EMOVIS: A Visual Approach to Tracking Emotional Sentiment Dynamics in Social Network Commentaries

Ismail Hossain[1], **Sai Puppala**[1], **Md Jahangir Alam**[1], **Sajedul Talukder**[1], **Zahidur Talukder**[2]
[1]School of Computing, Southern Illinois University Carbondale, IL, USA
[2]Department of Computer Science, University of Texas at Arlington, TX, USA
ismail.hossain@siu.edu, saimaniteja.puppala@siu.edu, mdjahangir.alam@siu.edu, sajedul.talukder@siu.edu, zahidurrahim.talukder@mavs.uta.edu

Abstract

The expansion of social media has unlocked a real-time barometer of public opinion. This paper introduces a novel framework to visualize sentiment shifts in social network comment sections, a reflection of the broader public discourse over time. Leveraging a pre-trained uncased $RoBERTa_{large}$ model, we predict emotional scores from user comments, mapping these to key sentiment trends such as Approval, Toxicity, Obscenity, Threat, Hate, Offensive, and Neutral. Our methodology employs machine learning techniques to train a dataset that connects emotional scores with these trends, generating trend probability scores. We utilize a bottom-up recursive algorithm to aggregate emotional scores within comment threads, enabling the prediction of trend scores using three distinct aggregation methods. The results demonstrate that our emotional prediction model achieves an AUC of 0.92, and XGBoost stands out with an F1 score exceeding 0.40. Our research elucidates the temporal evolution of online public sentiment, enhancing the understanding of digital social dynamics and offering insights for strategic online interaction, intervention, and content moderation.

Introduction

Social media platforms have become integral to modern society, serving as both conduits for communication and influential drivers of public sentiment. Penycook et al. (Pennycook et al. 2020) have explored the role of these platforms in molding public perspectives. With the continuous expansion and transformation of these digital spaces, the imperative to decode their impact intensifies. Platforms like Twitter, Reddit, and Facebook empower individuals to broadcast their thoughts, engage in dialogues on pertinent issues, and propel those discussions into wider circles. The resulting dialogues can weave through networks, connecting users across various degrees of separation. Sentiment analysis emerges as a key analytical tool, especially potent in parsing the nuanced exchanges within comment threads. These threads are fertile grounds for debate, rich in data that can be mined to discern public mood swings, highlight nascent trends, and gauge collective reactions to unfolding events. Bollen et al. (Bollen, Mao, and Zeng 2011) have also indicated the potential of emotion tracking in forecasting economic shifts. For the corporate sphere, such insights are invaluable for sculpting marketing campaigns, navigating crises, and steering product innovation (Petrović, Osborne, and Lavrenko 2010). Moreover, for decision-makers and researchers, the shifts in sentiment unearthed from these discussions can illuminate public stances on policy decisions, societal developments, or emergent health concerns.

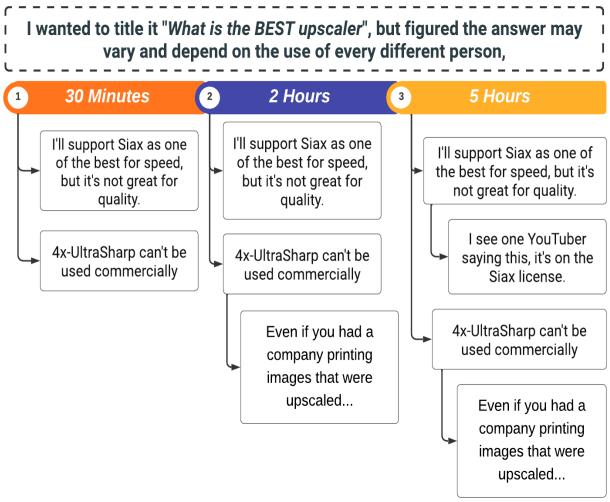

Figure 1: A sample of a hierarchy thread representing the possible increase in comments over the period of time for a user post. Each rectangular box represents a user's comment on the post or news.

This paper introduces a detailed methodology to monitor and depict sentiment changes within comment sections of social networks. We explore the application of this method across various fields to extract practical insights and deepen our grasp of how public opinion shifts in the digital age. The focus is on dissecting comment threads to predict sentiment trends based on individual comments' emotional content. We utilize a pre-trained uncased $RoBERTa_{large}$ model to evaluate comments and their responses, assigning emotional scores that correspond to trends like Approval, Toxicity, and Neutral, among others. We then map the $RoBERTa_{large}$

output to these seven identified trends and train a dataset to predict the likelihood of each trend, using a variety of machine learning models. The experiment processes comment threads through a bottom-up recursive method, aggregating emotional data to determine the overall sentiment. The resulting trend scores, indicative of our research goal, show that the $RoBERTa_{large}$ model achieves an AUC of 0.92, with XGB and Decision Tree models surpassing others with an F1 score over 0.40.

Figure 1 displays the growth of a comment section over time, demonstrating our data-driven approach to understanding online sentiment. This work leverages sophisticated sentiment analysis tools and visual techniques to illuminate public opinion trends, emphasizing the value of sentiment tracking across diverse sectors such as marketing, crisis response, political science, public health, and social research.

In summary, we introduce the following contributions:

- **Prediction of Emotion and Trend Scores:** We have fine-tuned two BERT-based models specifically for predicting both emotion and trend scores derived from social media comments.
- **Aggregation of Emotion and Trend Scores:** We employ two distinct aggregation methods to consolidate emotion scores, both at the sub-reply tree level and for the entire comment section.
- **Empirical Validation:** We present and scrutinize various trend prediction and aggregation approaches. Through empirical validation, we determine the most effective trend prediction technique and the most resilient aggregation method.
- **Dynamic Visualization of Emotions and Trends:** The innovative design of our user interface provides a dynamic, real-time visualization of changing emotions and trends within comment sections. This feature allows users to effectively track and understand the evolving sentiments in an interactive and engaging manner.

Research Objectives

In this article, we investigate the following research questions on sentiment analysis and trend prediction in social media comment sections:

(RQ1): Can we design a tool to effectively track and analyze the evolution of sentiment trends in social media comment sections across different time frames, focusing on both short-term fluctuations and long-term shifts in public opinion?

(RQ2): What challenges and opportunities arise when implementing real-time sentiment tracking and visualization in social media comment sections?

(RQ3): How effective is the BERT model at accurately predicting emotional scores from complex social media comments that are often ambiguous or multifaceted in nature?

(RQ4): Which is more effective in predicting trends in social media comments: a direct trend prediction approach using T-BERT or a two-step approach involving emotion-based E-BERT and traditional models?

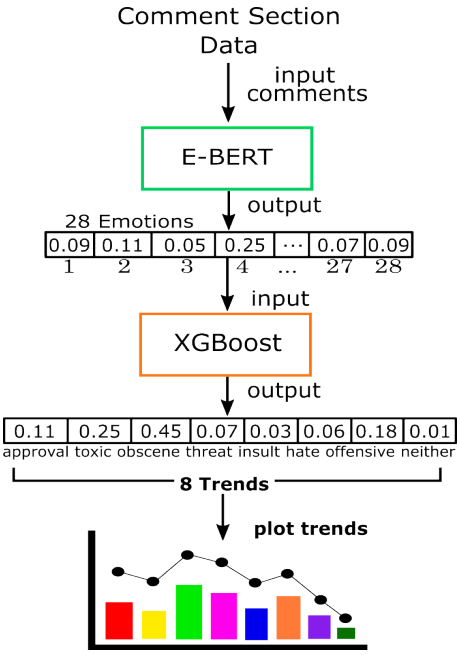

Figure 2: System flow architecture starting at data acquisitions from different sources to the machine learning model tuning and different algorithms.

(RQ5): How does the choice of aggregation method for consolidating emotion scores influence the subsequent prediction of sentiment trends, specifically in terms of precision and recall, in social media comment sections?

Our end goal with this research is to transform complex sentiment data into intuitive visual representations, making it easier for users to grasp and interpret the evolving nature of public opinion in the digital landscape and have a deeper understanding of online social interactions and public opinion dynamics.

