mathematics

MDPI

Article

Tracking Fine-Grained Public Opinions: Two Datasets from Online Discourse on Trending Topics

Haihua Xie * and Miao He

Beijing Institute of Mathematical Sciences and Applications (BIMSA), Beijing 101408, China; miaohe@bimsa.cn
* Correspondence: haihua.xie@bimsa.cn

Abstract

In this paper, we introduce two novel, publicly available datasets that capture public opinion on two highly salient and timely topics: the **perceived feasibility of artificial general intelligence (AGI)** and the **Hamas–Israel conflict (HIC).** Collected from social media posts on X(formerly Twitter), the datasets were curated and annotated using a structured methodology designed to ensure robustness, consistency, and interpretability. Whereas prior research on public opinion has primarily relied on theoretical models, controlled experiments, or narrowly scoped datasets, such approaches often fail to capture the complexity and dynamism of real-world discourse. In contrast, the datasets presented in this work provide temporally fine-grained, large-scale empirical data that reflect the evolving nature of public sentiment on contemporary global issues. In addition to detailed sentiment annotations, these resources support longitudinal analyses of opinion dynamics, offering a foundation for empirical studies on consensus formation, polarization, and social influence in digital environments. By releasing these datasets, we aim to advance the empirical study of public opinion in the age of social media.

Keywords: public opinion; fine-grained sentiment analysis; social computing; natural language processing; Twitter dataset; opinion dynamics

MSC: 68-11; 91D30; 62R07; 68T09; 94A16; 68T50



Academic Editor: Dan Vilenchik

Received: 9 August 2025
Revised: 23 September 2025
Accepted: 27 September 2025
Published: 28 October 2025

Citation: Xie, H.; He, M. Tracking Fine-Grained Public Opinions: Two Datasets from Online Discourse on Trending Topics. *Mathematics* **2025**, *13*, 3433. https://doi.org/10.3390/math13213433

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In sociological analysis, tracking public opinion aims to systematically understand how individual beliefs evolve within complex social networks [1]. This line of research provides both theoretical and practical insights for monitoring public sentiment and informing governance strategies [2]. Real-world phenomena such as opinion polarization and extremism [3] are deeply rooted in the mechanisms of social influence [4,5]. Developing accurate models of opinion diffusion not only enhances our ability to predict and assess these challenges, but also supports the design of targeted interventions to mitigate polarization and reduce societal harm [6].

However, progress in this field is hindered by critical data limitations [7]. Much existing research relies on small-scale surveys or cross-sectional datasets, which lack both temporal resolution and causal depth [8]. Moreover, while significant effort has been devoted to curating datasets centered on historical content, comparatively less attention has been paid to capturing contemporary discourse [9]. Real-time public sentiment on trending topics is inherently difficult to measure. Although social media platforms such

as X (formerly Twitter) provide unprecedented access to public opinion data, extracting nuanced sentiment from complex, informal language remains a substantial challenge [10].

In summary, existing resources for analyzing public opinion exhibit three major shortcomings:

(1) **Limited scale and temporal granularity**: Most datasets are based on small-scale simulations, limited voting records, or short-term annotations [8], lacking the longitudinal and large-scale structure required to track the evolution of public sentiment over time.
(2) **Narrow thematic scope**: Previous work has largely focused on a narrow range of historically prominent topics (e.g., political ideology, climate change) [11], overlooking fast-changing, socially salient debates that dominate modern digital discourse.
(3) **Imprecise opinion categorization**: Common annotation schemes often rely on coarse binary labels (e.g., positive/negative) or on continuous sentiment scores derived from subjective assessments [12]. Such approaches lack the expressiveness required to capture nuanced stances and subtle shifts in opinion.

These limitations constrain the development of robust models capable of capturing the dynamic, rapidly evolving nature of opinion formation in internet-mediated social environments.

To address these challenges, we introduce two novel and publicly available Twitter datasets centered on the feasibility of artificial general intelligence (AGI) and the Hamas–Israel conflict (HIC)—two highly socially salient topics collected from user-generated posts on X (https://x.com/home). To facilitate fine-grained analysis of public opinion, we perform sentiment annotation on the collected tweets using a multi-level classification scheme. Our approach leverages the semantic reasoning capabilities of large language models (LLMs) through few-shot in-context learning (ICL), enabling efficient and scalable sentiment analysis. To improve alignment with human judgment, we iteratively refine the LLM outputs using human-annotated samples and manually summarized classification criteria.

As a result, our dataset supports applications in both **fine-grained sentiment analysis** and **public opinion modeling**. Overall, the dataset's design introduces three key innovations:

(1) **Longitudinal Opinion Tracking:** Continuous data collection preserves temporal dependencies, allowing for the detection of evolving public attitudes.
(2) **Alignment with Real-Time Discourse:** The focus on viral narratives ensures relevance to modern, platform-mediated opinion formation mechanisms.
(3) **Multidimensional Annotation:** A dual-layer structure captures both aspect-specific (e.g., the feasibility of AGI) and intensity-graded sentiments, enabling precise modeling of polarization and consensus dynamics.

By integrating large-scale empirical social media data with computationally grounded methods, this work can be applied to sentiment analysis, public opinion analysis, and various applications in computational social science, extending beyond abstract modeling and theoretical simulation. The released datasets offer a valuable resource for interdisciplinary research spanning computational social science, AI ethics, and the study of geopolitical conflicts. The datasets are publicly available at: https://github.com/BIMSA-DATA/FGPO (accessed on 26 September 2025).

The remainder of this paper is organized as follows: Section 2 reviews related work on datasets for public opinion modeling and fine-grained sentiment analysis. Section 3 describes the data collection and processing procedures, with an emphasis on fine-grained sentiment annotation. Section 4 analyzes the temporal patterns and characteristics of public opinion across the two datasets and outlines potential application scenarios. Finally, Section 5 concludes the paper and discusses directions for future research.

2. Related Work

This section systematically reviews datasets on public opinion and fine-grained sentiment analysis, identifying critical gaps in current resources and methodological approaches.

2.1. Datasets for Public Opinion

With the growing prevalence of individuals expressing their views online, numerous studies have constructed public opinion analysis models that account for internet-specific characteristics by collecting and analyzing user-generated content (UGC).

Since our dataset is also derived from user-generated content, Table 1 highlights the key differences between our dataset and several representative, comparable datasets:

- **IsamasRed** [13]: A meticulously compiled dataset containing nearly 400,000 conversations and over 8 million comments from Reddit, spanning August 2023 to November 2023.
- **GoEmotions** [14]: A large-scale, manually annotated dataset of 58,000 English Reddit comments, each labeled with one or more of 27 emotion categories or marked as neutral.
- **TED-S** [15]: Covers two distinct domains—sports and politics—and includes complete subsets of Twitter data streams, annotated with both sub-event and sentiment labels, supporting research on event-based sentiment analysis.
- **Covid19-twitter** [16]: A large-scale dataset developed to evaluate Informed Machine Learning (IML) methods on contrastive discourse relations (CDRs). It contains approximately 100,000 tweets, symmetrically categorized for targeted evaluation.

Table 1. Comparison across datasets.

Attribute \ Dataset	IsamasRed	GoEmotions	TED-S	Covid19-tweet	Ours
User Metadata [a]	No	No	No	No	Yes
Post Interactions [b]	Yes	No	No	No	Yes
Topic-Based [c]	Yes	No	Yes	Yes	Yes
Temporal Continuity [d]	Yes	No	Yes	Yes	Yes
Annotation Method	Model [e]	Manual	Model + Manual	Model	Model + Manual
Label Type	Nominal	Nominal	Ordinal (3 classes)	Ordinal (2 classes)	Ordinal (5 classes)

[a] Whether user-level metadata is included, such as the number of followers/followings, total posts, etc. [b] Whether post-level interaction data is available, such as likes, retweets, and replies. [c] Whether the tweets were collected based on specific topics or themes. [d] Whether the collected data exhibits continuity over time. [e] "Model" refers to annotations generated automatically by a model, while "Manual" indicates annotations created by human annotators. "Model + Manual" means a combination of both.

