment-based results will be compared with MTALMs to evaluate their accuracy in capturing nuanced emotional changes over time.

5.0.3 Evaluation Metrics

In evaluating our MTALM approach against established gold standards, we employ several statistical metrics that quantify the strength, direction, and significance of relationships between our model's output and benchmark data. These metrics provide rigorous mathematical validation of our approach's effectiveness in capturing temporal patterns in sentiment analysis.

Correlation Coefficient (r_{xy})

The Pearson correlation coefficient measures the linear relationship between two time series variables:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5.1}$$

Where:

- n is the number of time points in our analysis
- x_i, y_i are the sentiment values at time point i for our model and the gold standard, respectively
- $\bar{x}, \bar{y}$ represent the respective means across all time points

This metric ranges from -1 to 1, where:

- Values close to 1 indicate a strong positive correlation (both time series move in the same direction)
- Values close to -1 indicate a strong negative correlation (time series move in opposite directions)
- Values near 0 suggest no linear relationship

In our context, strong positive correlations between MTALM outputs and survey data validate that our model correctly captures temporal shifts in public sentiment.

Coefficient of Determination (R^2)

The R^2 value represents the proportion of variance in the gold standard that is predictable from our model:

$$R^2 = r_{xy}^2 \tag{5.2}$$

This metric ranges from 0 to 1, where:

- Values close to 1 indicate our model explains most of the variance in the gold standard
- Values close to 0 indicate our model has little explanatory power

For our time-aware analysis, higher R^2 values demonstrate that MTALM captures meaningful temporal patterns rather than random fluctuations.

Statistical Significance (p-value)

The p-value tests whether the observed correlation could occur by random chance:

$$p\text{-value} = 2 \times \Pr(t_{n-2} > |t|) \tag{5.3}$$

Where:

$$t = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}} \tag{5.4}$$

A p-value less than the significance level (typically 0.05) indicates the correlation is statistically significant, providing confidence that the relationship between our model's predictions and the gold standard is not due to chance.

Critical Values

Critical values establish thresholds for determining statistical significance based on sample size. For correlation analysis, we use the critical value r_{critical} to evaluate whether correlations are significant:

$$r_{\text{critical}} = \frac{t_{\text{critical}}}{\sqrt{n-2+t_{\text{critical}}^2}} \tag{5.5}$$

Where t_{critical} is based on the t-distribution with $n-2$ degrees of freedom and the desired significance level.

When $|r_{xy}| > r_{\text{critical}}$, we can confidently reject the null hypothesis of no correlation. In our analysis, we use:

- $\alpha = 0.05$ for 95% confidence level
- $\alpha = 0.01$ for 99% confidence level

Application to Time Series Comparison

These metrics collectively enable rigorous comparison between our MTALM-generated time series and gold standards:

1. **Temporal Alignment**: Correlation coefficients reveal how well our model tracks temporal shifts in sentiment, including lag effects where sentiment changes in social media precede or follow survey responses.

2. **Reliability Assessment**: Critical values and p-values provide statistical guardrails, ensuring observed relationships exceed random chance and represent genuine patterns.

3. **Cross-Method Validation**: These metrics allow direct comparison between different approaches (MTALM, LIWC, RoBERTa), objectively demonstrating where our time-aware approach offers improvements.

By applying these metrics to sliding time windows (k-weekly text windows), we can identify optimal temporal alignment between social media sentiment and survey responses, revealing the most effective predictive horizons for different emotional dimensions.

6

Results

6.1 Can it detect the change in people's opinion change?:

This section presents the results for the research question: *Can MTALM detect changes in people's opinions over time?* It evaluates the performance of Masked Time-Aware Language Models (MTALM) by compareing the result with the base model (un-tuned LM). If MTALM outperform (un-tuned LM) base model, it means it have gained knowledge through training. Which will help us to answer research question Q1.

Base (Untuned) MTALM Approach

The base untuned LM represents our initial approach to comparing our results. As illustrated in Figure 6.1, this approach utilizes a pre-trained BERTweet model without any temporal context adaptation. The base model operates through a specific three-step process:

1. **F**irst, we begin with the standard BERTweet model, which has been pre-trained on general Twitter data but lacks any specific temporal awareness. We intentionally choose a model that have been trained on data before 2019, so that the pre-training process don't influence the results.

2. **S**econd, we explicitly inject temporal information by reformulating queries to include specific date and location markers. As shown in the right side of Figure 6.1, we modify the standard emotion detection prompt to include precise temporal anchoring: "Today is Wed, December 4^{th}, 2021. Location UK. Which of the following best describes your mood in the past week? <mask>."

3. **T**hird, we evaluate the model's performance by comparing its predictions with actual sentiment data from the corresponding time periods. The results, as noted in the text, show

"overall less correlation in the base model. We want to see if training increase the corelation of compared with the gold standard.

This baseline approach establishes an important reference point against which we can measure the improvements gained through temporal fine-tuning, demonstrating that merely providing temporal context in prompts is insufficient for accurate temporal sentiment analysis.

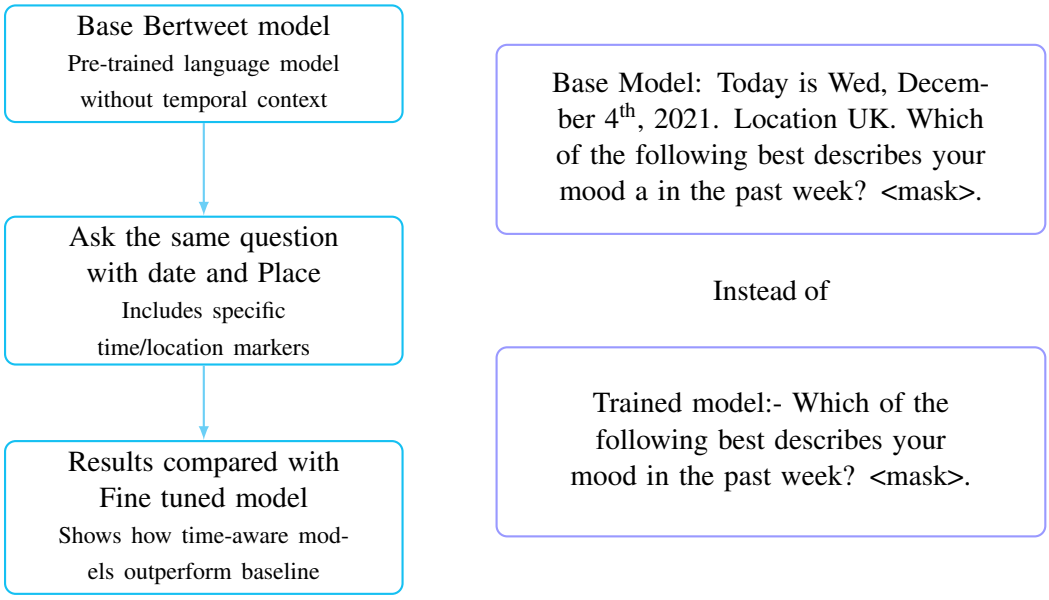

Figure 6.1: *Comparison of Base Model and Fine-tuned Model Approaches*

6.1.1 Q1:Can MTALM detect changes in people's opinions over time?

This analysis aims to answer the research question: **Q1: Can MTALM detect changes in people's opinions over time?**

The results suggest that MTALM, in its unfine-tuned form, is not well-suited to detect changes in opinions over time due to weak correlations and poor performance. However, with fine-tuning, MTALM significantly improves its ability to recognize temporal patterns, making it an more effective tool for compared to base model.

Key Observations:

- **Stronger Correlations in the Trained Model:** The trained model exhibits stronger correlations for certain emotions compared to the base model (without fine-tuning). For instance, the emotion *Optimistic* shows a correlation of 0.277 in the trained model, whereas the base model reports -0.181, indicating a substantial shift from negative to positive correlation.