Related Works

In the ever-evolving landscape of online interactions, researchers have embarked on a journey to unravel the intricacies of human behavior within digital realms. In this quest, Lee (Lee and Ryu 2019) takes center stage, delving into the exploration of human characteristics within the dynamic landscape of online news comments. As the digital narrative unfolds, Won (Oh and Park 2021) seamlessly extends the conversation, shifting the focus to the pervasive issue of online spam. With a keen eye for reliability, Won emphasizes innovative cleaning approaches to enhance the quality of digital discourse.

Amidst this exploration of the digital realm, Sharma (Sharma et al. 2017) emerges as a trailblazer, contributing significantly to the preliminary steps of hate speech identification in recent publications. In the intricate tapestry of online communication, Sharma's work marks a crucial step forward, shedding light on the challenges and advancements in identifying and combating hate speech.

As the story unfolds, these researchers stand at the fore-

front of understanding, cleaning, and securing the digital spaces we inhabit. Their collective efforts pave the way for a safer, more insightful online experience, challenging the complexities of human expression in the vast landscape of the internet.

On the other side, Rottger et al. (Röttger et al. 2020) introduced HATECHECK, a suite of functional tests designed for evaluating hate speech detection models. While continuous advancements are made in developing better models for hate speech detection, there has been limited research on the bias and interpretability aspects of hate speech. In a recent paper, Mathew et al. (Mathew et al. 2021) introduced HateXplain, the first benchmark hate speech dataset covering multiple dimensions of the issue.

Previous research efforts primarily focused on identifying and characterizing various forms of negative behaviors, such as hate speech (Mondal, Silva, and Benevenuto 2017), harassment (Founta et al. 2018), cyberbullying (Yao, Chelmis, and Zois 2019), and general toxic behavior (Wulczyn, Thain, and Dixon 2017a). These studies typically analyzed content in isolation and retrospectively (Fortuna and Nunes 2018). While valuable for monitoring and reducing exposure to toxic content, they have limited potential in predicting and preventing toxic behaviors beforehand (Jurgens, Hemphill, and Chandrasekharan 2019). To forecast toxicity, it is crucial to consider the social and conversational context in which such behaviors occur.

Other previous studies on conversational analysis explored various outcomes, including re-entry of conversations (Shugars and Beauchamp 2019), their productivity (Niculae and Danescu-Niculescu-Mizil 2016), controversy (Hessel and Lee 2019), and the likelihood of leading to disagreement (Wang and Cardie 2016). Recently, there has been a focus on predicting toxicity by considering pragmatic cues (Zhang et al. 2018) and learned representations (Chang and Danescu-Niculescu-Mizil 2019) of the language used in the initial exchanges of a conversation.

In a separate study, Saveski et al. (Saveski, Roy, and Roy 2021) utilized two prediction techniques on 1.18M Twitter conversations. The first technique predicts the toxicity of the first ten replies, while the second predicts whether the next reply from a specific user would be toxic or not. Additionally, Bollen et al. (Bollen, Mao, and Pepe 2011) conducted sentiment analysis on all public tweets broadcasted by Twitter users between August 1 and December 20, 2008. They extracted six dimensions of mood using an extended version of the Profile of Mood States (POMS) – tension, depression, anger, vigor, fatigue, and confusion.

Addressing the training dataset needs of researchers, Mohammad et al. (Mohammad and Kiritchenko 2018) presented a pioneering dataset that encompasses over 11,000 tweets. This dataset is meticulously designed to offer researchers a holistic comprehension of emotions. With a keen focus on impactful content, it features annotations encompassing both coarse classes (e.g., anger or no anger) and intricate real-valued scores, providing insights into the intensity of emotions, including anger, sadness, and valence.

Methodology

The online comments section, a ubiquitous feature on most social media posts, online blogs, news websites, and other digital platforms, provides a space where audiences can interact and express their views on the posted content. In addressing our first two research questions (**RQ1**, **RQ2**), we aim to design a tool that effectively tracks and analyzes the evolution of sentiment trends in these online comment sections. This tool is specifically tailored to capture both short-term fluctuations and long-term shifts in public opinion across various time frames. To effectively capture the sentiment evolution, we introduce two primary visualization metrics: the shift in emotions over time and the change in trends over time.

We have identified a comprehensive set of 27 different emotions (Demszky et al. 2020) with a neutral option, which is suitable for multi-class, multi-label emotion classification in online comments. Additionally, we have delineated eight distinct trend categories—approval, toxicity, obscenity, threat, insult, hate, offensiveness, and neutrality—that can be extracted from online comments (Davidson et al. 2017). To address **RQ3**, we utilize a pre-trained, $BERT_{base}$, $BERT_{large}$, and $RoBERTa_{large}$ model, fine-tuned with the GoEmotion dataset (Demszky et al. 2020) for the prediction of emotional scores. To classify trends, we implement two distinct approaches: 1) direct prediction of the trend from the comments, and 2) a two-step method involving first predicting emotions from the comments, and then deducing the trend based on these emotions. Figure 2 shows the system flow architecture starting at data acquisitions from different sources to the machine learning model tuning and different algorithms. The subsequent sections will detail our methodology and its application in depth.

Emotion Prediction

For the prediction of emotional scores, we utilize pre-trained $BERT_{base}$, $BERT_{large}$, and $RoBERTa_{large}$ models, which are fine-tuned using the GoEmotion dataset (Demszky et al. 2020), referred to as $D_{emotion}$. This approach is tailored for multi-class, multi-label emotion classification encompassing 27 distinct emotions and a neutral option. We engage in transfer learning by using the pre-trained BERT models, fitting our dataset, and finally updating the final classification layer to accommodate the number of classes present in $D_{emotion}$ (num_class=28). Detailed specifications regarding the fine-tuning process and parameter settings are provided in the sec:experiment section. After fine-tuning we call this model E-BERT (E for Emotion) as shown in Figure 5. E-BERT is subsequently employed to generate the probability of each emotion for every comment in an online comments section.

Trend Prediction

To classify trends in social media comments, we adopt two distinct approaches. The first involves directly predicting the trend from the comments themselves. The second approach is a two-step method, where we first predict emotions from the comments using E-BERT and then deduce the

trend based on these identified emotions. In the following sections, we will elaborate on the methodologies employed in both these approaches.

Direct Approach. For the direct prediction of trend probabilities from comments, we deploy pre-trained models such as $BERT_{base}$, $BERT_{large}$, and $RoBERTa_{large}$. These models are fine-tuned using a combined dataset of Twitter (Davidson et al. 2017) and Wikipedia (Davidson et al. 2017) comments, denoted as D_{trend}. This procedure is designed for multi-class, multi-label trend classification, covering eight distinct trend categories. We modify the final classification layer of the models to align with the eight classes (num_class=8) in D_{trend}. The finer details of this fine-tuning process, including specific parameter settings are discussed in the sec:experiment section. Upon completion of the fine-tuning, the resultant model is referred to as T-BERT (T for Trend), as depicted in Figure 5. T-BERT is then applied to assess the trend probability of each comment within online comment sections.

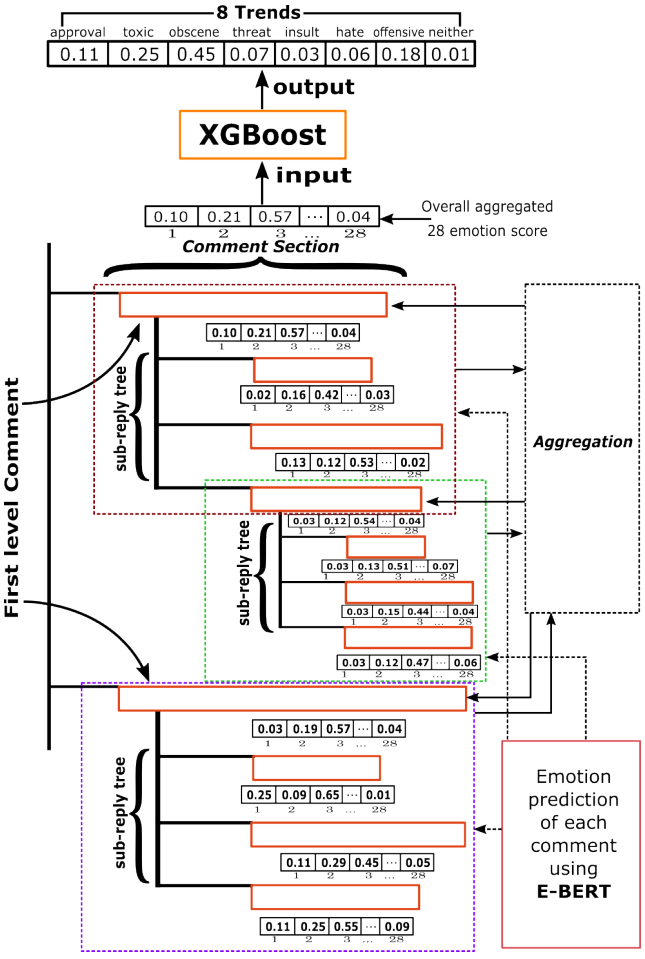

Figure 3: Score aggregation in comment threads.