In addition to UGC-based datasets, more traditional studies obtain data through structured and specialized methods such as voting records and surveys. Electoral records and polling data remain foundational resources due to their structured nature and societal relevance [17]. These datasets enable systematic comparisons between voting models [18] and analyses of population-level opinion distributions.

Notable examples include research on vote distribution patterns, which examines spatial–temporal regularities in candidate preference allocations [19]. For instance, U.S. presidential election datasets (1980–2012) have been used to validate the voter model [20], which successfully replicates county-level vote distributions. Likewise, the noisy voter model produces beta-distributed outputs consistent with Lithuanian parliamentary election data (1992–2012) [21].

While traditional datasets have supported theoretical validation and scenario-based simulations in the study of collective behavior, their usefulness for real-world opinion

analysis is constrained by several key limitations: (1) Scale Constraints: A reliance on small-sample experiments and aggregated statistics limits representativeness; (2) Topical Narrowness: A disproportionate focus on well-studied or conventional issues neglects the rapidly evolving topics emerging from online discourse; and (3) Limited Temporal Coverage: The absence of continuous, longitudinal observations of public opinion hinders the ability to track and analyze changes in attitudes or emotions over time [8].

A promising path toward bridging the gap between theoretical opinion models and empirical evidence lies in the growing use of computerized social experiments, often conducted in collaboration with the general public. These experiments are increasingly enhanced by advanced computational techniques for analyzing large-scale datasets of online content expressing public opinion [22].

Following this direction, the dataset introduced in this paper is sourced from online social platforms and captures large-scale, continuous public discourse on high-salience, real-world topics. It provides rich insights into the dynamic evolution of public opinion in the internet era, fostering a closer alignment between theoretical models and observed behavioral patterns.

2.2. Datasets for Fine-Grained Sentiment Analysis

Sentiment analysis is a fundamental area of research in computational sociology, with wide-ranging applications including public opinion analysis, business intelligence, and beyond [23,24]. It involves extracting users' nuanced perspectives on specific topics, typically achieved through the classification of sentiment in user-generated content [25]. This section provides an overview of existing datasets commonly used for sentiment analysis.

2.2.1. Traditional Sentiment Analysis Datasets

Conventional approaches focus on coarse polarity classification, typically categorizing text into binary (positive/negative) or ternary (positive/neutral/negative) classes. Notable examples include: (1) Stanford Sentiment Treebank (SST) [26]; (2) Twitter US Airline Sentiment Dataset [27]; and (3) IMDB Reviews Dataset [28].

While valuable for basic sentiment analysis, these resources lack the granularity required for detailed sentiment analysis due to their limited classification scope.

2.2.2. Aspect-Based Sentiment Analysis Datasets

Datasets supporting aspect-level sentiment detection are particularly relevant to research on public opinion modeling, as they enable the analysis of opinions directed toward specific entities or attributes. Key resources in this category include: (1) SemEval-2014 Task 4 [29]: Features restaurant and laptop reviews annotated with aspect-term sentiment pairs; (2) SemEval-2015 Task 12 [30]: Extends aspect-based analysis with improved annotation frameworks; and (3) SemEval-2016 Task 5 [31]: Expands the scope to multilingual and multi-domain aspect sentiment analysis.

2.2.3. Intensity-Graded Sentiment Analysis Datasets

Intensity-graded sentiment analysis extends beyond binary or ternary classifications by capturing varying degrees of sentiment strength—for example, distinguishing between *very positive* and *generally positive opinions*. Recent efforts have introduced datasets with expanded label sets to support this more nuanced classification, including: (1) Amazon Reviews Dataset [32]: 5-point opinion intensity classification across product categories; (2) TripAdvisor Hotel Reviews [33]: 5-point opinion intensity annotations covering service quality and location; (3) SST-5 [34]: 5-tier opinion sentiment classification for movie reviews; and (4) SemEval-2017 Task 4C [35]: 5-point opinion intensity classification for Twitter data.

The dataset introduced in this paper differs significantly from the aforementioned datasets in two key aspects: (1) Topical focus: Our work centers on controversial topics discussed on the internet, whereas most existing datasets primarily consist of reviews related to products or services. (2) Temporal continuity: Our dataset captures time-continuous discussions on specific topics, while the others lack this temporal dimension. Moreover, to support more precise analysis of public opinion, our sentiment annotation adopts the intensity-graded scheme, spanning from *very negative* to *very positive*.

3. Dataset Collection and Processing

3.1. Dataset Collection Method and Overview

To analyze opinion trends from two perspectives, we collected datasets on two topics as follows:

- **Artificial general intelligence (AGI)**: Captures relatively long-term, gradual opinion changes on an evolving technological topic.
- **Hamas–Israel Conflict (HIC)**: Focuses on rapid opinion shifts following sudden news events with intense public discourse.

Both datasets were collected using the Twitter API and Python 3.10 scripts from the Twitter platform (https://x.com/home (accessed on 1 May 2025)), employing hashtag- and keyword-based queries.

The **AGI dataset** (https://github.com/BIMSA-DATA/FGPO/blob/main/AGI.xlsx (accessed on 1 May 2025)) contains 53,649 tweets related to artificial general intelligence (AGI), collected over a 32-month period from 1 October 2022 to 1 May 2025. The data was retrieved using a combination of relevant hashtags (e.g., "#ArtificialGeneralIntelligence") and keywords (e.g., "Artificial General Intelligence", "Strong Artificial Intelligence", "Strong AI").

The **HIC dataset** (https://github.com/BIMSA-DATA/FGPO/tree/main/HIC (accessed on 1 November 2023)) contains 608,814 tweets related to the Hamas–Israel conflict that began on 7 October 2023. Data was collected from 6 October to 31 October 2023, using hashtags such as "#Gaza", "#Israel", "#Hamas", "#FreePalestine", "#SaveGaza", "#PrayForGaza", "#IsraelUnderAttack", and "#StopTheViolence", along with keywords including "Hamas", "Israel Defense Forces (IDF)", "Palestinian Authority", "Gaza", and "West Bank".

An example tweet entry is shown in Table 2, with key data fields detailed below.

Table 2. Key attributes of a tweet entry.

	Attribute	Example Value
Tweet Details	tweet_ID	AGI_TWID_00012983
	tweet_pub_time	2022-12-13 21:02:17
	tweet_text	Imagine the year 2035, and human-level artificial general intelligence (AGI) exists. What's one question you would ask it?
	tweet_replies	213
	tweet_reposts	43
	tweet_likes	247
Account Details	account_ID	AGI_ACID_00000837
	followers_num	257,458
	following_num	1706
	tweets_num	12,486

Dataset Attributes:

- tweet_ID: A unique identifier for each tweet. The original Twitter ID has been anonymized in our dataset.
- tweet_pub_time: The timestamp when the tweet was posted.
- tweet_text: The tweet's content, used for opinion analysis.
- tweet_replies, tweet_reposts, tweet_likes (please note that the recorded values for tweet_replies, tweet_ reposts, tweet_likes, followers_num, following_num, and tweets_num represent snapshots at the time of data collection and may have changed since then): The number of replies, reposts, and likes received by the tweet.
- account_ID: A unique identifier for the tweet's author. Original Twitter account IDs have been anonymized.
- followers_num, following_num Section 3.1: The number of followers and accounts followed by the user.
- tweets_num Section 3.1: The total number of tweets posted by the account.

3.2. Dataset Preprocessing

To ensure data quality and usability, we applied the following preprocessing steps to the raw data collected from Twitter.

3.2.1. Handling Missing Information

To maintain data integrity, we implemented an automated re-capturing process for missing values. If a data field was empty upon collection, we attempted to retrieve the data up to 10 times. As a result, entries with missing information account for less than 0.1% of the dataset, ensuring that data completeness does not significantly impact the overall analysis of opinion dynamics.

3.2.2. Data Filtering

To preserve the integrity of discussion data on a given topic, we did not artificially remove entries, except for those containing empty or garbled tweets. This approach ensures a more accurate analysis of public opinion and perspective shifts.