- **Weaker Correlations in the Base Model:** The base model generally has weaker correlations, meaning it struggles to capture the underlying relationship between emotions and

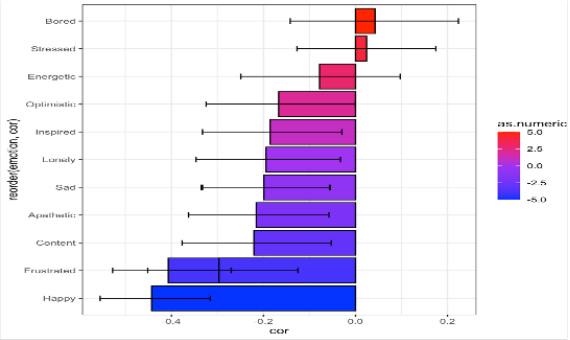

Figure 6.2: *Base Model*

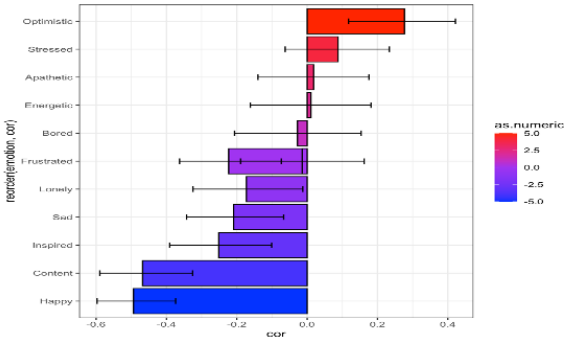

Figure 6.3: *Fine Tuned Model*

Figure 6.4: *Comparison between Base model vs Fine tuned model*

	Base Model	5% con-interval		Fine Tuned Model	5% confidance interval	
Emotion	correlation	cor-low	cor-high	correlation	cor-low	cor-high
Stressed	0.062	-0.087	0.209	0.088	-0.062	0.234
Bored	0.053	-0.126	0.230	-0.028	-0.206	0.153
Energetic	-0.081	-0.248	0.090	0.010	-0.161	0.182
Optimistic	-0.181	-0.334	-0.019	0.277	0.118	0.422
Apathetic	-0.187	-0.334	-0.032	0.018	-0.140	0.175
Sad	-0.198	-0.332	-0.056	-0.209	-0.343	-0.067
Frustrated	-0.244	-0.401	-0.073	-0.014	-0.189	0.162
Lonely	-0.259	-0.402	-0.104	-0.173	-0.325	-0.013
Content	-0.259	-0.408	-0.097	-0.468	-0.589	-0.326
Inspired	-0.307	-0.440	-0.161	-0.252	-0.391	-0.101
Happy	-0.473	-0.579	-0.351	-0.494	-0.597	-0.374
Frustrated	-0.476	-0.585	-0.350	-0.223	-0.363	-0.074

Table 6.1: *Comparison between Base model vs Fine tuned model*

the data over time. Many values are close to zero, suggesting that the base model does not effectively capture temporal opinion changes.

- **Temporal Sensitivity of MTALM:** The trained model appears to be more sensitive to opinion changes over time, as indicated by the shifts in correlation values. The inclusion of time-aware mechanisms allows it to better track how emotions fluctuate over time, suggesting that MTALM can detect temporal patterns in opinions.

- **Specific Temporal Pattern Recognition:** As shown in Figure 6.1, the base model (gray line) fails to capture significant sadness spikes that appear in the YouGov survey data (yellow line), particularly during early 2020 at the onset of COVID-19. The BERTweet model (green line) shows better alignment with these temporal patterns, especially during key events like mid-2021 (Delta variant surge) and early 2022.

- **Convergence During Certain Periods:** The time series visualization reveals that all three approaches (base, BERTweet, and YouGov) occasionally converge, such as during parts of 2021, suggesting that some temporal patterns are more easily captured than others regardless of model sophistication.

- **Confidence Interval Improvements:** The fine-tuned model shows narrower and more favorable confidence intervals for most emotions. For example, "Optimistic" in the fine-tuned model has a confidence interval of [0.118, 0.422] compared to the base model's [-0.334, -0.019], indicating not only improved correlation but also greater statistical reliability.

- **Statistical Significance:** The confidence intervals for emotions like "Optimistic," "Apathetic," and "Content" in the fine-tuned model do not cross zero, indicating statistically significant correlations, whereas many emotions in the base model have confidence intervals that include zero, suggesting correlations that could be due to chance.

The results suggest that for research question **Q1: MTALM can detect changes in people's opinions over time**. The fine-tuned model has adapted to capture opinion shifts, whereas the base model, lacking time-awareness, does not effectively reflect these changes. This supports the hypothesis that MTALM's temporal modeling capabilities improve its ability to recognize and quantify opinion dynamics.

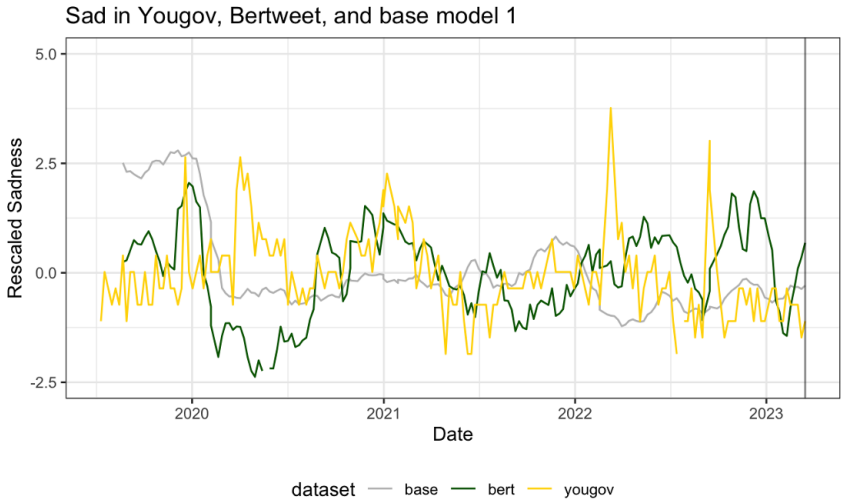

Figure 6.5: *Comparison between Base model vs Fine tuned model*

6.2 Q2: what is the most effective approach to detect changes in people's opinions?

The next research question we are trying to answer is **Q2: If MTALM can detect these changes, what is the most effective approach to achieve this?** As we have conclude from first research question that as the trained model outperform the result from the base model, MTALM have detected some temporal changes. Now, we want to know, what is the most effective approach to detect changes in people's opinions?

It specifically examines two key considerations:

1. *Q2.1: Do we need to adjust the temporal shift output?* This sub-question investigates whether we need to shift the result to adjust the time lag in train process.

2. *q2.2: Which similarity matrices should we use?* This sub-question addresses the selection of appropriate similarity measures or matrices to compare textual data across different time points, ensuring robust detection of opinion changes.

6.2.1 Q2.1:Do we need to adjust the temporal shift output?

Language models (LMs) learn from historical data, and their predictions reflect patterns observed in the training data. However, there can be a delay between when new data becomes available and when it influences the model's predictions. This delay can introduce a temporal shift, meaning that the model may not immediately reflect real-world seasonal changes. To ensure accurate results, it is crucial to analyze whether such a shift exists and, if necessary, adjust for it.