Two-Step Approach. In the two-step method, we first predict emotions from the comments using E-BERT and then deduce the trend based on these identified emotions using traditional models. In the first step, we predict the emotion scores of the comments using our fine-tuned E-BERT. The output from E-BERT then serves as input for the traditional models in the second step. For this step, we train a variety of traditional machine learning models using the dataset $D_{derived}$ (described in the next paragraph). This training process involves fine-tuning the models with various multi-level classifiers and implementing the GridSearch technique for optimization (Kumari, Suresh, and Dhananjaya 2022). This process results in a fine-tuned model, which is then utilized to predict the trend probabilities for each comment in the comment section.

In Figure 4, we illustrate the execution of the two-step process. In the "Generating Emotion" phase, we initially employ the dataset D_{trend} to generate emotion probabilities using E-BERT. Subsequently, the output of E-BERT undergoes the Training Phase. Before being fitted into XGBoost, each text in D_{trend} is replaced by its corresponding E-BERT output, resulting in a derived dataset referred to as $D_{derived}$. In this dataset, we have a total of 28 feature sets and eight trends as target columns. We utilize 80% of this derived dataset to train XGBoost. In the trend prediction phase, 20% of $D_{derived}$ serves as the test dataset for predicting trend probabilities.

Score Aggregation

To obtain the collective emotional and trend probability scores for an entire comment section, it is essential to aggregate individual comment emotion probabilities into a final comprehensive score. Consider a scenario where we have n replies to a single comment. The aggregation technique for each emotion score is applied as follows: Here, $e_{(i,1)}$ denotes the ith reply's first emotion of 28 emotions

$$i \in [1, ..., n]$$
$$[e_{(i,1)}, e_{(i,2)}, e_{(i,3)}, e_{(i,4)}, e_{(i,5)}, ..., e_{(i,28)}]$$
$$[e_{(i+1,1)}, e_{(i+1,2)}, e_{(i+1,3)}, e_{(i+1,4)}, e_{(i+1,5)}, ..., e_{(i+1,28)}]$$
$$.....$$
$$[e_{(n-1,1)}, e_{(n-1,2)}, e_{(n-1,3)}, e_{(n-1,4)}, e_{(n-1,5)}, ..., e_{(n-1,28)}]$$
$$[e_{(n,1)}, e_{(n,2)}, e_{(n,3)}, e_{(n,4)}, e_{(n,5)}, ..., e_{(n,28)}]$$

The aggregation is performed for each emotion individually, creating a distribution d_e for each emotion. For example, the distribution for the first emotion across all n comments is:

$$d_e \leftarrow [e_{(i,1)}, e_{(i+1,1)}, e_{(n-1,1)}, ..., e_{(n,1)}]$$
$$E_1 \leftarrow Aggregation(d_e)$$

Upon aggregating the emotion distribution d_e, we obtain a final score for each emotion. This process is repeated for all 28 emotions, resulting in a comprehensive array of final scores: $[E_1, E_2, E_3, ..., E_{28}]$. These aggregated scores provide an overall emotional landscape of the comment section, encapsulating the sentiment dynamics in a consolidated format.

Z-score Approach. The distribution of emotional scores in online comment threads is crucial for accurate sentiment analysis. In lengthy comment threads with numerous replies, the use of arithmetic mean for sentiment aggregation (Hossain et al. 2023) may be insufficient, particularly when significant outliers or emotional shifts occur, as these can skew

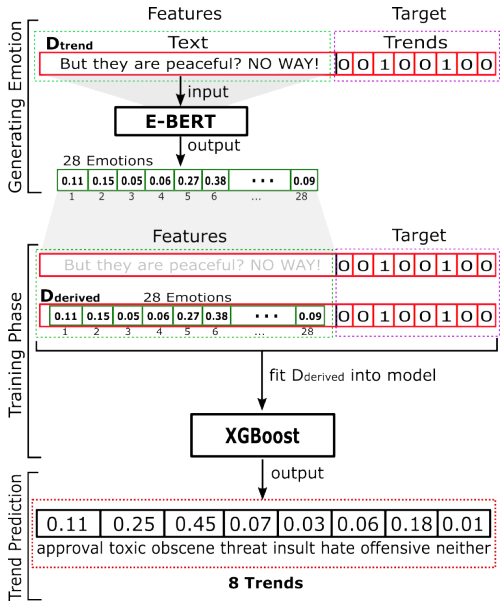

Figure 4: Text in D_{trend} is updated with emotions probability scores to create $D_{derived}$ dataset which is used to train XGBoost.

the overall sentiment interpretation (Anusha et al. 2019; Seo 2006). To mitigate this, we incorporate the Z-score in our aggregation methodology.

Z-scores standardize the sentiment scores across comment threads, neutralizing the impact of outliers. This normalization is essential in threads with diverse emotional expressions, where certain comments may significantly deviate from the average sentiment. For example, within a sub-reply tree, varying emotional intensities (such as higher offensive or hate scores in specific replies) can lead to outliers. This method is critical for distinguishing between genuine sentiment trends and irregularities caused by atypical replies. These outliers, if not addressed, can distort the aggregated sentiment score. The Z-score approach recalibrates these scores to a common scale, ensuring a more accurate reflection of the collective sentiment in the comment thread (FasterCaptial 2017). The Z-score is computed as follows:

$$i \in [1, ..., n]$$

$$\sigma_{i,[0,n]} \leftarrow \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j - \mu)^2}$$

$$\mathcal{Z}_{i,[0,n]} \leftarrow [\frac{x_0-\mu}{\sigma_{i,[0,n]}}, ..., \frac{x_n-\mu}{\sigma_{i,[0,n]}}]$$

$$\mathcal{N}_{i,[0,n]} \leftarrow \frac{\mathcal{Z}_{i,[0,n]} - min(\mathcal{Z}_{i,[0,n]})}{max(\mathcal{Z}_{i,[0,n]}) - min(\mathcal{Z}_{i,[0,n]})}$$

$$A_{[1,28]} \leftarrow [mean(\mathcal{N}_{1,[0,n]}), ..., mean(\mathcal{N}_{28,[0,n]})]$$

The standard deviation (SD), represented by $\sigma_{i,[0,n]}$, is computed for each specific emotion among a set of n replies, including the comment itself. Subsequently, Z-scores ($Z_{i,[0,n]}$) are calculated, followed by normalizing the scores. The normalized outcomes are then utilized to determine the average score for each emotion among the n responses. This average score is designated as the final aggregated score for a particular sub-reply tree. This process is repeated for all 28 emotions, resulting in an array of size 1x28 containing the aggregated scores that will address on final emotions scores.

Weighted Average Approach. In datasets derived from real-world scenarios, outliers are common and can often distort the overall data trend due to their disproportionate impact (Dash et al. 2023). These outliers may arise from contextually misplaced comments or those with reduced relevance. To address this issue, we introduce a method that assigns lesser weight to such outliers, thereby moderating their influence on the overall sentiment analysis.

Consider an emotion distribution set expressed as $[x_1, x_2, x_3, x_4, \ldots, x_n]$, where n is the total number of replies to a specific comment. Defining the median of this set as $\mathcal{M}$, we calculate the deviation of each emotion score from this median, denoted as $\delta_i = |M - x_i|$. This deviation is crucial in determining the weight for each score, given by $\omega_i = \frac{1}{\delta_i} / \sum_{j=1}^{n} \frac{1}{\delta_j}$ (Talukder and Islam 2022).