Additionally, we did not filter tweets based on language, allowing the dataset to retain its **multilingual features**. Emoticons, emojis, hashtags, and other common social media symbols were also preserved, as these elements can provide valuable context for opinion analysis using LLMs.

For subsequent opinion analysis, we removed URLs from tweet text prior to annotation. URLs were excluded because they rarely provide direct, in-text opinion signals that are accessible to LLMs, and they cannot be reliably crawled or previewed at scale. By removing URLs, we ensured that sentiment judgments were based primarily on the explicit and interpretable content of the tweet itself, thereby minimizing potential noise introduced by external links.

Additionally, to handle excessively long texts, we implemented a truncation strategy by retaining only the first 3000 tokens of each tweet. This threshold was chosen as it is sufficient for accurate opinion classification, while also reducing computational costs during analysis.

3.2.3. Privacy Anonymization

To standardize identifiers and protect user privacy, all Twitter tweet IDs were converted into custom dataset-specific *tweet_IDs*, as explained in Section 3.1.

While author information is valuable for in-depth research, such as evaluating an article's impact based on the author's influence (e.g., number of followers) or tracking opinion

shifts across multiple tweets, we ensured privacy protection by replacing original Twitter account IDs with custom identifiers *account_IDs*. However, we retained metadata such as the number of followers (*followers_num*), accounts followed (*following_num*), and total tweets (*tweets_num*) to support further analysis.

3.3. Sentiment Annotation for Tweets

To enable comprehensive and nuanced opinion analysis, we performed **fine-grained sentiment classification** on the collected tweets. The labeling scheme was tailored to the specific objectives of each dataset:

- For the **AGI dataset**, sentiment classification targeted the tweet author's **perceived feasibility of AGI realization**.
- For the **HIC dataset**, sentiment was evaluated separately toward the entities **"Hamas"** and **"Israel"**.

All tweets were annotated using a five-level sentiment scale: *very positive*, *mildly positive*, *neutral*, *mildly negative*, and *very negative*. An additional label, **NA** (not applicable), was assigned to tweets that did not express sentiment toward the specified target. While a more fine-grained sentiment scale (e.g., 7 or 9 categories) could, in theory, capture more nuanced opinions and potentially benefit certain applications, our pilot studies indicated that increasing the number of sentiment levels substantially reduced annotation reliability. The five-level scheme offered a better balance between granularity and annotation quality, resulting in higher inter-annotator agreement.

This task is inherently challenging due to its fine-grained nature and involves two primary dimensions of complexity:

1. It is a **multi-class classification problem**, with opinions categorized into five or more sentiment levels rather than the conventional ternary (positive/neutral/negative) scheme. This granularity supports more precise modeling of opinion dynamics.
2. It requires **aspect-specific opinion classification**, in which both the sentiment polarity and the specific topic or aspect being evaluated must be identified. This necessitates contextual understanding of the sentiment expression and its associated aspect, significantly increasing the difficulty of classification.

Given the large scale of the data, full manual annotation was infeasible. Furthermore, traditional sentiment analysis techniques lack the precision necessary for such fine-grained categorization. Recent advances have shown that **LLMs** excel at capturing nuanced sentiment [36]. For example, Flan-T5 has been applied to this task using prompt engineering techniques to achieve strong performance [37].

Accordingly, we designed an LLM-based strategy to automate sentiment labeling, employing GPT-4.1 (https://openai.com/index/gpt-4-1/ (accessed on 15 June 2025)) as the underlying model. To ensure alignment between LLM-generated labels and human judgment, we implemented two complementary procedures: (1) **Few-Shot In-Context Learning with Manual Annotations** (Section 3.3.1); and (2) **Iterative Refinement via Human-in-the-Loop Evaluation** (Section 3.3.2).

3.3.1. Few-Shot In-Context Learning with Manual Annotations

We employed **in-context learning (ICL)**, where the LLM is prompted with labeled example tweets to guide its sentiment classification. The demonstration pool consists of **1200 manually annotated tweets**, with **200 tweets per sentiment category**, including NA. All annotations were performed independently by three annotators. To minimize annotator bias, all annotators received detailed written guidelines and representative examples for each sentiment category, and were selected to represent diverse backgrounds. Annotation

was performed independently by the annotators, who were blinded to tweet metadata such as user identity and timestamp. Only examples with unanimous agreement were retained for the demonstration pool. For tweets with disagreement, annotators discussed the cases to resolve ambiguities and further refine the guidelines and criteria. If agreement could not be achieved, the tweets were discarded. The resulting consensus annotations provide clear classification references for LLMs.

For each tweet to be classified, **five of the most semantically similar** examples were retrieved from the demonstration pool using cosine similarity over SBERT embeddings [38]. These examples, along with the classification schema, were included in a unified prompt format. The final prompt template is shown in Figure 1.

Instruction: Analyze the sentiment expressed in the given **tweet** regarding {*target*} into one of the following categories:
- **Very Negative**: {*classification_criteria_1*}
- **Negative**: {*classification_criteria_2*}
- **Neutral**: {*classification_criteria_3*}
- **Positive**: {*classification_criteria_4*}
- **Very Positive**: {*classification_criteria_5*}
- **NA**: The target {*target*} is not mentioned.

Refer to the following examples:
1) Tweet text: {*tweet_text_1*}
   Sentiment: {*sentiment_1*}
2) Tweet text: {*tweet_text_2*}
   Sentiment: {*sentiment_2*}
3) Tweet text: {*tweet_text_3*}
   Sentiment: {*sentiment_3*}
4) Tweet text: {*tweet_text_4*}
   Sentiment: {*sentiment_4*}
5) Tweet text: {*tweet_text_5*}
   Sentiment: {*sentiment_5*}

Now analyze the following **tweet**:
{*tweet_text*}

Figure 1. Prompt template of in-context learning for fine-grained sentiment analysis.

In the prompt, italicized content enclosed in {...} indicates variable components. For instance, {*target*} refers to dataset-specific targets, such as "perceived feasibility of AGI realization" (for the AGI dataset) or "Hamas"/"Israel" (for the HIC dataset). The {*classification_criteria*} variables provide concise textual definitions for each sentiment category, which are manually derived from the annotated data. Five in-context examples are selected based on their semantic similarity to the input tweet. Representative classification criteria and example prompts are provided in Appendix A.1.

Because the HIC dataset includes two distinct targets, additional measures were implemented. Specifically, sentiment was evaluated independently for each entity ("Hamas" and "Israel") on a per-tweet basis. Annotators assigned separate sentiment labels for each target entity mentioned or implied in the tweet, following entity-specific classification guidelines (see Appendix A.1). For tweets expressing sentiment toward both entities, annotators evaluated and recorded sentiment independently for "Hamas" and "Israel." If a tweet did not mention or allude to one of the entities, the label "NA" (not applicable) was assigned for that target.

3.3.2. Iterative Refinement via Human-in-the-Loop Evaluation

Instead of finalizing the annotation strategy in a single pass, we adopted an iterative human-in-the-loop procedure to progressively refine both the prompt template and the ICL demonstration pool. The overall algorithmic workflow is illustrated in Figure 2.