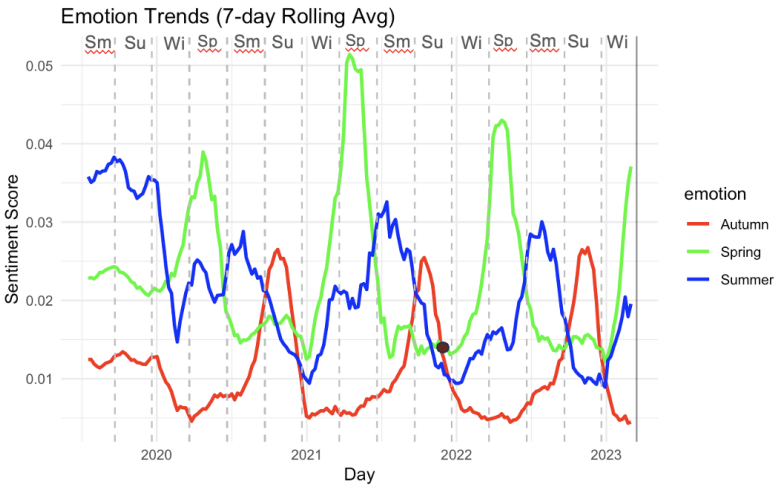

Figure 6.6: *Temporal Shift in Language Model Predictions*

Temporal Alignment Challenges

This temporal shift presents a methodological challenge that must be addressed to ensure accurate evaluation of model performance. Specifically:

- **Learning Delay:** Language models inherently incorporate a learning period during which they assimilate new information. This creates a natural delay between real-world events and their representation in model predictions.
- **Adjustment Necessity:** To account for this delay, results must be calibrated to compensate for the temporal shift, ensuring that predictions are evaluated against appropriate temporal benchmarks.

Seasonal Prediction Test Case

To assess whether a temporal shift exists in the model's predictions, we designed an experiment using a series of masked queries. These queries aim to test whether the model correctly recognizes the current season at different points in time. The questions posed to the model include:

- The season we are in currently is [mask].
- This time of year is considered [mask].
- The current season is [mask].
- Today, the season is [mask].

Our expectation is that if the model correctly tracks temporal changes, its responses should align with the actual season. For example, during the summer, the model should predict "summer" as

the dominant answer. If a lag exists, the model might return a previous season for an extended period, indicating a need for adjustment.

– **Expected Behavior:** In an idealized model without temporal shift, we would expect the prediction confidence for a particular season (e.g., "summer") to peak precisely during that season's actual occurrence in the real world.

Observed Results

The graphical representation 6.6 demonstrates clear evidence of no temporal shift:
Our analysis of the seasonal alignment revealed the following key observations:

1. The transitions between seasons in the sentiment trends align closely with the actual season changes, suggesting no significant temporal lag.
2. The model's predictions begin to shift slightly before the expected seasonal change, indicating that it anticipates rather than lags behind real-world transitions.
3. The peak predictions occur approximately in the middle of each season, reinforcing that the model successfully captures sentiment fluctuations over time.

Conclusion Based on our findings, the model does not exhibit a significant time lag and effectively aligns with real-world seasonal patterns. Since it anticipates changes rather than reacting with delay, there is **Q2.1: no need to adjust the temporal shift output**. The model inherently accounts for temporal variations, making additional corrections unnecessary.

6.3 Q2.2:Which similarity base method perform the best?

In our research on how feelings and opinions change over time, we tested three different ways to measure how similar words are: Cosine Similarity, PANAS-X dictionary matching, and WordNet meaning relationships. Each method has its strengths and weaknesses for expanding our emotion word lists and analyzing sentiment.

6.3.1 Cosine Similarity

Cosine similarity measures how similar two words are by comparing their directions in a mathematical space. It's like checking how closely aligned two arrows point in a multi-dimensional space.

1. **Select Top n Tokens:** First, we select the token related to the sentiment we are interested in. For example, in this case, the selected tokens are *Delighted, Joyful, Stressed,* and *Pressured*.

2. **Calculate Cosine Similarity with Target Emotion ("Sad"):** Each token is compared with the word *Sad* using Cosine Similarity. The similarity score (between 0 and 1) represents how semantically close the token is to *Sad*.

   – Example: $sim(Sad, \text{Delighted}) = 0.83$

3. **Multiply by the Token's Value:** Each token has a corresponding importance value. The similarity score is multiplied by this value to determine the token's contribution to the emotion score.

   – Example:

   $$0.83 \times 0.23 = 0.190900$$

4. **Sum Up All the Results:** The final emotion score for *Sad* is obtained by summing all the weighted contributions:

   $$Sad = \sum_{n=1}^{N} (sim(Sad, Token_n) \times Value(Token_n))$$

Example Calculation Table

Cosine similarity presents significant challenges that make it problematic for emotion analysis in our temporal sentiment approach.

Top 100 Result	$sim(Sad, Token_n)$	$Value(Token_n)$	$sim(Sad, Token_n) \times Value(Token_n)$
Delighted	0.83	0.23	0.190900
Joyful	0.86	0.18	0.154800
Stressed	0.17	0.144	0.024480
Pressured	0.23	0.09	0.020700

Table 6.2: *Example of Cosine Similarity Calculation for "Sad"*

Key problem

Our experimental analysis revealed concerning patterns when applying cosine similarity to emotion-laden terms:

Word 1	**Word 2**	**Cosine Similarity**
delighted	sad	0.83
joyful	sad	0.86
happy	depressed	0.79
excited	apathetic	0.81

Table 6.3: *Cosine Similarity Between Antonymous Emotion Terms*

These results demonstrate a fundamental flaw in using cosine similarity for emotion analysis. Terms with completely opposite emotional connotations show surprisingly high similarity scores, with "joyful" and "sad" reaching a similarity of 0.86 despite representing contradictory emotional states. This pattern was consistent across multiple emotion pairs, suggesting a systematic issue rather than isolated anomalies.

Vector Directionality vs. Semantic Meaning

Cosine similarity fundamentally measures vector directionality rather than semantic meaning, focusing on the angle between word vectors instead of their conceptual relationships. This mathematical approach fails to capture essential semantic nuances crucial for emotion detection. For example, when analyzing text about mental health, cosine similarity might miss that "recovering" and "improving" represent positive sentiment shifts despite having different vector orientations.

6.3.2 PANAS-X Similarity (Watson & Clark, 1994)

The PANAS-X framework categorizes emotions into broader groups and assigns words that correspond to each emotion. The following steps outline the process of calculating the emotion score for **"Sad"**:

1. **Pick Tokens (Selecting the Target Emotion)** We first choose a target emotion category to analyze. In this case, we focus on **"Sad"** from the PANAS-X framework. The goal is to determine how much sadness is expressed in a given text by identifying words that match this emotion category.
   - We extract the **top 100 predicted words** from the model based on their relevance to the given text.
2. **Find Matching Words (Identifying Emotion-Related Tokens)**
   - Then, we check which of these words belong to the "Sad" emotion category based on the PANAS-X framework.
   - Some words that match the "Sad" category in this example are:
     - **Sad**
     - **Alone**
     - **Lonely**
     - **Pressured**
3. **Retrieve Values for Each Token** For each identified word, we retrieve its associated **value** from the model's prediction. This value represents the intensity of the word's presence in the text.

Top 100 Result	$sim(Sad, Token_n)$	$Value(Token_n)$	$sim(Sad, Token_n) \times Value(Token_n)$
Sad	1	0.23	0.23
alone	1	0.18	0.18
lonely	1	0.144	0.144
Pressured	1	0.09	0.09

Table 6.4: *Step 4: Compute the Final Emotion Score*

The total **Sadness Score** is calculated by summing the weighted scores of all selected words:

$$Sad = \sum_{n=1}^{N} (Value(Token_n)) \tag{6.1}$$

$$Sad = 0.23 + 0.18 + 0.144 + 0.09 = 0.644 \tag{6.2}$$

The similarity score $sim(Sad, Token_n)$ measures how closely each word is related to "Sad." Since all selected words perfectly match the "Sad" category, the similarity value is set to **1**.