The weighted average for each emotion symbolized as $\bar{a}$, is computed by summing the products of each weight ω_i with its corresponding emotion score x_i. This process is replicated for all 28 emotions, leading to an array of weighted averages $\mathcal{A} = [\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_{28}]$. This array $\mathcal{A}$ effectively represents the aggregated emotion scores for the entire sub-reply tree, ensuring a balanced assessment that minimizes the impact of outliers.

$$\delta_i \in [\delta_1, \delta_2, \delta_3, ..., \delta_n]$$

$$\omega_i \in [\frac{\delta_1^{-1}}{\sum_{i=1}^{n}\delta_i^{-1}}, \frac{\delta_2^{-1}}{\sum_{i=1}^{n}\delta_i^{-1}}, ..., \frac{\delta_n^{-1}}{\sum_{i=1}^{n}\delta_i^{-1}}]$$

$$\bar{a} = \sum_{i=1}^{n} \omega_i * x_i$$

$$\mathcal{A} = [\bar{a}_1, \bar{a}_2, ..., \bar{a}_{28}]$$

Hybrid Approach. In our pursuit of a more nuanced score aggregation, we introduce a hybrid approach that synergistically combines the strengths of the Z-score and weighted average methods. In our hybrid approach, we commence with the application of the Z-score method to our dataset, formulated as:

$$\mathcal{Z}_{i,[0,n]} \leftarrow \left[\frac{x_0 - \mu}{\sigma_{i,[0,n]}}, \ldots, \frac{x_n - \mu}{\sigma_{i,[0,n]}}\right],$$

to systematically identify and quantify outliers within the sentiment data. This step delineates the data distribution's extremities, particularly in discerning emotionally significant deviations from the mean (μ).

Subsequently, we transition to the weighted average methodology. We innovate by employing an inverse-proportional weighting scheme based on the computed Z-scores. Specifically, the weights (ω_i) for each data point are inversely correlated with their Z-score magnitude, cal-

Algorithm 1: **Recursive method for score aggregation**

Require: $\mathcal{P}_c$: Parent Comment, $\mathcal{S}_i$: scores of ith reply comment, $\mathcal{S}_a$: list of $\mathcal{S}_i$, t : output of the tokenizer, $\mathcal{S}_c$: scores of parent comment
Ensure: A: Overall scores for whole comment thread

RECUR($\mathcal{P}_c$) :
if $replies[\mathcal{P}_c] == 0$ **then**
  return //if a comment doesn't have any reply.
end if
for $i \in replies[\mathcal{P}_c]$ **do**
  RECUR(i) //recursively call function for traversing the whole reply tree
  if $replies[\mathcal{P}_c].size == 0$ **then**
    $t \leftarrow tokenizer(comments[i])$
    $\mathcal{S}_{i,[1,28]} \leftarrow$ TextClassifier(t)
    $\mathcal{S}_a.append(\mathcal{S}_{i,[1,28]})$
  else
    $\mathcal{S}_a.append(scores[i])$
  end if
end for
$t \leftarrow tokenizer(comments[\mathcal{P}_c])$
$\mathcal{S}_c \leftarrow$ TextClassifier(t)
$A \leftarrow$ Aggregation($\mathcal{S}_a, \mathcal{S}_c$) //aggregating the scores of all reply comments.
$scores[\mathcal{P}_c] \leftarrow A$

culated as:

$$\omega_i = \frac{1}{\mathcal{Z}_{i,[0,n]}} / \sum_{j=1}^{n} \frac{1}{\mathcal{Z}_{j,[0,n]}}.$$

$$\bar{a} = \sum_{i=1}^{n} \omega_i * x_i$$

This strategy effectively moderates the influence of outliers, ensuring a more representative sentiment analysis.

The final sentiment score for each emotion is derived by calculating the weighted mean ($\bar{a}$) of the emotion scores across the dataset, with weights adjusted as per the inverse Z-score scheme. This process results in a comprehensive array:

$$\mathcal{A} = [\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_{28}],$$

representing a balanced aggregation of the sentiment scores.

This hybrid approach leverages the outlier detection strengths of the Z-score method and the balanced representation of the weighted average method.

Recursive Method for Score Aggregation. In Figure 3, a section of a reply tree is depicted, where each node represents a reply comment. Our machine-learning model has generated scores for each emotion category. Initially, all leaf nodes have received emotional scores, and by applying our Hybrid Approach, we obtain the aggregated score for a particular parent node, including the node itself. This calculation is carried out for every node in the reply tree. To facilitate this process, we have designed algorithm 1, which functions recursively to calculate the score of each node and ultimately returns the aggregated score for the entire comment thread. In this context, $\mathcal{P}_c$ represents an identifier for a comment that has replies, while the scores of each comment are indicated as $\mathcal{S}_{i,[1,28]}$, where i represents each individual reply comment and $[1, 28]$ refers to the range of emotions (a total of 28). To explore the entire tree of replies, a recursive function is utilized, enabling iteration over each reply comment for every comment. When a comment has no replies, it undergoes tokenization and classification techniques. The scores for each emotion in the comment are then added to the collection $\mathcal{S}_a$. Conversely, if a comment has at least one reply, the previously calculated scores are appended to the list. Once all iterations are complete, the comment identified by $\mathcal{P}_c$ is chosen to generate the scores, with $\mathcal{S}_c$ representing the scores for the comment with the identifier $\mathcal{P}_c$. Finally, the scores of comments and their corresponding reply comments are aggregated and saved.

Experiment

Dataset

Our study utilizes two distinct corpora: $D_{emotion}$ (GoEmotion dataset), D_{trend} (Twitter and Wikipedia datasets combined) that are obtained from three different data sources: Reddit, Twitter, and Wikipedia. Additionally, we crawled users' comments from Fox News articles to evaluate our UI.

GoEmotion. The GoEmotion dataset (denoted as $D_{emotion}$), as described by Demszky et al. (Demszky et al. 2020), is a multi-class, multi-label emotion dataset containing 58k carefully curated and manually annotated Reddit comments. The dataset is labeled with 27 emotion categories along with a neutral label, totaling 28 different classes. Researchers at Amazon Alexa, Google Research, and Stanford curated this dataset, which includes train, validation, and test splits containing 43,410, 5,426, and 5,427 examples respectively. The source language producers are English-speaking Reddit users, with annotations provided by 3 English-speaking crowdworkers. Demszky et al. performed transfer learning experiments using established emotion benchmarks, demonstrating the dataset's robustness across various domains and emotion taxonomies.

Twitter. Sourced from Hugging Face, this dataset contains over 24k tweets labeled for hate speech, offensive language, and neutrality (Davidson et al. 2017). It provides insights into the categorization of different types of offensive language, including racist, homophobic, and sexist content.

Wikipedia. This dataset, obtained from a Kaggle competition (cjadams 2017), features approximately 160k comments from English Wikipedia Talk pages, labeled for various types of toxicity (Wulczyn, Thain, and Dixon 2017b). Annotations were gathered from around 5000 crowdworkers assessing comment toxicity. The dataset helps to identify a broader range of toxic behaviors, including five distinct trend categories, which we restructured to align with our eight-trend framework. Specifically, we added three more trends: 'offensive' and 'neither' from (Davidson et al. 2017) and 'approval' from (Demszky et al. 2020) with the already available five distinct trends making a total of eight trend classes named [approval, toxic, obscene, threat, insult,

hate, offensive, neither]. So, our final dataset D_{trend} consists of Twitter and Wikipedia comments where each text is labeled with either of the 8 trends mentioned above.

FoxNews. To evaluate our UI with real-world data, we extracted user comments from Fox News articles using a custom Python script integrated with Selenium. This script adeptly navigates the website's hierarchical comment structure, capturing essential details such as content, timestamps, and user IDs. These elements are pivotal for our system's functionality, enabling dynamic trend display and chronological organization. Each comment is precisely timestamped to reflect its original posting time on the Fox News site, ensuring real-time data representation. Utilizing a tree-structured approach, we efficiently aggregate and store the scraped data, setting the stage for thorough preprocessing and subsequent analysis.

Preprocessing

Our preprocessing strategy involves merging the Twitter and Wikipedia datasets to form D_{trend}, structured with columns [text, approval, toxic, obscene, insult, threat, hate, offensive, neither]. Texts from both datasets are consolidated into the 'text' column, and corresponding values are aligned in the eight output columns, each representing a trend category. Where column data is not shared between the datasets, we assign zero values.