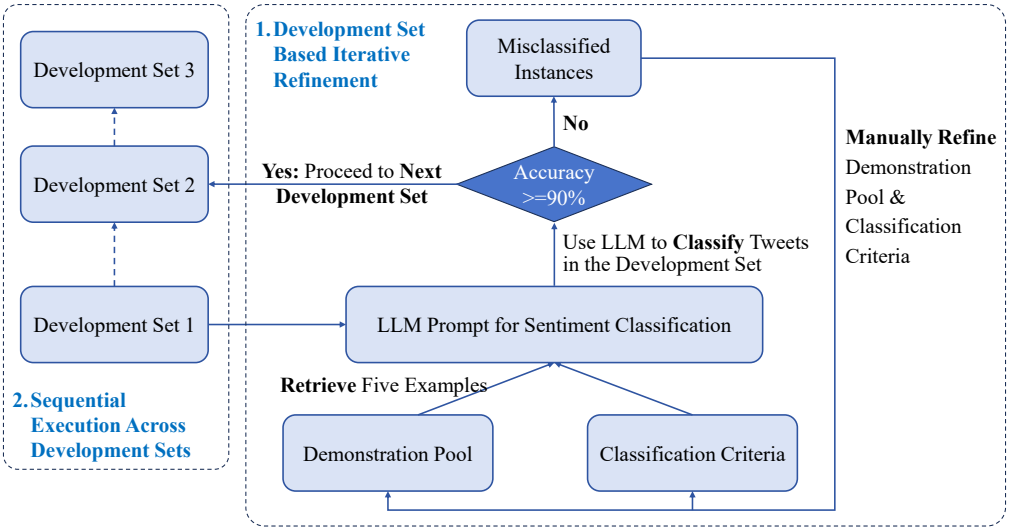


Figure 2. Iterative refinement via human-in-the-loop evaluation for sentiment annotation of tweets.

(1) Development Set–Based Iterative Refinement

Each iteration begins with the construction of a dedicated development set for evaluating the current prompt and demonstration configuration. Given the existing version of the demonstration pool and classification criteria, the following steps are carried out:

- The LLM, guided by the current prompt, is used to classify tweets in the development set.
- Misclassified instances are manually reviewed to identify issues in either the classification criteria or the representativeness of demonstration examples.
  - The classification criteria descriptions are refined to improve the delineation between sentiment categories.
  - For each misclassified tweet, three additional semantically similar tweets, retrieved via SBERT-based cosine similarity, are annotated and incorporated into the demonstration pool. The annotation procedure for the new tweets follows the same methodology as that used for the initial demonstration pool.

This refinement cycle is repeated until the model achieves at least **90% agreement** with human annotations on the current development set. Once this performance threshold is met, the evaluated development set is merged into the demonstration pool, thereby enhancing its coverage and diversity for subsequent iterations.

(2) Sequential Execution Across Development Sets

To enhance generalizability and robustness, the iterative refinement process is extended across multiple development sets, executed sequentially. Upon completing refinement on one development set, the updated prompt and demonstration pool are used as the basis for evaluating the next. This sequential design ensures that improvements—such as refined classification criteria and expanded demonstrations—accumulate over time and benefit subsequent iterations. The total number of sequentially executed development sets is denoted by λ_t, which is set to 3 in practice.

To evaluate the effectiveness of our sentiment annotation method, we conducted experiments using a subset of manually labeled data. The experimental details and results are detailed in Appendix A.2.

3.3.3. Subjectivity Filtering

Beyond sentiment classification, we also categorized tweets based on their subjectivity: **subjective, objective, or both**. This was essential for filtering out objective tweets in public opinion analysis, as they do not reflect personal attitudes.

Given the relatively straightforward nature of this classification task, we employed GPT-4.1 with a direct prompting approach, as detailed in Appendix A.3.

3.3.4. Example of Labeled Data

A sample annotation follows:

[subjective]; [AGI: very positive]

This label indicates that the tweet expresses a subjective opinion, with a very positive stance toward the "perceived feasibility of AGI realization".

4. Dataset Exploration

This section focuses on presenting an exploratory analysis of the constructed dataset, highlighting its key characteristics to provide a comprehensive overview. For illustrative purposes, in the AGI dataset, the sentiment subject is the "feasibility of AGI", while in the HIC dataset, it pertains to "Hamas".

4.1. Textual Analysis Based on Sentiment

For each dataset, we classified each instance into one of five sentiment categories based on the tweet's content, as outlined previously. As an initial step, we visualized the most frequently occurring keywords and phrases using word clouds for each dataset (see Figures 3 and 4). The creation of these visualizations involved the following steps:

(1) **Filtering by Tweet Subjectivity**: We removed instances labeled as "objective" and retained those marked as "subjective" or "both" to focus on tweets expressing personal opinions.

(2) **Stop-Word Removal**: Common words such as "and" and "is" often appear frequently across texts but do not contribute significant semantic meaning. We filtered out these stop-words from the word clouds. Additionally, we manually curated and removed high-frequency words that contribute little semantic value. For instance, in the AGI dataset, terms such as "AI" and "AGI" were excluded, while in the HIC dataset, words like "Hamas" and "Israel" were filtered out.

These steps ensure that the resulting word clouds effectively highlight the most relevant and distinctive terms within each dataset.



(**a**) (**b**)

Figure 3. Word clouds generated from tweets in the AGI dataset: (**a**) tweets labeled as "Very Positive" sentiment; (**b**) tweets labeled as "Very Negative" sentiment. The size of each word reflects its relative frequency within the corresponding sentiment class.



(**a**) (**b**)

Figure 4. Word clouds generated from tweets in the HIC dataset: (**a**) tweets labeled as "Very Positive" sentiment toward "Hamas"; (**b**) tweets labeled as "Very Negative" sentiment toward "Hamas". The size of each word reflects its relative frequency within the corresponding sentiment class.

Figure 3 showcases word clouds of tweets in the AGI dataset reflecting different attitudes toward AGI. Specifically, Figure 3a illustrates optimism about AGI's potential, with words such as "will", "future", and "advanced". On the other hand, Figure 3b reflects skepticism about the realization of AGI, with terms like "impossible" and "never", suggesting that its achievement is nearly unfeasible.

Figure 4 displays word clouds of tweets in the HIC dataset representing two distinct sentiments toward "Hamas": "very positive" and "very negative,". In Figure 4a, the word cloud evokes a more humanistic tone with words like "people." In contrast, Figure 4b emphasizes negative sentiment, with terms like "terrorist."

Additional word clouds representing other sentiments are provided in Appendix B.

4.2. Opinion Evolution Analysis Based on Sentiment Labels

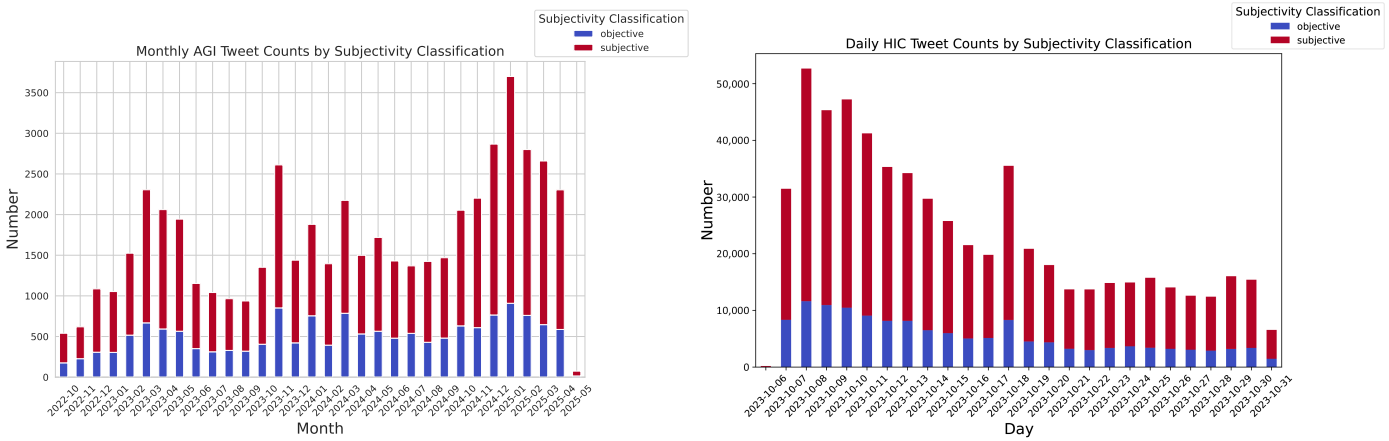
We propose that the volume of public discussion can serve as an indicator of the public impact of an event. Figure 5 illustrates the temporal evolution of tweet subjectivity related to each topic, revealing several key patterns (in the following sections, tweets labeled as "both" or "subjective" are collectively referred to as "subjective"):

(1) Subjective tweets generally outnumber objective tweets, indicating that personal opinions are more frequently expressed in the discussion.