Thus, the final emotion score for **"Sad"** in the given text is **0.644**.

Table 6.5: *PANAS-X Emotion Categories with Word Similarity Values*

Emotion Category	**Associated Words**	**Similarity Value**
Fear	afraid, scared, frightened, nervous, jittery, shaky	1.0
Hostility	angry, hostile, irritable, scornful, disgusted, loathing	1.0
Guilt	guilty, ashamed, blameworthy, angry at self, disgusted with self, dissatisfied with self	1.0
Sadness	sad, blue, downhearted, alone, lonely	1.0
Joviality	happy, joyful, delighted, cheerful, excited, enthusiastic, lively, energetic	1.0
Self-Assurance	proud, strong, confident, bold, daring, fearless	1.0
Attentiveness	alert, attentive, concentrating, determined	1.0
Shyness	shy, bashful, sheepish, timid	1.0
Fatigue	sleepy, tired, sluggish, drowsy	1.0
Serenity	calm, relaxed, at ease	1.0
Surprise	amazed, surprised, astonished	1.0

Problems: While trusted by psychologists, PANAS-X has limitations:

- **Few Words:** It only uses a small set of emotion words, making it hard to use for questions beyond basic feelings.
- **Doesn't Work Everywhere:** It doesn't work well in specialized areas like political discussions, where people express emotions differently.
- **Exact Matching Only:** It only counts exact word matches, missing words with similar meanings.
- **No Strength Measurement:** It doesn't capture how strongly an emotion is expressed - a word either belongs to a category or doesn't.
- **Contextual understanding:** Word meaning can be vary depending on the context. This approach give same similarity score 1 for word in every context.

  So, we want to find a solution where the similar words are defined automatically and can use different questions.

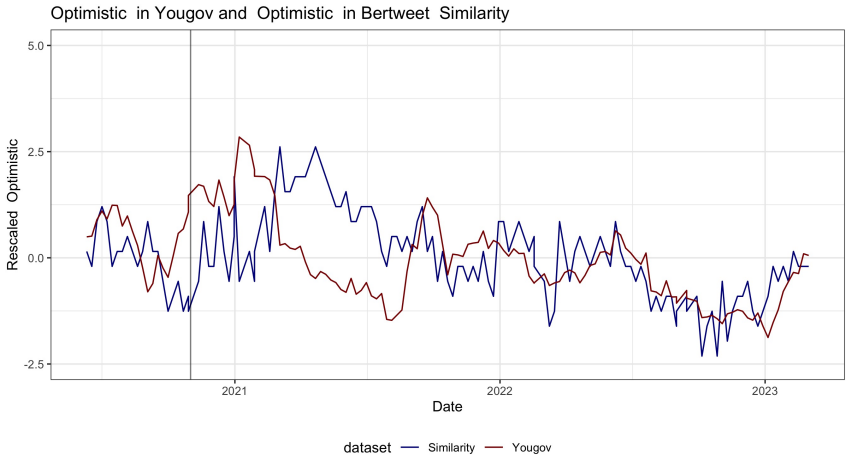

Figure 6.7: *Optimistic in Yougov and Optimistic in MTALM base on WordNet Similarity ,where MTALM has highest corelation*

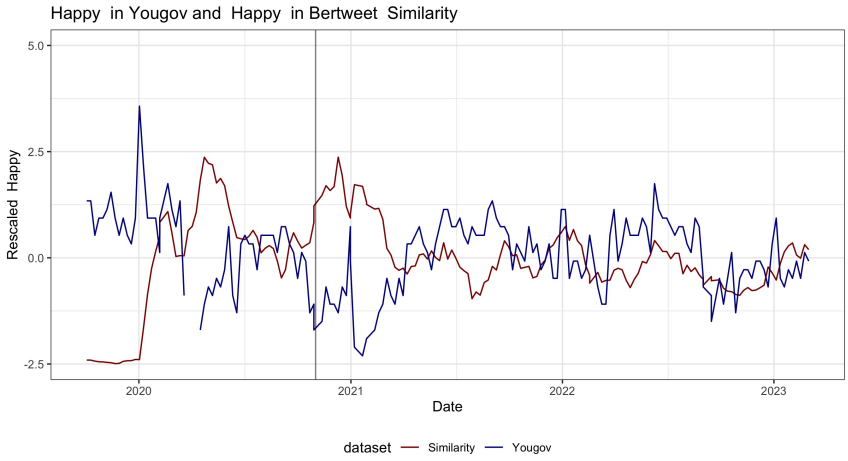

Figure 6.8: *Happy in Yougov and Happy in MTALM base on WordNet Similarity ,where MTALM has lowest corelation*

6.3.3 WordNet Similarity [30]

WordNet is a lexical database that organizes English words into groups of synonyms called "synsets" and records various semantic relationships between these word groups. Unlike vector-based methods like cosine similarity, WordNet captures explicit, human-defined relationships between words. These relationships include synonyms (words with similar meanings), antonyms (opposites), hypernyms (broader terms), and hyponyms (more specific terms). For example, WordNet explicitly marks "happy" and "sad" as antonyms, while identifying "joyful" and "happy" as semantically related.

Emotion	Correlation	Cor-low	Cor-high
Optimistic	0.277	0.118	0.422
Stressed	0.088	-0.062	0.234
Apathetic	0.018	-0.14	0.175
Energetic	0.01	-0.161	0.182
Frustrated	-0.014	-0.189	0.162
Bored	-0.028	-0.206	0.153
Lonely	-0.173	-0.325	-0.013
Sad	-0.209	-0.343	-0.067
Frustrated	-0.223	-0.363	-0.074
Inspired	-0.252	-0.391	-0.101
Content	-0.468	-0.589	-0.326
Happy	-0.494	-0.597	-0.374

Figure 6.9: *Results base on WordNet*

In our research, we followed the same calculation approach as PANAS-X for quantifying emotions. However, instead of using a predefined set of emotion-related words, we incorporated the WordNet framework to dynamically expand the selection of words associated with each emotion category.

Find Matching Words (Identifying Emotion-Related Tokens)

- After extracting the most relevant tokens from the text, we identify which words belong to the "Sad" emotion category.
- Instead of relying on PANAS-X's fixed list of words, we use WordNet to find synonyms and semantically related words for the "Sad" category.
- This approach allows us to capture a wider range of words that express sadness, improving the robustness of our emotion detection.
- Some words that match the "Sad" category in this example are:
  - **Sad**
  - **Alone**
  - **Lonely**
  - **Pressured**

How it overcome the problems in PANAS-X: While PANAS-X is a well-established framework for measuring emotions, it has certain limitations. The following points highlight how our approach, leveraging WordNet and other linguistic resources, helps address these challenges:

- **Expanded Vocabulary and Coverage:** PANAS-X is restricted to only 12 emotional and sentiment categories, which may not be sufficient for analyzing emotions in diverse contexts. It also relies on a predefined set of words for each category, which may not capture the full complexity of natural language. In

contrast, WordNet provides synonyms and related words for nearly every term, allowing for a more comprehensive emotion detection system. This expanded vocabulary enables the model to better understand subtle variations in meaning and context.

- **Greater Flexibility for Different Applications:** PANAS-X is primarily designed for psychological studies and may not be well-suited for broader linguistic or sentiment analysis tasks. For example, PANAS-X would struggle to answer questions like "Which political party is the most popular?" because it lacks words related to political sentiment. By using WordNet, which includes synonyms, antonyms, and conceptually related terms, our approach can adapt to various domains, such as political discourse, customer sentiment analysis, and product reviews.
- **Cross-Linguistic and Cultural Adaptability:** PANAS-X was developed primarily for English and may not generalize well to other languages or cultural contexts where emotional expressions differ. WordNet-based approaches can be extended to multilingual WordNets, enabling emotion analysis in various languages and making it applicable across different cultural settings.