For both $D_{emotion}$ and D_{trend}, preprocessing involves a series of text-cleaning steps. Before tokenizing, we perform the word or sign elimination technique, and the elimination is applied to each text for both $D_{emotion}$ and D_{trend} datasets. The GoEmotion dataset $D_{emotion}$ requires minimal cleaning, including the removal of punctuation, stopwords, non-ASCII characters, bad quotes, and contraction fixes. In contrast, D_{trend} undergoes more extensive cleaning. This includes the aforementioned steps plus the elimination of numbers, emails (Atagün, Hartoka, and Albayrak 2021), hashtags, emojis, user handles, URLs (Pota et al. 2021), HTML tags, multiple spaces, terms in a bracket, date, and short words by using python NLP package NeatText (Vidhya 2021; Jesse E.Agbe 2020). After this data cleaning phase, we use a BERT tokenizer to tokenize the dataset to fit into the model.

Model Finetuning

BERT. We performed fine-tuning of the $BERT_{base}$, $BERT_{large}$, and $RoBERTa_{large}$ models both for dataset $D_{emotion}$ and dataset D_{trend} separately. Subsequently, we engaged in transfer learning, involving the addition of the number of classes (for $D_{emotion}$ num_class=28, for D_{trend} num_class=8) in the final classification layer. Key parameters for this process include a maximum token size of 30 (Demszky et al. 2020), batch sizes of 16 (Demszky et al. 2020; Mosbach, Andriushchenko, and Klakow 2020) and 32 (Zhang et al. 2020), a learning rate of 2e-5 (Demszky et al. 2020; Mosbach, Andriushchenko, and Klakow 2020), and a fixed number of epochs set to 3 (Demszky et al. 2020; Mosbach, Andriushchenko, and Klakow 2020; Zhang et al. 2020) for all our experiments. Post-fine-tuning, we refer to these models as E-BERT (E for Emotion) and T-BERT (T for Trend), as depicted in Figure 5.

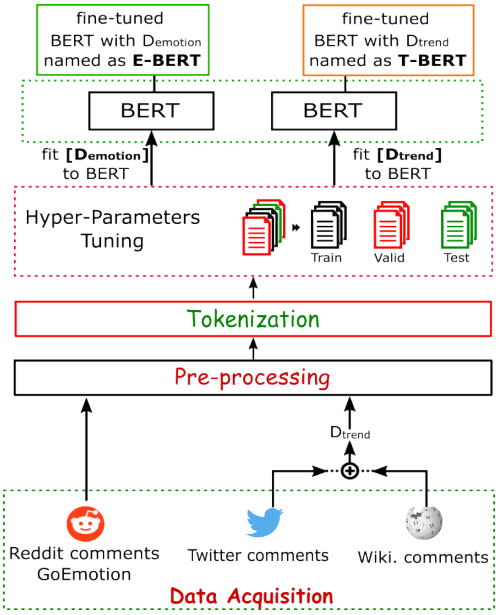

Figure 5: Overview of data processing and fine-tuning BERT model using $D_{emotion}$ and D_{trend} datasets.

Traditional Model. We have fine-tuned several machine learning models, including Logistic Regression, Naive Bayes, Decision Tree, Support Vector Machine, Boosting, Bagging, Random Forest, and XGBoost, with a variety of multi-level classifiers such as MultiOutputClassifier (MOC), OneVsRestClassifier (ORC), ClassifierChain (CC), LabelPowerset (LP), and BinaryRelevance (BR). This selection is tailored to predict eight distinct trend probabilities. The models are tuned using the GridSearch technique (Kumari, Suresh, and Dhananjaya 2022), utilizing the dataset $D_{derived}$. This dataset comprises 28 feature sets and eight trend categories as target columns. Furthermore, we employ k-fold cross-validation, setting k=10 (Kumari, Suresh, and Dhananjaya 2022), to ensure a thorough and reliable assessment of the model's performance.

Results

We evaluated the performance of three fine-tuned BERT models ($BERT_{base}$, $BERT_{large}$, and $RoBERTa_{large}$) using the datasets $D_{emotion}$ and D_{trend}. The outcomes, as detailed in Table 1 and Table 2, reveal that $RoBERTa_{large}$ consistently outperforms the other models, achieving F1 scores of 0.75 and 0.72 for $D_{emotion}$ and D_{trend}, respectively. Additionally, $RoBERTa_{large}$ registers the highest AUC values among the models. Consequently, we designate the fine-tuned $RoBERTa_{large}$ models as E-BERT for $D_{emotion}$ and T-BERT for D_{trend}.

Table 3 presents the performance metrics of all models fine-tuned with the $D_{derived}$ dataset. In this comparison, Decision Tree and XGBoost exhibit notably higher micro-average Recall scores. However, Random Forest achieves

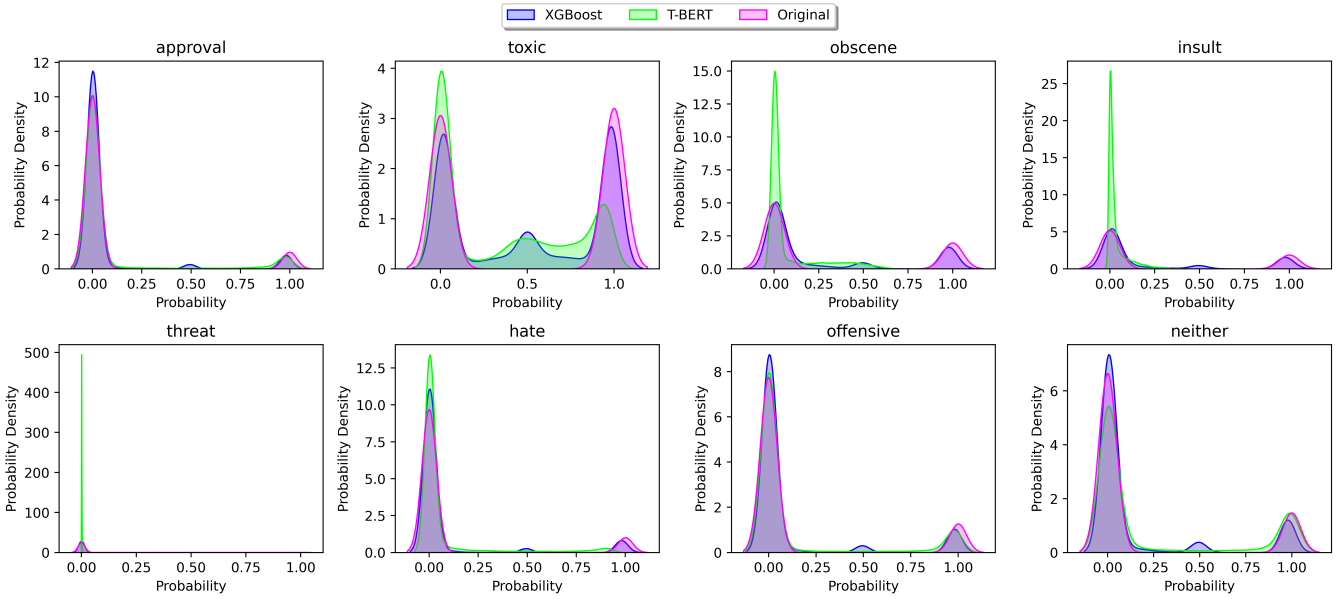

Figure 6: Deviance of T-BERT and XGBoost with the original baseline data.

Model	F1	Recall	Precision	AUC
BERT$_{base}$	0.74	0.73	0.77	0.90
BERT$_{large}$	0.74	0.72	0.78	0.91
RoBERTa$_{large}$	**0.75**	**0.72**	**0.79**	**0.92**

Table 1: Performance of E-BERT model on $D_{emotion}$ dataset.

Model	F1	Recall	Precision	AUC
BERT$_{base}$	0.69	0.65	0.75	0.83
BERT$_{large}$	0.71	0.66	**0.78**	0.84
RoBERTa$_{large}$	**0.72**	**0.68**	**0.78**	**0.86**

Table 2: Performance of T-BERT model on D_{trend} dataset.