(2) The trends in objective and subjective tweet counts follow similar temporal patterns. Increases in objective tweet volume are often accompanied by corresponding rises in subjective tweet volume, and vice versa.

(3) Sudden surges in objective tweets, typically driven by news reports, seem to trigger even larger increases in subjective tweets. For instance, in the AGI dataset, spikes in objective tweets from 2022-11 to 2022-12, 2023-02 to 2023-03, and 2024-11 to 2025-01 were followed by substantial increases in subjective tweets. Similar patterns are observed in the HIC dataset during 2023-10-06 to 2023-10-07, and 2023-10-16 to 2023-10-17. Additionally, while the volume of objective tweets tends to decline over time, this decrease is more gradual compared to the faster decline observed in subjective tweet volume (e.g., from 2023-03 to 2023-06 in AGI and from 2023-10-10 to 2023-10-16 in HIC).

(4) In both datasets, tweet volumes frequently exhibit a rapid rise followed by a sharp decline, reflecting the "explosive" nature of certain public opinion trends, described by Sun et al. [39]. For instance, in Figure 5a, AGI tweet counts surge between 2023-09 and 2023-11, and again between 2024-11 and 2025-01. Similarly, in Figure 5b, HIC tweet counts spike between 2023-10-16 and 2023-10-18. Furthermore, the pronounced decrease in tweet volume observed between 2023-10-31 and 2023-11-06 in Figure 5b is consistent with the "annealing effect" on social media, as discussed by Volchenkov and Putkaradze [40].

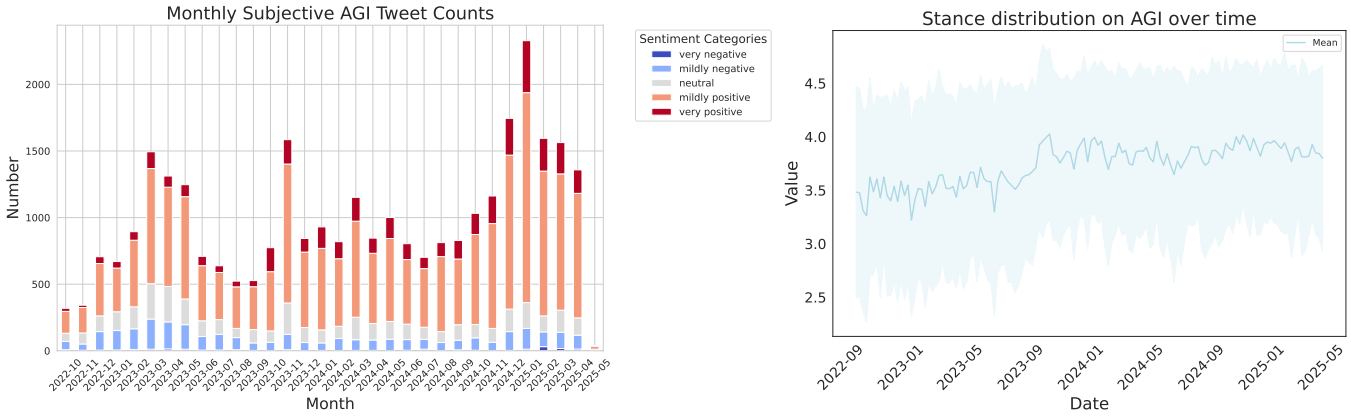


(**a**) (**b**)

Figure 5. Temporal distribution of subjectivity classification in two datasets: (**a**) AGI dataset; (**b**) HIC dataset.

Next, we examine potential shifts in public opinion following the onset of significant events.

Figure 6 illustrates the evolution of public opinion on the feasibility of AGI. Figure 6a presents the distribution of tweet counts across five sentiment categories, while Figure 6b tracks the temporal changes in mean public opinion and its standard deviation from September 2022 to May 2025.



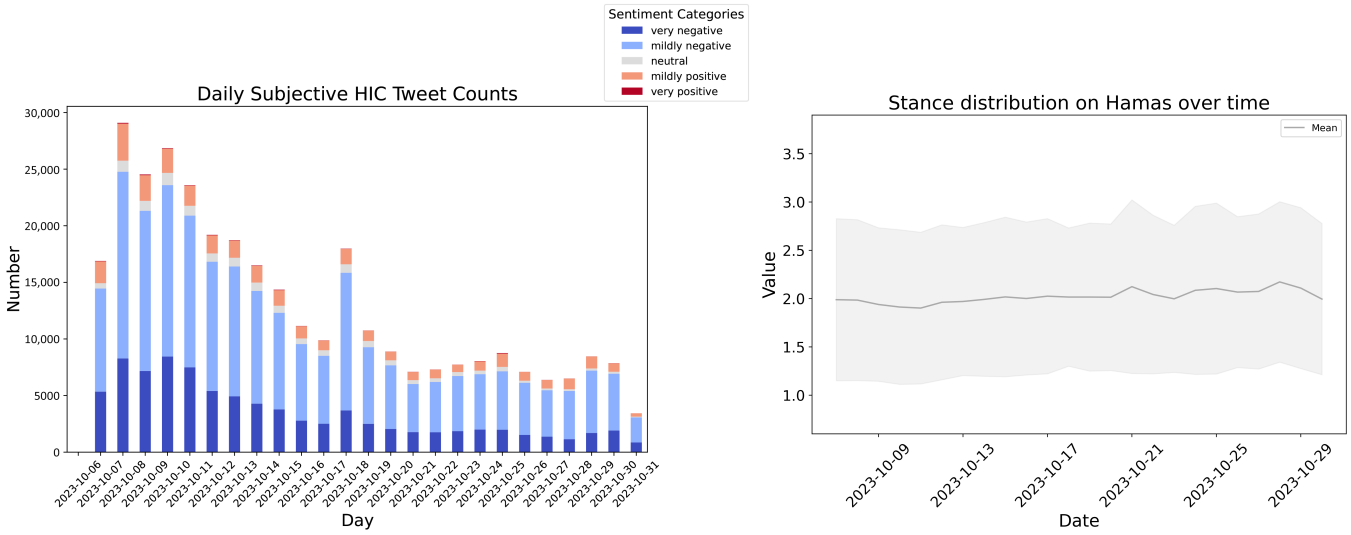
(**a**) (**b**)

Figure 6. Evolution of public opinion on "the feasibility of AGI" in the AGI dataset: (**a**) temporal distribution of opinions; (**b**) mean opinion shift. The solid line indicates the average public stance, and the shaded area represents one standard deviation above and below the mean.

As shown in Figure 6a, more than half of the tweets convey an optimistic outlook on AGI realization. In Figure 6b, sentiment categories are mapped to integer scores, where 1 represents the highest level of pessimism and 5 represents the greatest optimism. A sharp increase in optimism is observed following the release of ChatGPT in late November 2022. While the mean stance fluctuates over time, it maintains a mild ongoing trend and generally stays slightly above 3, suggesting a sustained moderate optimism regarding AGI feasibility throughout the period.

Figure 7 illustrates the evolution of sentiments surrounding the Hamas–Israel conflict. Specifically, Figure 7a depicts the distribution of tweet counts across five sentiment cate-

gories, while Figure 7b presents the mean stance value along with its standard deviation over a 25-day period following the initial attack.



(**a**) (**b**)

Figure 7. Evolution of public opinion toward "Hamas" in the HIC dataset: (**a**) temporal distribution of opinions; (**b**) mean opinion shift. The solid line indicates the average public stance, and the shaded area represents one standard deviation above and below the mean.