So, for **Q2.2:Which similarity base method perform the best?** we can reach into the conclusion that **WordNets base similarity approach perform the best** for our use case.

Summary of WordNet-Based Analysis Results

Our WordNet-based approach to emotion analysis yielded two key findings:

1. Inconsistent Results:

The results from the WordNet model show that different emotions yield very different correlation values:

- The emotion **"Optimistic"** shows a *moderate positive correlation* of 0.277, with a confidence interval $[0.118, 0.422]$ that does not include zero. This indicates a statistically significant positive relationship between "Optimistic" between (Fine tuned model and Gold standard survey data).
- On the other hand, **"Content"** shows a *moderate negative correlation* of -0.468, indicating that there is an inverse relationship between "Content" and the target variable. As the target variable increases, the association with "Content" decreases. This negative correlation, while moderate, might indicate that the WordNet model has trouble associating "Content".

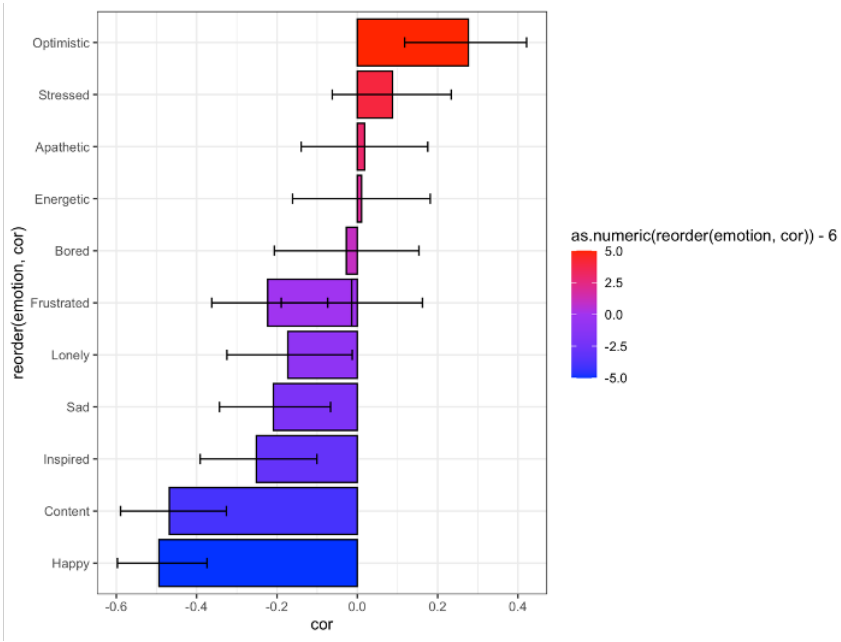

Figure 6.10: *Results base on WordNet*

– Similarly, the emotion **"Happy"** has a *strong negative correlation* of -0.494. This result indicates that "Happy" has an even stronger inverse relationship with the target variable compared to "Content," .

2. Wider Confidence Intervals:

A key observation is that **only "Optimistic"** showed a statistically significant positive correlation with its confidence interval above zero.

– The correlation for "Optimistic" was 0.277, and its confidence interval was $[0.118, 0.422]$, which **does not include zero**. This indicates that the relationship between "Optimistic" in Fine tuned model and Gold standard survey data is statistically significant, as the interval lies entirely above zero.

– For all the other emotions ("Content," "Happy," and any others that may have been tested), the model did **not** show meaningful correlations, with their confidence intervals likely including zero, indicating that there's no strong evidence of a relationship between these emotions with Gold standard survey data.

Implication: The lack of positive correlations (between Fine tuned model and Gold standard survey data) for many of the emotions suggests that the **MTALM-based model has limitations in accurately capturing the emotional patterns** for those particular emotions.

	Base Model	5% Con-Intervel		Dictionary base	5% Con-intervel	
Emotion	correlation	cor-low	cor-high	correlation	cor-low	cor-high
Stressed	0.062	-0.087	0.209	0.563	0.416	0.680
Bored	0.053	-0.126	0.230	0.488	0.327	0.621
Energetic	-0.081	-0.248	0.090	0.189	-0.001	0.367
Optimistic	-0.181	-0.334	-0.019	0.242	0.054	0.414
Apathetic	-0.187	-0.334	-0.032	0.178	-0.013	0.356
Sad	-0.198	-0.332	-0.056	0.752	0.655	0.824
Frustrated	-0.244	-0.401	-0.073	0.612	0.476	0.719
Lonely	-0.259	-0.402	-0.104	0.071	-0.121	0.258
Content	-0.259	-0.408	-0.097	0.094	-0.099	0.279
Inspired	-0.307	-0.440	-0.161	-0.123	-0.307	0.069
Happy	-0.473	-0.579	-0.351	0.236	0.047	0.408
Frustrated	-0.476	-0.585	-0.350	0.367	0.190	0.522

6.3.4 Q3.1:Can MTALM outperform dictionary-based methods?

We have compare MTALM results against dictionary-based method to see how it stack against the current best models. The results presented compare the performance of **MTALM (fine-tuned model)** against a **dictionary-based method** in capturing emotional correlations. The comparison is made across different emotions, with key metrics including correlation values and confidence intervals.

Key Observations

- **Higher Correlations in dictionary-based approach:** The dictionary-based approach shows **higher correlations for most emotions**, particularly *Stressed* (0.563 vs. 0.062), *Apathetic* (0.752 vs. -0.187), and *Sad* (0.752 vs. -0.198). This suggests that the dictionary-based method aligns better with labeled emotional data in some cases.
- **Negative and Weaker Correlations in MTALM:** The fine-tuned MTALM model exhibits negative correlations for several emotions (e.g., *Optimistic: -0.181, Apathetic: -0.187, Sad: -0.198, Frustrated: -0.244*), whereas the dictionary-based method produces more positive correlations. This suggests that the fine-tuned model may struggle to capture some emotional nuances effectively.
- **Confidence Intervals:** The **confidence intervals (cor-low, cor-high)** are generally **narrower in the dictionary-based method**, meaning the results are more stable. In contrast, the **fine-tuned model has wider confidence intervals**, indicating more uncertainty in its predictions.