	Micro-Avg			Macro-Avg		
Models	Precision	Recall	F1	Precision	Recall	F1
Bagging-ORC	0.86	0.70	0.77	0.58	0.33	0.38
Boosting-ORC	0.85	0.69	0.76	0.55	0.32	0.35
DecisionTree	0.73	**0.73**	0.73	0.43	**0.42**	0.42
LR-CC	0.79	0.65	0.72	0.22	0.24	0.23
LR-LP	0.80	0.67	0.73	0.38	0.26	0.26
MultiNB-BR	0.82	0.57	0.68	0.23	0.19	0.20
MultiNB-CC	0.78	0.62	0.69	0.40	0.26	0.29
MultiNB-LP	0.72	0.59	0.65	0.22	0.19	0.19
RandomForest	**0.88**	0.70	0.78	**0.68**	0.32	0.36
SVC-MOC	0.86	0.63	0.72	0.46	0.24	0.25
XGB-ORC	0.86	**0.73**	**0.79**	0.65	0.38	**0.44**

Table 3: Performance of ML models.

the highest Precision scores, recording 0.88 and 0.68 in micro and macro averages, respectively. XGBoost stands out in terms of F1 scores, leading with 0.79 in micro-average and 0.44 in macro-average. Given that the F1 score is the harmonic mean of Precision and Recall, XGBoost is identified as the most effective model. Consequently, it is deployed for trend prediction in the comment section following the emotion prediction phase, underscoring its superior performance in our model evaluations.

Direct vs. Two-Step Approaches for Trend Prediction

To ascertain the most effective method for trend prediction that addresses **RQ4**, we analyzed the outcomes of both the direct and two-step approaches. We trained T-BERT and XGBoost models and then evaluated them using a test set, which comprises 20% of the D_{trend} dataset. The comparative results are depicted in Figure 6, showcasing the model outputs against the baseline.

In this figure, the x-axis represents the probability scores, and the y-axis denotes the probability density. The visual comparison highlights the differences in probability density among the original data, T-BERT, and XGBoost outputs. Notably, XGBoost, depicted as a purple mountain in the graph, closely aligns with the original probability density. In contrast, T-BERT, represented in green, shows a divergence, either exceeding or falling short of the original data's distribution. This comparison suggests that XGBoost, within the two-step approach, demonstrates a more accurate trend prediction, closely matching the baseline probability density compared to the direct approach employed by T-BERT.

Selecting the Best Aggregation Method

In our analysis, we explored the effectiveness of Z-score, weighted-average, and hybrid aggregation methods, as men-

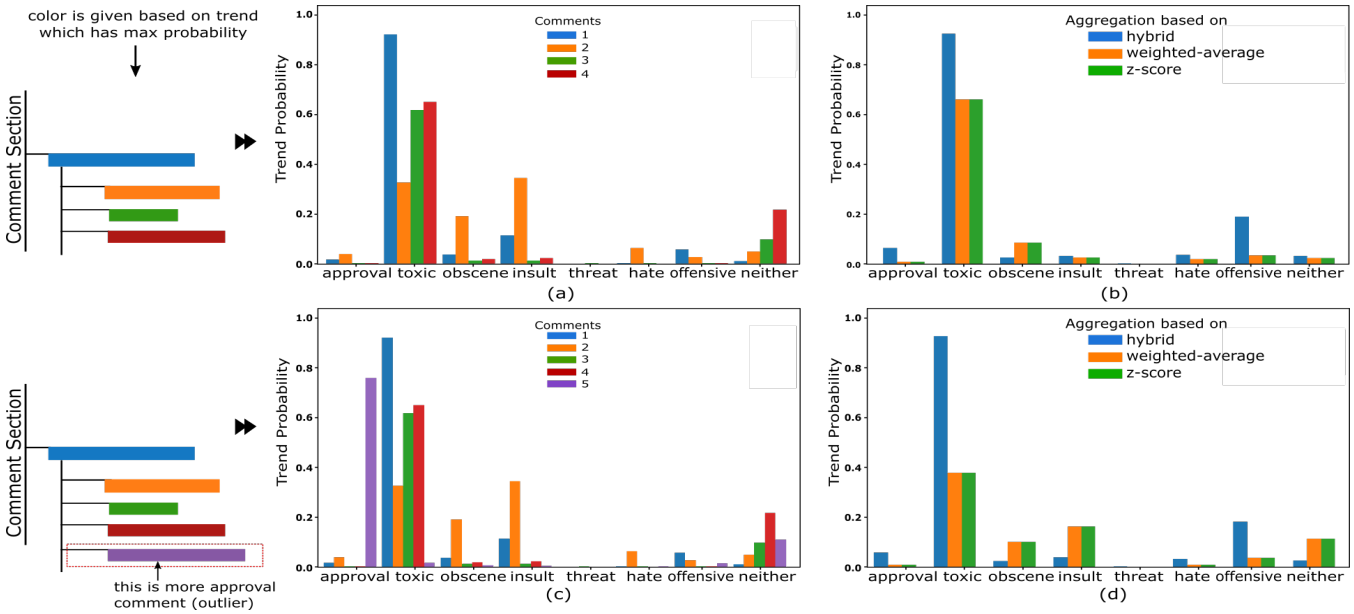

Figure 7: Plots a & b show the trend probability of a portion of the comment section without an outlier, and c & d show the probability when there is an outlier.

tioned earlier to address **RQ5**. To determine the most robust aggregation technique, we compare different aggregation methods' resiliency to outliers by intentionally introducing the outlier. We focused our experiments on a subset of the comment tree, comprising a parent comment and a minimum of three replies (as shown in Figure 7). In this figure, plot (a) displays the trend output for each comment using a bar graph, while plot (b) presents the aggregated results for these comments using the three different aggregation techniques.

Plot (a) represents the trend probability for each individual comment, where each comment's dominant trend is indicated by the color of its corresponding bar. In this plot, one comment—visually distinguished by a blue bar—displays a significantly higher toxicity level than the rest.

Plot (b) provides a comparison of the aggregated results using Z-score, weighted-average, and hybrid methods. Despite the heightened toxicity of one comment, the weighted-average, and hybrid methods aggregation do not reflect a significant deviation, resulting in a uniform appearance across all trends. The Z-score approach, however, reveals a clear disparity, with a pronounced probability for the 'toxicity' trend, which most comments incline toward.

When a new reply with extreme 'approval' (purple color) is introduced, as shown in plot (c), this comment becomes an outlier, depicted by the dashed red outline. This addition tests the robustness of each aggregation method. In plot (d), the hybrid method exhibits its resilience, effectively incorporating the outlier while still prioritizing the 'toxicity' trend. On the other hand, the Z-score method overemphasizes the outlier's impact, disproportionately increasing the aggregated 'approval' score.

The results underscore the hybrid approach's superior management of outliers, ensuring a more balanced trend representation, even when faced with extreme data points. This efficacy in mitigating outlier impact positions the hybrid method as the preferred choice for trend aggregation in sentiment analysis within comment sections. It adeptly preserves the integrity of the overarching trend distribution, despite the presence of individual extreme values. This outcome encourages us to persist with the hybrid approach for aggregation, as it not only captures the essence of the majority but also accommodates the extremities without letting them skew the overall analysis.

User Interface and Evaluation

We have developed an intuitive user interface to dynamically monitor sentiment trends in online discussions. Our system, which aggregates comment threads from Fox News, integrates a bar graph visualization to reflect sentiment changes over time, as demonstrated in Figure 8. This visualization is particularly revealing when tracking the sentiment evolution as new comments are added to the system.

The interface is equipped with tabs for emotion, trend, and sentiment analysis, offering a multifaceted view of user interactions within the comment section. The emotion tab, for instance, displays a time-sensitive sentiment scale across 28 different emotions, allowing for an analysis of how specific events or comments alter the sentiment landscape.

An additional feature is the trend-wise metric tracker, accessible via a dropdown menu, which delineates the sentiment progression for various emotions over time. This feature corresponds to the timeline presented in the graph, offering a dynamic representation of sentiment shifts.

For a more engaging experience, users can utilize the 'play' function, which animates the trend transitions, nar-

Figure 8: This is a demonstration of the user interface in our trend or emotion prediction application, which is embedded above the user comments. The users has the ability to view the sentiments of the comment thread over a certain time period.

rating the sentiment story of the comment section.

Testing in the Wild

To comprehensively assess the user experience within our interactive comment section UI, we conducted a survey involving 50 diverse participants. The survey aimed to gather both qualitative feedback and quantitative insights on satisfaction, usability, graph visibility, engagement, and overall user sentiments.

Participants. The participant group represented a broad spectrum, including individuals from various age groups, genders, and professional backgrounds. Each participant's frequency of interaction with similar comment sections or interfaces added depth to the analysis.