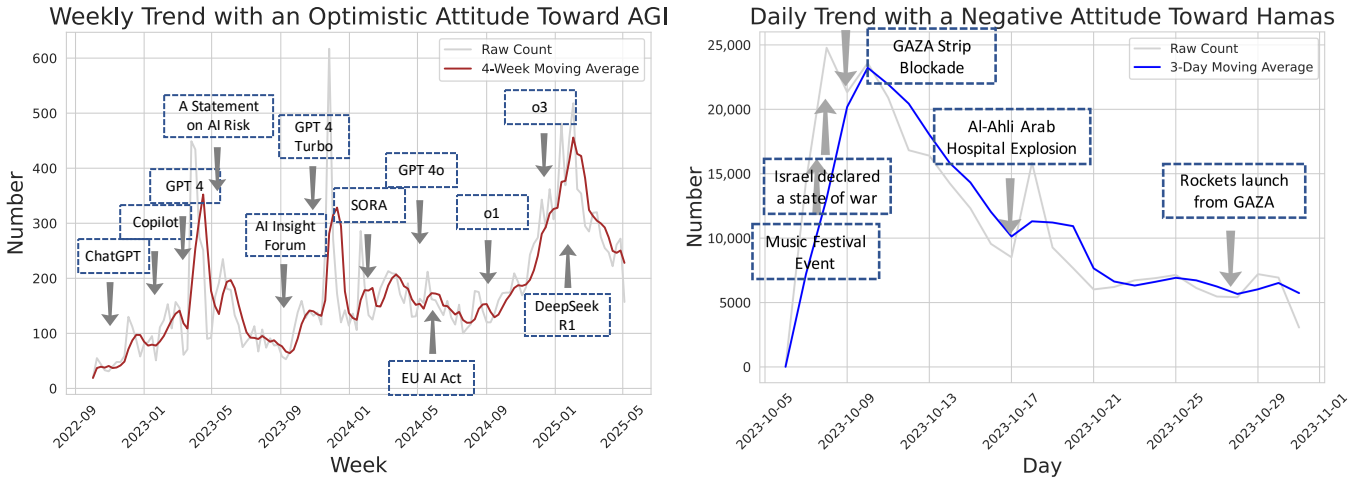
In Figure 7a, a substantial portion of tweets express negative sentiment. In Figure 7b, sentiment categories are mapped to integer scores from 1 to 5, where 1 represents a "very negative" stance and 5 represents a "very positive" stance. A clear downward trend in the mean stance is observed during the first week after the attack, reflecting a rise in negative sentiment, with occasional minor rebounds followed by rapid declines. Overall, the mean stance stabilizes around 2, indicating a predominantly negative sentiment.

In addition, Figures 6a and 7a reveal that the proportion of tweets expressing neutral opinions is minimal, while the shares of "very positive" (in Figure 6a) and "very negative" (in Figure 7a) sentiments are notably high. This distribution highlights pronounced political polarization, consistent with the findings reported by Galam [6]. Furthermore, the sharp rises and falls in the average sentiment values observed in Figures 6b and 7b closely mirror the rapid shifts in collective opinion states reported by Qian et al. [41]. These patterns underscore the dynamic and polarized nature of public discourse on both topics.

4.3. Impact of Major News Events

Building on our previous analysis, we hypothesize that news coverage plays a crucial role in shaping public engagement. To explore this, we compiled date information for key related events and examined their potential correlation with shifts in public sentiment.

Figure 8a illustrates the temporal evolution of opinions on AGI feasibility within the AGI dataset. To capture overall optimism, we aggregated tweets labeled as "very positive" and "positive". Notable surges in optimistic sentiment were observed following the introduction of competitive LLMs, such as ChatGPT (GPT-3.5) GPT-4 Turbo, o3, and DeepSeek R1. Significant events that may have influenced these fluctuations are annotated along the timeline.



(**a**) (**b**)

Figure 8. Temporal evolution of opinions with key events marked in two datasets: (**a**) weekly trend of optimistic attitudes toward AGI with key event annotations; (**b**) daily trend of negative attitudes toward Hamas with key event annotations. The gray arrows indicate the corresponding event dates. Note that the moving average window differs between the two subfigures to accommodate the distinct temporal characteristics of each dataset: a 4-week moving average is applied to the AGI dataset (subfigure (**a**)) due to its longer time span, while a 3-day moving average is used for the HIC dataset (subfigure (**b**)) to reflect its shorter duration and higher temporal resolution. These choices were made to enhance the clarity of the visualized trends and do not influence the interpretation of the underlying attitude dynamics.

Figure 8b illustrates the trajectory of negative sentiment in the HIC dataset. In this figure, we aggregate tweets classified as "very negative" and "mildly negative" toward Hamas and track their daily counts. The data reveal a sharp increase in negative tweets starting on 6 October 2023, followed by a gradual decline until 17 October 2023, after which a renewed upward trend emerges. Similar to Figure 8a, key events are annotated along the timeline for contextual reference.

These findings indicate that major news events may directly influence public sentiment, with media coverage potentially shaping shifts in public opinion. To move beyond these descriptive observations and rigorously assess the causal impact of major news events, we next turn to a quantitative approach. Specifically, we employ an Interrupted Time-Series (ITS) analysis, which enables us to decompose an event's effect into two core components: an immediate level shift (measured by the β_2 coefficient) and a sustained change in the post-event trend (measured by the β_3 coefficient). To illustrate this approach, we examine two representative cases: the release of GPT-4 (https://en.wikipedia.org/wiki/GPT-4 (accessed on 1 June 2025)) in the AGI dataset and the Al-Ahli Arab Hospital explosion (https://en.wikipedia.org/wiki/Al-Ahli_Arab_Hospital_explosion (accessed on 15 November 2023)) in the HIC dataset. For each event, we model three distinct metrics: the count of positive posts, the count of negative posts, and the proportion of positivity (defined as the count of positive posts divided by the sum of positive and negative posts). The full results are presented in Table 3, and the following discussion focuses on the core findings that were statistically significant ($p < 0.05$).

For the GPT-4 release event in the AGI dataset, we employed an ITS analysis. To mitigate confounding from other potential events, we defined a minimized effective analysis window spanning 60 days surrounding the event (14 February–14 April 2023), with the intervention effects measured starting from 15 March. We modeled the three core metrics,

using Negative Binomial regression to analyze the counts of positive and negative posts due to their nature as count data. A binary variable was included in the model to control for weekend versus weekday differences.

Table 3. Interrupted Time-Series (ITS) analysis for two representative events.

Variable	AGI			HIC		
	Model 1: Positives (NB)	**Model 2: Negatives (NB)**	**Model 3: Proportion (OLS)**	**Model 1: Positives (NB)**	**Model 2: Negatives (NB)**	**Model 3: Proportion (OLS)**
Intercept	3.599 (0.375) ***	2.390 (0.392) ***	0.763 (0.047) ***	4.406 (0.334) ***	6.712 (0.331) ***	0.091 (0.005) ***
time	−0.045 (0.023)	−0.043 (0.024)	−0.001 (0.005)	−0.006 (0.014)	−0.005 (0.014)	−0.000 (0.000)
event (β_2)	1.618 (0.531) **	1.103 (0.564)	0.052 (0.098)	0.394 (0.487)	0.324 (0.483)	0.006 (0.008)
post_event (β_3)	0.049 (0.031)	0.060 (0.032)	0.002 (0.005)	0.006 (0.024)	0.022 (0.024)	−0.001 (0.000) **

Note: Standard errors are in parentheses. ** $p < 0.01$, *** $p < 0.001$.

The results indicate that the event had a significant, immediate impact on discussion volume. On the day following the event, the count of positive posts surged in a statistically highly significant manner ($\beta_2 = 1.618$, $p = 0.002$). The increase in negative posts was not statistically significant, nor was there a significant change in the positivity proportion.

For the Al-Ahli Arab Hospital explosion in the HIC dataset, we also conducted an ITS analysis. Similarly, to mitigate confounding from other potential events, we selected a minimized analysis window that was effective for model estimation: a 72 h period surrounding the event (16–18 October 2025), with the intervention at 17 October, 16:00. We modeled three metrics: the counts of positive and negative posts, and the positivity proportion. For the positivity proportion series, which yielded the core finding, we used an OLS regression with HAC robust standard errors and controlled for the time of day (grouped into Night, Morning, Afternoon, and Evening) to account for periodic fluctuations.