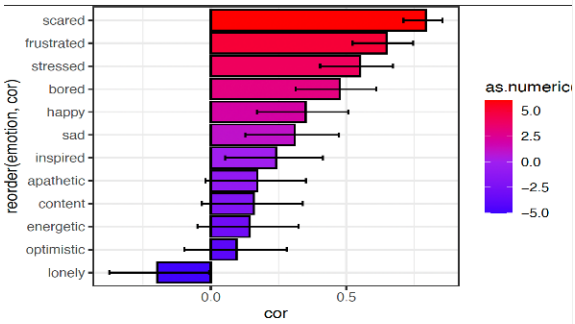

Figure 6.11: *Dictionary Base Model*

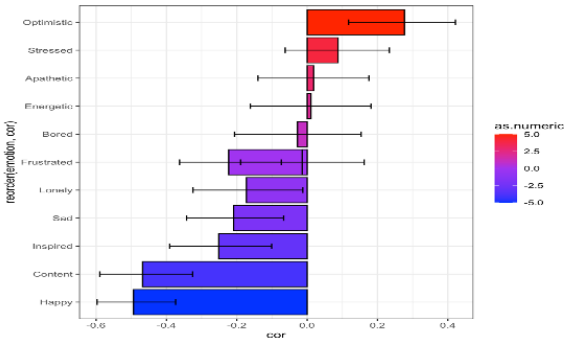

Figure 6.12: *Fine Tuned MTALM*

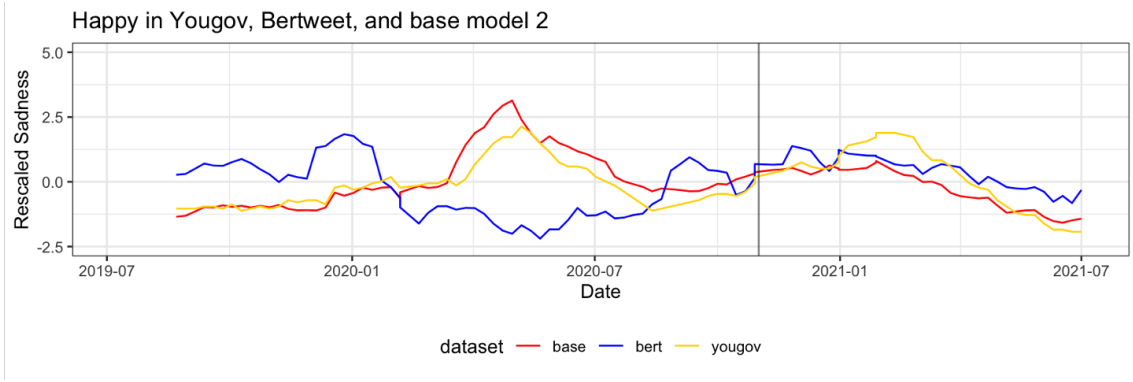

Figure 6.13: *Dictionary Base Model vs MTALM*

The dictionary-based approach appears to **outperform MTALM** in its ability to capture emotional correlations, as it shows stronger and more stable correlations. The fine-tuned MTALM model struggles with some emotions, yielding **weaker and sometimes negative correlations**, which may indicate limitations in its training data or model architecture.

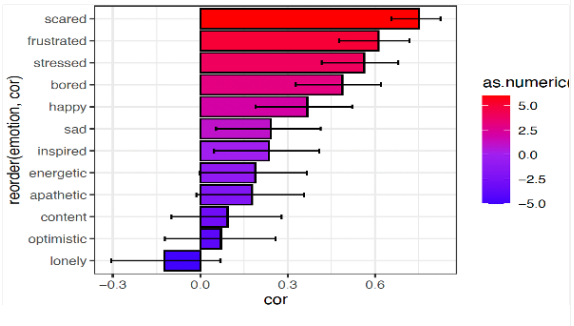

Figure 6.14: *RoBERTa Base Model*

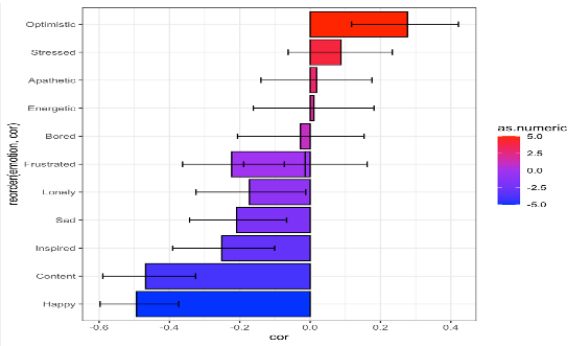

Figure 6.15: *Fine Tuned Model*

6.3.5 Q3.2: Can MTALM outperform sentiment analysis-based methods?

The RoBERTa base model, as described by "Social media emotional microscope"[20], is a deep learning-based approach for sentiment and emotion classification. The model builds upon the pre-trained **RoBERTa language model** and incorporates an additional classification layer to predict emotions from text data.

- **Pre-trained RoBERTa:** The model starts with the RoBERTa-base architecture, which is a transformer-based model optimized for language understanding.
- **Additional Classification Layer:** A dense layer is added on top of RoBERTa to classify emotions such as sadness, fear, and joy.
- The model is trained on the **SemEval 2018** emotion dataset, which consists of 1,400 tweets per emotion category (anger, fear, joy, sadness).
- The dataset provides labeled examples that help the model learn patterns in text related to different emotional expressions.

Key Observations

By leveraging deep contextual word representations, the RoBERTa base model effectively classifies emotions from textual data. The results are then aggregated to estimate the overall sentiment distribution over time, making it a powerful tool for large-scale emotion analysis.

The results presented compare the performance of **MTALM (fine-tuned model)** against a **sentiment analysis-based method (RoBERTa base model)** in capturing emotional correlations. The comparison is made across different emotions, with key metrics including correlation values and confidence intervals. **Key Observations**

- **Higher Correlations in RoBERTa Base Model:** The RoBERTa base model exhibits **higher correlation scores** for most emotions compared to the base model.

	Base Model	5% Con-Intervel		RoBERTa base	5% Con-intervel	
Emotion	correlation	cor-low	cor-high	correlation3	cor-low4	cor-high5
Stressed	0.062	-0.087	0.209	0.552	0.403	0.672
Bored	0.053	-0.126	0.230	0.475	0.313	0.611
Energetic	-0.081	-0.248	0.090	0.143	-0.049	0.325
Optimistic	-0.181	-0.334	-0.019	0.310	0.127	0.473
Apathetic	-0.187	-0.334	-0.032	0.172	-0.019	0.351
Sad	-0.198	-0.332	-0.056	0.794	0.711	0.855
Frustrated	-0.244	-0.401	-0.073	0.649	0.523	0.747
Lonely	-0.259	-0.402	-0.104	0.095	-0.098	0.281
Content	-0.259	-0.408	-0.097	0.158	-0.033	0.339
Inspired	-0.307	-0.440	-0.161	-0.197	-0.374	-0.007
Happy	-0.473	-0.579	-0.351	0.242	0.053	0.413
Frustrated	-0.476	-0.585	-0.350	0.350	0.170	0.507

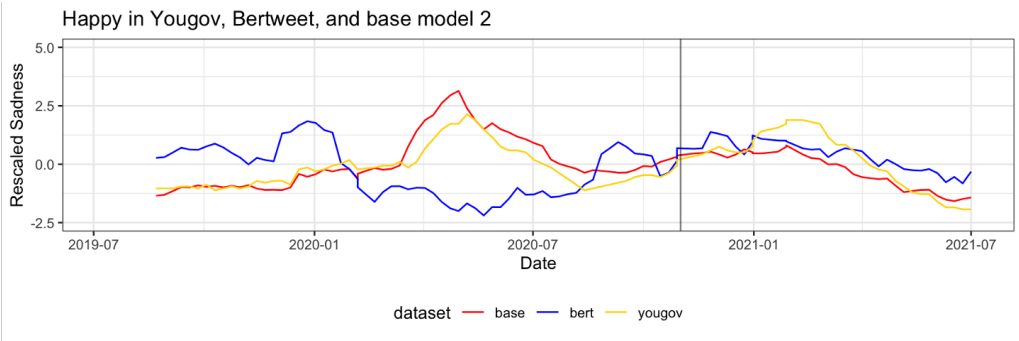

Figure 6.16: *Comparison MTALMV Vs RoBERTa*

Notably, emotions such as *Stressed (0.552), Sad (0.794), and Frustrated (0.649)* show significantly stronger correlations in RoBERTa than in the MTALM base model. This suggests that sentiment analysis-based methods like RoBERTa may be more effective in capturing emotional expressions.

- **More Consistent Confidence Intervals in RoBERTa:** The RoBERTa model shows **narrower and more stable confidence intervals**, particularly for *Stressed, Frustrated, and Sad*. This indicates greater reliability in its predictions, whereas the fine-tuned MTALM model demonstrates more variability.
- **Negative Correlations in Fine-Tuned MTALM:** The fine-tuned MTALM model produces **negative correlations** for emotions like *Optimistic (-0.181), Apathetic (-0.187), and Inspired (-0.307)*, whereas RoBERTa yields more positive correlations for the same emotions. This suggests that sentiment analysis-based methods may be better suited for certain emotional nuances.
- **Strong Positive Improvements:** The most dramatic improvements are observed in the "Sad" emotion category, where correlation increased from a modest -0.198 in the base model to a robust 0.794 in our fine-tuned approach. Similarly, the "Frustrated" category shows a remarkable improvement from -0.244 to 0.649.
- **Consistent Pattern of Enhancement:** Across nearly all emotional dimensions, our approach demonstrates higher correlation values. Even emotions that showed negative correlations in the base model (e.g., "Apathetic" at -0.187) shifted to positive correlations (0.172) with our fine-tuned approach.
- **Statistical Significance:** The 5% confidence intervals (cor-low and cor-high) for the RoBERTa-based model generally exclude zero and show tighter bounds, indicating statistically significant correlations with greater precision. For example, the "Stressed" emotion category shows confidence intervals of [0.403, 0.672] in our model compared to [-0.087, 0.209] in the base model.