Satisfaction and Usability Analysis. Participants were prompted to rate their satisfaction and perceived usability of the UI on a Likert scale. The mean satisfaction rating across all participants was 3.5, indicating a generally positive sentiment. Similarly, the mean usability rating was 3.5, reflecting a favorable evaluation. Figures 9(a)(b) provide a visual representation of the distribution of satisfaction and usability ratings.

Graph Visibility and Engagement. Quantifying participants' responses regarding graph visibility and engagement revealed noteworthy patterns. Approximately 80% of participants found the sentiment and emotional graphs to be clear and visible, while 60% engaged with the graphs frequently. These values indicate a positive user perception of the visual elements in our UI. Figures 9(c)(d) present a detailed breakdown of these metrics.

Recommendations and Future Use. Quantitative insights into participants' likelihood to recommend the UI and intentions for future use were obtained. An overwhelming 85% of participants expressed a likelihood to recommend the UI, and 70% indicated a positive inclination towards using it again in the future. These percentages demonstrate a strong endorsement of the UI. Figures 9(e)(f) visually represent these recommendations and future use intentions.

Ethical Considerations

We have developed our protocols to collect and analyze our data in an ethical, IRB-approved manner. For analysis purposes, we only stored anonymized data that we collected from the different corpora whose handling does not fall within the PII definition of NIST SP 800-122 (NIS 2021). Under GDPR (GDP 2021), the use of the information without context, e.g., name or personal identification number, is not considered to be "personal information".

Discussion and Limitations

This study, while pioneering in its approach, acknowledges several limitations and ethical considerations. Primarily, the visual depiction of sentiment dynamics, despite its innovation, may not encompass the entire spectrum of human emotions. This risk of oversimplification could lead to the omission of nuanced sentiments inherent in specific comments. Moreover, the applicability of our methodology across diverse social media platforms is not guaranteed due to the variable nature of user interactions and the format of content presentation. There is also an ethical dimension to consider, particularly in the interpretation and visualization of user sentiments without explicit consent, which may raise privacy concerns.

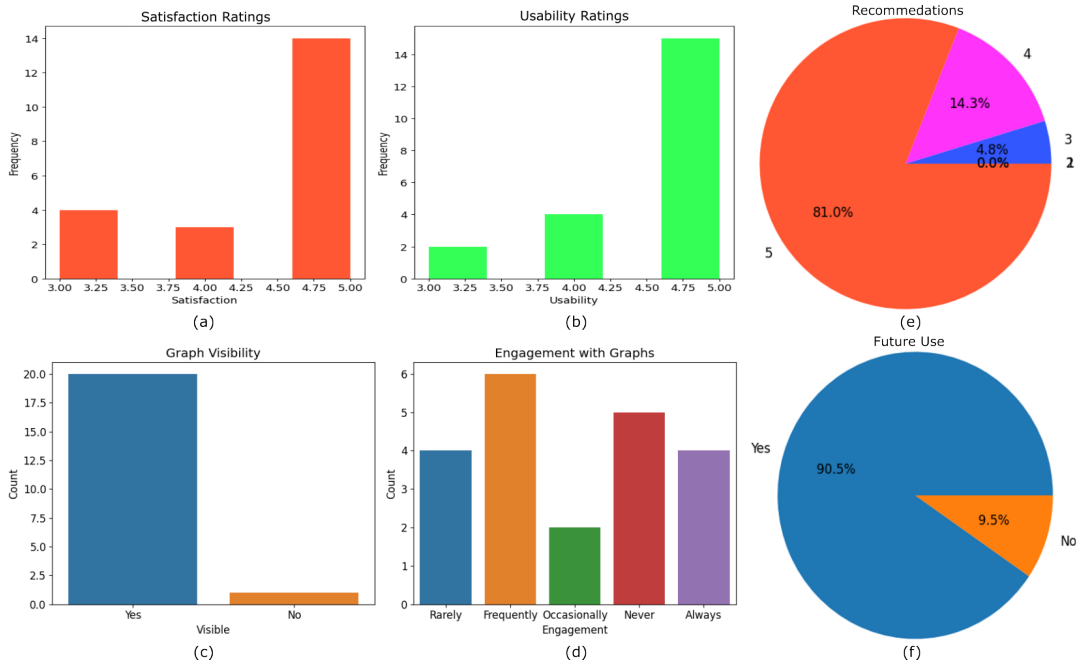

Figure 9: The evaluation metrics represent the user feedback for the application we built for trend or emotion prediction. These results were collected as part of sending the feedback form to all of our users. We have used google forms as a feedback templates and generated graph based on the output we received.

Another point of consideration is the potential for misinterpretation of our visual sentiment analysis. Without proper contextual understanding or sufficient background knowledge, users may draw incorrect conclusions from the visual data. Additionally, inherent biases in sentiment analysis represent a significant challenge. These biases might stem from the training data or the algorithms interpreting sentiments, potentially leading to skewed or unrepresentative insights. Lastly, while we utilized a hybrid approach for neutralizing outliers during the aggregation process in comment sections, the efficacy of this system may vary depending on the depth and nature of the discussions. For instance, on platforms like Reddit, subtrees within comment threads often diverge significantly, sometimes discussing topics that are entirely unrelated to each other.

Conclusion

The rapid growth of social media platforms has provided a dynamic gauge of public sentiment, capturing the pulse of global discourse in real time. This study has introduced a cutting-edge framework for the meticulous analysis of sentiment shifts within social media comment sections—vital indicators of the evolving public conversation. By harnessing the power of a pre-trained uncased $RoBERTa_{large}$ model, we have successfully predicted emotional scores from vast swathes of user comments, classifying these scores into pivotal sentiment trends that range from Approval to Neutral. Our approach has integrated sophisticated machine learning models to train a novel dataset, linking emotional scores with corresponding sentiment trends, and in doing so, has facilitated the generation of trend probability scores. A bottom-up recursive algorithm was instrumental in aggregating these emotional scores across comment threads, culminating in the prediction of trend scores through three innovative aggregation methods. The empirical evidence from our research underscores the robustness of our emotional prediction model, which boasts an AUC of 0.92. Notably, XGBoost emerged as a high performer, achieving an F1 score that surpasses 0.40. The implications of our findings extend beyond the academic, providing practical insights for enhancing online interactions, guiding interventions, and refining content moderation strategies. Ultimately, this research contributes a granular understanding of the digital social fabric, charting the trajectory of sentiment in an increasingly interconnected world.

Acknowledgments

This research was supported by NSF grant CNS-2153482.

References

2021. General Data Protection Regulation (GDPR). https://gdpr-info.eu/. Accessed: 2021-02-12.

2021. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII). https://tinyurl.com/ylyjst5y. Accessed: 2021-02-12.

Anusha, P. V.; Anuradha, C.; Murty, P. S. C.; and Kiran, C. S. 2019. Detecting outliers in high dimensional data sets using Z-score methodology. *International Journal of Innovative Technology and Exploring Engineering*, 9(1): 48–53.

Atagün, E.; Hartoka, B.; and Albayrak, A. 2021. Topic Modeling Using LDA and BERT Techniques: Teknofest Example. In *2021*

6th International Conference on Computer Science and Engineering (UBMK), 660–664. IEEE.

Bollen, J.; Mao, H.; and Pepe, A. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the international AAAI conference on web and social media*, volume 5, 450–453.

Bollen, J.; Mao, H.; and Zeng, X. 2011. Twitter mood predicts the stock market. *Journal of computational science*, 2(1): 1–8.

Chang, J. S.; and Danescu-Niculescu-Mizil, C. 2019. Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop.

cjadams, J. E. L. D. M. M. n. W. C., Jeffrey Sorensen. 2017. Toxic Comment Classification Challenge.

Dash, C. S. K.; Behera, A. K.; Dehuri, S.; and Ghosh, A. 2023. An outliers detection and elimination framework in classification task of data mining. *Decision Analytics Journal*, 6: 100164.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, 512–515.

Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; and Ravi, S. 2020. GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

FasterCaptial, S. 2017. Z-Scores and Their Significance. http://tinyurl.com/5n8psvc6. Dataset.

Fortuna, P.; and Nunes, S. 2018. A Survey on Automatic Detection of Hate Speech in Text.

Founta, A.-M.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior.

Hessel, J.; and Lee, L. 2019. Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features.