Among the three models, we found no statistically significant immediate shock or trend change in the absolute counts of positive or negative posts. The most significant impact, however, was on the relative balance of sentiment. The analysis revealed that the long-term trend of the positivity proportion experienced a significant negative reversal after the event ($\beta_3 = -0.001$, $p = 0.003$). This result suggests that the event was not merely a one-time emotional shock, but rather a key turning point that initiated a persistent trajectory of deteriorating public mood.

4.4. Discussion on Applications of the Dataset

The temporal analyses in Sections 4.2 and 4.3 serve as illustrative examples, showcasing the dataset's potential to support longitudinal studies on the evolution of public opinion. By publishing this dataset, we aim to foster a wide range of research directions. Below, we outline several key applications enabled by the dataset's richness in tweet metadata, user attributes, interaction information, and fine-grained sentiment annotations.

(1) **Fine-Grained Sentiment Analysis:** Unlike most existing textual datasets, which typically limit sentiment to binary or ternary categories, our dataset provides five distinct sentiment levels. This granularity, combined with the topical focus of each dataset, reduces confounding factors and enables more precise sentiment modeling. Researchers can thus explore nuanced emotional expressions and subtle shifts in public mood with greater fidelity.

(2) **Stance Detection:** Stance detection aims to determine whether an author is *in favor of*, *against*, or *neutral toward* a specific target [42]. Our dataset is well-suited for this task in several respects. First, the five-level sentiment annotations offer a graded perspective on users' attitudes toward central topics, supporting fine-grained stance tracking over time. Second, the dataset's temporal breadth and scale (hundreds of thousands of

annotated entries) enable longitudinal analyses of stance dynamics at both individual and population levels, including the study of evolving alignments and polarization. Third, the inclusion of user identifiers and timestamps facilitates the construction of interaction networks, which are essential for structural or embedding-based stance detection frameworks (e.g., STEM [43]).Collectively, these features make the dataset a valuable resource for advancing both supervised and unsupervised stance detection methodologies in the context of social media discourse.

(3) **Applications in Computational Social Science:** Our dataset offers substantial value for computational social science research by combining textual content, temporal metadata, and user-level attributes. Key application areas include:

- **Opinion Dynamics and Temporal Shifts:** Public sentiment evolves in response to both internal discourse and external events. Our dataset provides annotated category and stance labels for tweets, along with anonymized user IDs and timestamps. This enables researchers to construct temporal sentiment trajectories, analyze how public opinion fluctuates over time, and examine responses to major events. Such analyses can offer insight into the onset of opinion shifts and inform strategies to mitigate issues like social polarization or emotional escalation during crises.
- **Social Behavior and Expression Patterns:** Individuals express opinions based on their perspectives, social roles, and habitual communication styles. Our dataset allows investigation of how users convey stance or subjectivity in relation to their social context. For example, researchers can explore whether users with larger follower counts are more likely to express strong opinions or subjective views, and whether certain expression patterns correspond with specific stance categories. These insights can contribute to understanding patterns of opinion expression, identity signaling, and social influence in digital discourse.

5. Conclusions

In conclusion, the dataset presented in this paper provides a novel and nuanced resource for analyzing public sentiment and tracking opinion shifts over time. By addressing key limitations of existing datasets, such as limited scale, narrow topical coverage, and coarse-grained sentiment labels, we offer a more detailed and contextually rich foundation for studying complex social phenomena. By focusing on two highly salient and polarized topics, the perceived feasibility of artificial general intelligence (AGI) and the Hamas–Israel conflict (HIC), our dataset captures both the diversity of public sentiment and the temporal dynamics that characterize modern digital discourse.

We carefully curated and processed the AGI and HIC datasets, implementing fine-grained sentiment annotations to support robust analyses of public opinion. Our findings reveal several distinctive patterns in opinion evolution, particularly the pronounced influence of major events, underscoring the need for models that can capture such real-time shifts.

Ultimately, this dataset holds broad potential to advance research in sentiment analysis, opinion dynamics, computational social science, and AI ethics. By enabling empirical studies of opinion diffusion and supporting the development of predictive models, our work contributes to a deeper and more accurate understanding of public sentiment in the digital age.

6. Limitations

Although we have highlighted the richness of our dataset, the rigor of our processing pipeline, and its potential applications, several limitations should be acknowledged.

(a) **Potential inaccuracies in sentiment annotation.** We employed an LLM to perform fine-grained sentiment annotation of tweets. Inevitably, discrepancies remain compared to full human annotation. Given the scale of the dataset, complete manual annotation is infeasible; instead, we focus on improving model accuracy, which remains an important direction for future work. Moreover, as described in Section 3.2.2, we removed URLs from tweet text prior to annotation to concentrate on the textual content for subsequent opinion analysis. While this approach simplifies annotation and modeling, it introduces a limitation: URLs often provide crucial context, particularly in political discourse. Excluding them risks losing interpretive cues, which may lead to misclassification or less accurate sentiment labeling. Future work could explore strategies for incorporating external content referenced by URLs, such as linked previews, metadata, or content retrieval, to enhance annotation accuracy and contextual depth.

(b) **Fixed sentiment granularity.** Our annotation scheme divides sentiment into five discrete categories, which we regard as a key contribution. However, the choice of five levels is ultimately heuristic, and it is unclear whether this granularity is optimal across all applications. Some tasks may benefit from coarser or finer sentiment distinctions (e.g., seven or nine categories), or even continuous sentiment scores. Fixed granularity may therefore limit the dataset's adaptability for use cases requiring more nuanced sentiment differentiation. Future work could investigate adaptive or context-sensitive sentiment scales, potentially informed by external signals or annotator confidence, to support finer distinctions where reliability permits. Comparative evaluations of different levels of sentiment granularity on downstream tasks would also be valuable.

(c) **Limited scope of temporal analysis.** Although the temporal analyses in Sections 4.2 and 4.3 highlight the dataset's potential for longitudinal research, their scope is currently constrained. Our study primarily focuses on aggregate patterns, such as mean opinion shifts, tweet volume correlations, and Interrupted Time-Series (ITS) analyses of the impact of key events on public sentiment, rather than conducting systematic, event-driven, or fine-grained longitudinal investigations. While we provide illustrative cases of sentiment responses to major real-world events (e.g., the release of ChatGPT and a large-scale attack, as depicted in Figures 6–8), these are intended as demonstrations rather than exhaustive analyses. A more comprehensive longitudinal approach could systematically track sentiment dynamics across a wider array of events, evaluate the influence of various external factors (such as media coverage, policy changes, or social movements), and monitor the evolution of divergent viewpoints and user communities over extended periods.

In summary, while our dataset provides a valuable resource for opinion and sentiment analysis, addressing these limitations, particularly with respect to context preservation, sentiment granularity, and comprehensive temporal analysis, will be essential for maximizing its utility and impact in future research.

7. Ethical Statement

All data used in this study were collected from publicly available posts on Twitter/X in accordance with the platform's terms of service. No private or direct message content was accessed. To protect user privacy, we anonymized all account identifiers and removed personally identifying information during preprocessing. The dataset is released in a form that preserves analytical utility while minimizing risks of re-identification.

For sentiment annotation, we employed LLMs exclusively as a data processing tool. Importantly, LLMs were not used to generate or create any new data, content, or synthetic

records. Their sole role was to assist in labeling existing tweets with sentiment categories. All subsequent analyses were conducted on this annotated dataset.

Our study is designed to advance the understanding of public opinion dynamics in an aggregated and responsible manner. We report results only at the collective level and do not make claims about individual users. The dataset and analyses are intended for research purposes only, and care has been taken to avoid stigmatization or harm to any specific individuals or communities.

Author Contributions: Conceptualization, H.X. and M.H.; methodology, H.X. and M.H.; software, H.X. and M.H.; validation, H.X. and M.H.; formal analysis, H.X. and M.H.; investigation, H.X. and M.H.; resources, H.X. and M.H.; data curation, H.X. and M.H.; writing—original draft, H.X.; writing—review and editing, H.X. and M.H.; visualization, H.X. and M.H.; supervision, H.X. and M.H.; project administration, H.X.; funding acquisition, H.X. and M.H. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by Beijing Institute of Mathematical Sciences and Applications (BIMSA), Beijing, China.