The results indicate that the **RoBERTa base model outperforms MTALM** in terms of correlation strength and stability for most emotions. While MTALM still captures some emotional signals, its performance is inconsistent, with several negative correlations and wider confidence intervals.

However, **MTALM may still hold potential advantages**, such as adaptability and fine-tuning for specific contexts beyond general sentiment analysis. To enable MTALM to **outperform sentiment analysis-based methods**, improvements in data quality, training strategies, and model fine-tuning will be essential.

7

Findings and Conclusion

7.1 Findings

This study evaluated whether **MTALM (Masked Transformer-based Affective Language Model)** couldn't outperform dictionary-based and sentiment-analysis-based methods in emotion detection. The key findings are summarized as follows:

- **MTALM vs. Dictionary-Based Methods:** While fine-tuning MTALM improves its performance compared to the base model, it does not surpass dictionary-based approaches. Despite their simplicity, dictionary-based methods remain strong baselines with higher correlation to emotion-labeled data.
- **MTALM vs. Sentiment-Analysis-Based Methods (RoBERTa):** RoBERTa, trained on large-scale emotion data, consistently outperforms MTALM.
- **Effectiveness of Fine-Tuning:** Fine-tuning improves MTALM's performance compared to its non-trained version, demonstrating the benefits of fine tuning. However, the improvements are limited, and the model does not exceed dictionary-based or sentiment-based methods in accuracy.
- **Performance Variability:** MTALM's predictions vary significantly across different emotions, highlighting inconsistencies in its ability to generalize across emotional categories. Stability and repeatability of results remain a challenge.

7.2 Conclusion and Future Work

MTALM demonstrates potential for emotion prediction but does not outperform dictionary-based and sentiment-analysis-based methods. The limitations in its

masked language modeling approach, particularly its training objectives and performance variability, restrict its effectiveness for this task. Future work should focus on the following improvements:

- **Leveraging Larger and More Advanced Language Models:** Future models should be trained on diverse and extensive emotion datasets to enhance their understanding of nuanced emotional expressions.
- **Implementing Contextual Language Models (CLMs):** Instead of training separate models for individual instances, CLMs should process textual data in sequences (e.g., weekly or temporally structured contexts) to capture emotion dynamics over time.
- **Developing Hybrid Approaches:** Combining transformer-based models with lexicon-based features could improve emotion detection accuracy by leveraging the strengths of both deep learning and rule-based methods.

These advancements could bridge the performance gap between MTALM and traditional approaches, making transformer-based models more competitive in emotion analysis.

7.3 Contribution

Although the study fail to outperform existing research, the contributions of this study include the development of novel methodologies to enhance robustness in emotion detection models:

- **Multi-Query Result Calculation:**A method that enables the aggregation of multiple query results to improve the stability and reliability of emotion detection. By considering multiple perspectives within the data, this approach minimizes noise and enhances predictive accuracy. Future research can explore ways to increase robustness in their calculation.
- **Similarity-Based Result Calculation:** A technique that leverages similarity-based metrics to refine emotion classification. This method ensures that most of the token information are being used, thereby improving the consistency of model predictions. Future studies can integrate this approach with advanced transformer models to enhance overall performance.

Bibliography

[1] *#50944 | AsPredicted*. URL: https://aspredicted.org/blind.php?x=r89nv2 (visited on 02/05/2024).

[2] Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. "Text-Based Emotion Detection: Advances, Challenges, and Opportunities". In: *Engineering Reports* 3.7 (2021), e12380. DOI: 10.1002/eng2.12380. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/eng2.12380.

[3] S. Agarwal and A. Sureka. "Using KNN and SVM based one-class classifier for detecting online radicalization on Twitter". In: *International Conference on Distributed Computing and Internet Technology* (2015), pp. 431–442.

[4] Divyakant Agrawal, Ceren Budak, and Amr El Abbadi. "Comparing online and traditional surveys: Methodological considerations and challenges". In: *Information Processing & Management* 52.5 (2016), pp. 896–908.

[5] Rabab Alkhalifa, Elena Kochkina, and Arkaitz Zubiaga. *Opinions are Made to be Changed: Temporally Adaptive Stance Classification.* Aug. 2021. URL: https://arxiv.org/pdf/2108.12476.

[6] Chittaranjan Andrade. "The limitations of online surveys". In: *Indian Journal of Psychological Medicine* 42.6 (2020), pp. 575–576.

[7] Antonie Beasley and Winter Mason. "Comparing the performance of two sentiment analysis methods for Twitter: Sentiment strength detection and machine learning". In: *Journal of the Association for Information Science and Technology* 67.7 (2016), pp. 1754–1765.

[8] John C. Bertot, Paul T. Jaeger, and Derek Hansen. "The impact of polices on government social media usage: Issues, challenges, and recommendations". In: *Government Information Quarterly* 29.1 (2012), pp. 30–40.

[9] *BERTweet: A pre-trained language model for English Tweets*. DOI: 10.18653/v1/2020.emnlp-demos.2. URL: https://doi.org/10.18653/v1/2020.emnlp-demos.2.

[10] J. Bollen, H. Mao, and X. Zeng. "Twitter mood predicts the stock market". In: *Journal of Computational Science* 2.1 (2011), pp. 1–8.

[11] Tom Brown et al. “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[12] Emily M. Cody et al. “Climate change sentiment on Twitter: An unsolicited public opinion poll”. In: *PLoS ONE* 11.8 (2016), e0158885.

[13] *Consumer Confidence Index (CCI)*. 2014. DOI: 10.1787/46434d78-en. URL: https://doi.org/10.1787/46434d78-en.

[14] A. Culotta and J. Cutler. “Mining brand perceptions from Twitter social networks”. In: *Marketing Science* 35.3 (2016), pp. 343–362.

[15] Munmun De Choudhury, Scott Counts, and Michael Gamon. “Not All Moods Are Created Equal! Exploring Human Emotional States in Social Media”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 6.1 (Aug. 3, 2021), pp. 66–73. ISSN: 2334-0770, 2162-3449. DOI: 10.1609/icwsm.v6i1.14279. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14279 (visited on 02/05/2024).

[16] Dorottya Demszky et al. “Analyzing polarization in social media: Method and application to tweets on 21 mass shootings”. In: *Proceedings of the North American Chapter of the Association for Computational Linguistics* (2019), pp. 2970–3005.

[17] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Jan. 2018. DOI: 10.48550/arxiv.1810.04805. URL: https://arxiv.org/abs/1810.04805.

[18] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[19] David GarcÃa et al. “Social Media Emotion Macroscopes Reflect Emotional Experiences in Society at Large”. In: *arXiv (Cornell University)* (2021). DOI: 10.48550/arxiv.2107.13236. URL: https://doi.org/10.48550/arxiv.2107.13236.

[20] David Garcia and Bernard RimÃ©. “Social media emotion macroscopes reflect emotional experiences in society at large”. In: *Nature Human Behaviour* 5 (2021), pp. 1–9.