Hossain, I.; Puppala, S.; Alam, M. J.; and Talukder, S. 2023. Monitoring Dynamics of Emotional Sentiment in Social Network Commentaries.

Jesse E.Agbe, J. S. 2020. NeatText: A simple NLP package for cleaning textual data and text preprocessing. Simplifying Text Cleaning For NLP MLa. https://github.com/Jcharis/neattext?tab=readme-ov-file. Accessed: 2020-12-31.

Jurgens, D.; Hemphill, L.; and Chandrasekharan, E. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse.

Kumari, H. V.; Suresh, D.; and Dhananjaya, P. 2022. Clinical Data Analysis and Multilabel Classification for Prediction of Dengue Fever by Tuning Hyperparameter using GridsearchCV. In *2022 14th International Conference on Computational Intelligence and Communication Networks (CICN)*, 302–307. IEEE.

Lee, S. Y.; and Ryu, M. H. 2019. Exploring characteristics of online news comments and commenters with machine learning approaches. *Telematics and Informatics*, 43: 101249.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 14867–14875.

Mohammad, S. M.; and Kiritchenko, S. 2018. Understanding Emotions: A Dataset of Tweets to Study Interactions between Affect Categories. In *International Conference on Language Resources and Evaluation*.

Mondal, M.; Silva, L. A.; and Benevenuto, F. 2017. *A Measurement Study of Hate Speech in Social Media*.

Mosbach, M.; Andriushchenko, M.; and Klakow, D. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

Niculae, V.; and Danescu-Niculescu-Mizil, C. 2016. Conversational Markers of Constructive Discussions.

Oh, Y. W.; and Park, C. H. 2021. Machine cleaning of online opinion spam: Developing a machine-learning algorithm for detecting deceptive comments. *American behavioral scientist*, 65(2): 389–403.

Pennycook, G.; Bear, A.; Collins, E. T.; and Rand, D. G. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management science*, 66(11): 4944–4957.

Petrović, S.; Osborne, M.; and Lavrenko, V. 2010. Streaming first story detection with application to twitter. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*, 181–189.

Pota, M.; Ventura, M.; Fujita, H.; and Esposito, M. 2021. Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets. *Expert Systems with Applications*, 181: 115119.

Röttger, P.; Vidgen, B.; Nguyen, D.; Waseem, Z.; Margetts, H.; and Pierrehumbert, J. B. 2020. HateCheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*.

Saveski, M.; Roy, B.; and Roy, D. 2021. The structure of toxic conversations on Twitter. In *Proceedings of the Web Conference 2021*, 1086–1097.

Seo, S. 2006. *A review and comparison of methods for detecting outliers in univariate data sets*. Ph.D. thesis, University of Pittsburgh.

Sharma, H. K.; Singh, T.; Kshitiz, K.; Singh, H.; and Kukreja, P. 2017. Detecting hate speech and insults on social commentary using nlp and machine learning. *Int J Eng Technol Sci Res*, 4(12): 279–285.

Shugars, S.; and Beauchamp, N. 2019. Why Keep Arguing? Predicting Engagement in Political Conversations Online:. *SAGE Open*.

Talukder, Z.; and Islam, M. A. 2022. Computationally Efficient Auto-Weighted Aggregation for Heterogeneous Federated Learning. In *2022 IEEE International Conference on Edge Computing and Communications (EDGE)*, 12–22. IEEE.

Vidhya, A. 2021. Cleaning and Pre-processing Textual Data with Neattext Library. https://www.analyticsvidhya.com/blog/2021/10/cleaning-and-pre-processing-textual-data-with-neattext-library/. Accessed: 2021-10-16.

Wang, L.; and Cardie, C. 2016. A Piece of My Mind: A Sentiment Analysis Approach for Online Dispute Detection.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017a. Ex Machina: Personal Attacks Seen at Scale.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017b. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, 1391–1399.

Yao, M.; Chelmis, C.; and Zois, D.-S. 2019. *Cyberbullying Ends Here: Towards Robust Detection of Cyberbullying in Social Media*.

Zhang, J.; Danescu-Niculescu-Mizil, C.; Sauper, C.; and Taylor, S. J. 2018. Characterizing Online Public Discussions through Patterns of Participant Interactions.

Zhang, T.; Wu, F.; Katiyar, A.; Weinberger, K. Q.; and Artzi, Y. 2020. Revisiting few-sample BERT fine-tuning. *arXiv preprint arXiv:2006.05987*.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? Yes
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? No
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA

APPENDIX

Z-Score Example

To illustrate a practical example of how the z-score can guide us through significant changes, we will examine a few sample arrays representing three emotional spectrum's.

[0.09, 0.81, 0.10], **[0.33, 0.33, 0.33]**, **[0.70, 0.15, 0.15]**

We could see that element in first array has a dominant values and if we apply average on above array. This dominant value could significantly impact the overall scores. Applying above strategy, To calculate the Z-scores for each set of numbers, we'll use the formula:

$$Z = \frac{(X - Mean)}{StandardDeviation}$$

Let's calculate the Z-scores for each row of numbers:
For the first row $[0.09, 0.81, 0.10]$:

$$\text{Mean} = \frac{(0.09+0.81+0.10)}{3} = 0.33$$

$$\sigma_{i,[0,3]} = \sqrt{\frac{((-0.24)^2+(0.48)^2+(-0.23)^2)}{3}} \approx 0.34$$

$$\text{Z-scores: } \left[\frac{(0.09-0.33)}{0.34}, \frac{(0.81-0.33)}{0.34}, \frac{(0.10-0.33)}{0.34}\right]$$

$$\text{Z-scores} \approx [-0.89, 1.29, -0.88]$$

For the second row $[0.33, 0.33, 0.33]$:

$$\text{Mean} = \frac{(0.33+0.33+0.33)}{3} = 0.33$$

$$\sigma_{i,[0,3]} = 0$$

Z-scores: $[0, 0, 0]$

For the third row $[0.70, 0.15, 0.15]$:

$$\text{Mean} = \frac{(0.70+0.15+0.15)}{3} = 0.33$$

$$\sigma_{i,[0,3]} = \sqrt{\frac{((0.17)^2+(-0.08)^2+(-0.08)^2)}{3}} \approx 0.12$$

$$\text{Z-scores: } \left[\frac{(0.50-0.33)}{0.12}, \frac{(0.25-0.33)}{0.12}, \frac{(0.25-0.33)}{0.12}\right]$$

So, the Z-scores for the given sets of numbers are approximately:

$$[-0.89,\ \mathbf{1.29}, -0.88], [0, 0, 0], [\mathbf{1.42}, -0.67, -0.67]$$

The initial observations are evident, showing that values initially considered as 0.81 and 0.70 resulted in 1.29 and 1.42, respectively, through our z-score calculation. We may decide to either exclude these arrays or opt for normalization. Upon further normalization, the resulting output for the aforementioned array set of emotions would be:

[0.48, 0.51, 0.00]

As observed, the z-scores effectively capture shifts when there is a substantial change in the trend. For example, focusing on the last column with values 0.10, 0.33, 0.25, we notice stability, indicating less significance compared to the first column with values 0.09, 0.33, 0.50. The emotional values in the first column provide insights into the how and why of the emotional change, given their notable differences.

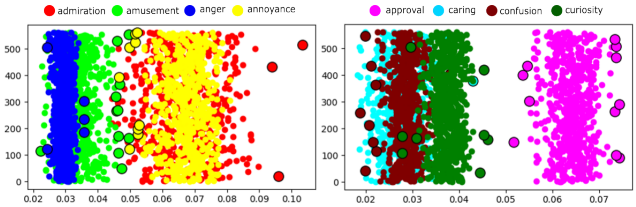

Figure 10: Eight different emotion data are plotted and outliers are shown with the larger circles with black border.

THE QUESTIONNAIRE

User Information Age: .....
Gender:

- Male
- Female

Occupation: .....
How often do you use similar interfaces?

- Daily
- Weekly
- Monthly
- Rarely

Graph Visibility:

- Yes
- No

How often do you interact with the trend and emotion graphs?

- Frequently
- Occasionally
- Rarely
- Never

Comments on UI:

.....
.....
.....
.....

Future Use Will you use this interface again in the future?

- Yes
- No

Likelihood to Recommend ☆☆☆☆☆

Overall Satisfaction ☆☆☆☆☆

Usability Rating ☆☆☆☆☆