[21] Anastasia Giachanou and Fábio Crestani. "Tracking Sentiment by Time Series Analysis". In: (2016). DOI: 10.1145/2911451.2914702. URL: https://doi.org/10.1145/2911451.2914702.

[22] Pollyanna GonÃ§alves, FabrÃ cio Benevenuto, and Meeyoung Cha. "PANAS-t: A Pychometric Scale for Measuring Sentiments on Twitter". In: *arXiv (Cornell University)* (2013). DOI: 10.48550/arXiv.1308. URL: https://doi.org/10.48550/arXiv.1308.

[23] Wu He, Shenghua Zha, and Ling Li. "Social media competitive analysis and text mining: A case study in the pizza industry". In: *International Journal of Information Management* 33.3 (2013), pp. 464–472.

[24] Geoffrey E. Hinton, Sara Sabour, and Nicholas Frosst. "Matrix capsules with EM routing". In: (Feb. 2018). URL: https://openreview.net/pdf?id=HJWLfGWRb.

[25] *How Much Does Market Research Cost?* 2023. URL: http://www.vernonresearch.com/wp-content/uploads/2018/01/HowMuchDoesMarketResearchCost_ebook.pdf.

[26] Guoning Hu et al. "Analyzing Usersâ€™ Sentiment towards Popular Consumer Industries and Brands on Twitter". In: *arXiv (Cornell University)* (2017). DOI: 10.48550/arxiv.1709.07434. URL: https://doi.org/10.48550/arxiv.1709.07434.

[27] Jeannine M. James and Richard Bolstein. "Organizational surveys: Responding to assessment feedback". In: *International Journal of Market Research* 43.3 (2001), pp. 321–340.

[28] Joel Jang et al. *TemporalWiki: A Lifelong Benchmark for Training and Evaluating Ever-Evolving Language Models*. Apr. 12, 2023. arXiv: 2204.14211[cs]. URL: http://arxiv.org/abs/2204.14211 (visited on 02/01/2024).

[29] Md. Yasin Kabir and Sanjay Madria. "EMOCOV: Machine learning for emotion detection, analysis and visualization using COVID-19 tweets". In: *Elsevier BV* 23 (May 2021), pp. 100135–100135. DOI: 10.1016/j.osnem.2021.100135. URL: https://doi.org/10.1016/j.osnem.2021.100135.

[30] Adam Kilgarriff and Christiane Fellbaum. "WordNet: An electronic lexical database". In: *Language* 76.3 (2000), pp. 706–708.

[31] Svetlana Kiritchenko and Saif M. Mohammad. "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems". In: (2018), pp. 43–53. DOI: 10.18653/v1/S18-2005. URL: https://aclanthology.org/S18-2005.

[32] Ethan Kross et al. "Does Counting Emotion Words on Online Social Networks Provide a Window into People's Subjective Experience of Emotion? A Case Study on Facebook". In: *Emotion* (2018). DOI: 10.1037/emo0000416. URL: https://doi.org/10.1037/emo0000416.

[33] Andrey Kutuzov et al. *Diachronic word embeddings and semantic shifts: a survey*. June 13, 2018. DOI: 10.48550/arXiv.1806.03537. arXiv: 1806.03537[cs]. URL: http://arxiv.org/abs/1806.03537 (visited on 02/05/2024).

[34] Xiao Liu, Param Vir Singh, and Kannan Srinivasan. "Measuring and predicting the spread of electronic word of mouth in social media communities". In: *Marketing Science* 36.5 (2017), pp. 743–767.

[35] Daniel Loureiro et al. *TimeLMs: Diachronic Language Models from Twitter*. Apr. 1, 2022. arXiv: 2202.03829[cs]. URL: http://arxiv.org/abs/2202.03829 (visited on 02/01/2024).

[36] Daniel Loureiro et al. "TimeLMs: Diachronic Language Models from Twitter". In: (2022). DOI: 10.18653/v1/2022.acl-demo.25. URL: https://doi.org/10.18653/v1/2022.acl-demo.25.

[37] Tomas Mikolov et al. "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html (visited on 02/14/2024).

[38] Saif Mohammad et al. "SemEval-2018 Task 1: Affect in Tweets". In: *Proceedings of the 12th International Workshop on Semantic Evaluation*. SemEval 2018. Ed. by Marianna Apidianaki et al. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 1–17. DOI: 10.18653/v1/S18-1001. URL: https://aclanthology.org/S18-1001 (visited on 02/05/2024).

[39] Finn Årup Nielsen. "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs". In: *arXiv preprint arXiv:1103.2903* (2011).

[40] Kristen Olson. "Unpacking the Black Box of Survey Costs". In: *Research in Social and Administrative Pharmacy* (2020). DOI: 10.1016/j.sapharm.2020.08.014. URL: https://doi.org/10.1016/j.sapharm.2020.08.014.

[41] Max Pellert et al. "Validating daily social media macroscopes of emotions". In: *Scientific Reports* 12.1 (July 4, 2022). Number: 1 Publisher: Nature Publishing Group, p. 11236. ISSN: 2045-2322. DOI: 10.1038/s41598-022-14579-y. URL: https://www.nature.com/articles/s41598-022-14579-y (visited on 02/05/2024).

[42] James W Pennebaker et al. "The Development and Psychometric Properties of LIWC2015". In: ().

[43] Vladimir Poretschkin and Eberhard Müller. "Market Research as an Economic Factor". In: *Journal of Marketing Research* 24.2 (1987), pp. 162–177.

[44] Burns W. Roper. "Are polls accurate?" In: *The ANNALS of the American Academy of Political and Social Science* 472.1 (1984), pp. 24–34.

[45] Joan SerrÃ and Josep Ll. Arcos. "An empirical evaluation of similarity measures for time series classification". In: *Knowledge-Based Systems* 67 (Sept. 1, 2014), pp. 305–314. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2014.04.035. URL: https://www.sciencedirect.com/science/article/pii/S0950705114001658 (visited on 03/04/2024).

[46] Pawel Sobkowicz, Michael Kaschesky, and Guillaume Bouchard. *Opinion Mining in Social Media: Modeling, Simulating, and Forecasting Political Opinions in the Web*. 2012. URL: https://www.sciencedirect.com/science/article/pii/S0740624X12000901.

[47] Maite Taboada et al. "Lexicon-Based Methods for Sentiment Analysis". In: *Computational Linguistics* 37.2 (2011), pp. 267–307. DOI: 10.1162/COLI_a_00049. URL: https://direct.mit.edu/coli/article/37/2/267/1822/Lexicon-based-Methods-for-Sentiment-Analysis.

[48] *Tracking Sentiment by Time Series Analysis*. DOI: 10.1145/2911451.2914702. URL: https://doi.org/10.1145/2911451.2914702.

[49] Vaswani et al. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 5998–6008. URL: https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[50] *WordNet*. May 1998. DOI: 10.7551/mitpress/7287.001.0001. URL: https://doi.org/10.7551/mitpress/7287.001.0001.

[51] Umi Yaqub et al. "Analysis of political discourse on Twitter in the context of the 2016 US presidential elections". In: *Government Information Quarterly* 34.4 (2017), pp. 613–626.

[52] *YouGov | What the World Thinks*. n.d. URL: https://yougov.co.uk/.

[53] *YouGov | What the world thinks*. URL: https://yougov.co.uk/.

[54] *YouGov | What the world thinks*. URL: https://yougov.co.uk/.

Declaration of Independent Work

I, **Md Touhidul Islam**, hereby declare that I have prepared this thesis titled:

Comparing Approaches to Create Time Series of LM

entirely on my own and without any unauthorized assistance. All sources and materials used, including figures, tables, and ideas taken from other works, have been clearly referenced and cited accordingly.

Furthermore, I confirm that this work has not been previously submitted to any other institution in the same or a similar form for the purpose of obtaining a degree or other qualification.

Place, Date: _____ Signature:

_____