Data Availability Statement: The data presented in this study are openly available in GitHub at https://github.com/BIMSA-DATA/FGPO/tree/main (accessed on 1 June 2025).

Acknowledgments: We are grateful to our colleagues at the Beijing Institute of Mathematical Sciences and Applications (BIMSA) for their valuable guidance during the processes of data collection, processing, and analysis. We also sincerely appreciate the insightful feedback and constructive suggestions provided by the anonymous reviewers.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Details and Results of Sentiment Annotation of Twitter Data

Appendix A.1. Example Classification Criteria

In this section, we present the classification criteria ultimately used as the basis for sentiment classification in our LLM prompts, specifically those labeled as {*classic_criteria_**} in Figure 1. Given the large number of categories, we illustrate only a subset to help readers understand the rationale behind the classification.

1. For the AGI dataset

The classification target is the perceived feasibility of AGI realization. As examples, we present the criteria for the categories **Very Positive**, **Mildly Positive**, and **Neutral**.

(1) **Very Positive**

- **Tone**: Enthusiastic, confident, visionary.
- **Grammar/Syntax**: Strong imperative and declarative forms; exclamations; metaphorical and poetic language.
- **Lexical Markers**: Assertive and confident language (e.g., “cannot be stopped”, “is happening now”, “immediately deployed”), superlatives and absolutes (e.g., “revolutionize”, “drastically”, “civilisation altering”), direct calls to action or celebration (e.g., “get in the game”, “must learn AGI”).
- **Semantic Indicators**:
  - AGI framed as transformative and beneficial.
  - Implicit or explicit belief in AGI’s rapid or inevitable success.
  - Often invokes large-scale societal impact.

(2) **Mildly Positive**

- **Tone**: Hopeful, cautiously optimistic.
- **Grammar/Syntax**: Balanced and explanatory.
- **Lexical Markers**: Modal verbs (e.g., could, might), hedging expressions (e.g., "could reasonably be viewed as", "expected to see it", "may arrive").
- **Semantic Indicators**:
  - Belief in AGI's feasibility, but acknowledgment of risks or uncertainty.
  - Often includes quotes or summaries of expert opinions.
  - Emphasis on ethical development or regulatory needs.

(3) **Neutral**

- **Tone**: Informative, speculative, ambiguous.
- **Grammar/Syntax**: Interrogative or conditional forms; open-ended phrasing.
- **Lexical Markers**: Objective or questioning language (e.g., "difficult to predict", "unclear when, or even if"), descriptive or definitional.
- **Semantic Indicators**:
  - Describes AGI without expressing judgment.
  - Discusses timelines, definitions, or open questions.
  - Includes doubts without commitment to feasibility or failure.

2. HIC Dataset Classification Criteria

The classification target is either Hamas or Israel. To complement the earlier examples from the AGI dataset, we provide below the criteria for the **Mildly Negative** and **Very Negative** categories as illustrative examples.

For the target "Hamas," the classification criteria are as follows:

(1) **Mildly Negative**

- **Grammatical/Syntactic Cues**: (1) more declarative and restrained in tone; (2) irony or rhetorical questioning used to express disagreement.
- **Lexical and Phrase Indicators**: (1) criticism is present but less emotionally intense (e.g., "can't be trusted", "sympathizer", "lied", "dangerous", "problematic"); (2) political sarcasm, skepticism, or accusation (e.g., "Who believes she would say something like this?").
- **Semantic Features:**
  - Critique of Hamas is often embedded in broader political commentary (e.g., on U.S. or Israeli policy).
  - Some recognition of complexity, but negative view of Hamas remains dominant.
  - Less visceral than "very negative," often focused on strategic or political implications.

(2) **Very Negative**

- **Grammatical/Syntactic Cues**: (1) frequent use of exclamatory sentences, imperative tone; (2) use of all caps and hashtags emphasizing condemnation (e.g., #HamasTerrorists, #IStandWithIsrael); (3) direct address (e.g., rhetorical questions to political figures demanding condemnation).
- **Lexical & Phrase Indicators**: (1) strong, graphic, emotionally charged words (e.g., "terrorists", "savages", "monsters", "murdered", "beheaded", "burned alive", "raped"); (2) use of explicit dehumanization or demonization (e.g., "evil", "heartless", "inhumane", "disgusting"); (3) calls for violent retribution (e.g., "must be eliminated", "wish death", "destroy Hamas").
- **Semantic Features:**

  - Total moral condemnation; Hamas framed as purely evil, non-human, beyond redemption.
  - Zero tolerance or nuance; no mention of context, motives, or broader conflict.
  - Often includes graphic or disturbing detail to shock or invoke rage.

For the target "Israel," the classification criteria are as follows:

(1) **Mildly Negative**

- **Grammatical/Syntactic Cues**: (1) statements often include "but" or "while" clauses, expressing ambivalence or conditional support; (2) emotional appeals focused on empathy for Palestinian civilians.
- **Lexical and Phrase Indicators**: (1) supportive but measured terms: "resistance", "support", "solidarity", "Palestine must defend itself"; (2) may critique Israel more than praise Hamas directly.
- **Semantic Features:**
  - Support for Palestinian cause may include implicit or indirect defense of Hamas actions.
  - Framing Hamas as a response to Israeli oppression rather than an aggressor.
  - Often emphasizes ceasefire, peace, or opposition to violence—without full endorsement of Hamas.

(2) **Very Negative**

- **Grammatical/Syntactic Cues**: (1) exclamatory, celebratory, or prayer-based tone; (2) hashtags praising Hamas or calling for Israel's destruction.
- **Lexical and Phrase Indicators**: (1) explicit glorification of Hamas (e.g., "heroes", "liberators", "bravery", "job well done", "long live Hamas"); (2) derogatory terms for Israel (e.g., "terrorist state", "oppressor", "satanic pilots").
- **Semantic Features:**
  - Hamas seen as freedom fighters or divine agents.
  - Often includes calls for resistance, revolution, or the fall of Israel.
  - May assert that violence is justified or even holy.

Appendix A.2. Details and Results of the Verification Experiment for the Sentiment Annotation Method

To validate the effectiveness of our sentiment annotation method, we manually annotated 300 Twitter data entries (50 per category) to serve as a test set. We also designed several comparative and ablation methods as follows:

- **Zero-Shot:** prompts contained only the task description and classification criteria, with no in-context learning (ICL) demonstrations.
- **No-CC:** prompts omitted the classification criteria.
- **No-IR:** the iterative refinement step involving human-in-the-loop evaluation was excluded from the prompt design.
- **Varying Number of Development Sets:** in sequential execution across development sets, we varied the number of development sets (i.e., the value of λ_t) to 1, 2, and 3, respectively.

GPT-4.1 was used as the large language model (LLM), accessed via API. We selected GPT-4.1 for its strong performance on sentiment analysis tasks (https://llm-stats.com/models/gpt-4.1-2025-04-14 (accessed on 1 June 2025)) and its relatively low API cost. Given the need to process hundreds of thousands of data entries, cost-effectiveness was an important consideration. While recent studies have demonstrated the effectiveness

of various LLMs for sentiment analysis (e.g., Flan-T5 [37]), we chose GPT-4.1 after both empirical evaluation and practical assessment of several candidate models. In preliminary experiments, we also evaluated other LLMs, including Flan-T5 (XXL) (https://huggingface.co/google/flan-t5-xxl (accessed on 1 June 2025)), GPT-o3-mini (https://openai.com/index/openai-o3-mini/ (accessed on 1 June 2025)), and Qwen3 (https://qwenlm.github.io/blog/qwen3/ (accessed on 1 June 2025)), which have also shown good performance in sentiment analysis and relatively low API costs (https://llm-stats.com/ (accessed on 1 June 2025)). All experiments were conducted using identical prompt templates, demonstration pools, and experimental setting ($\lambda_t = 3$) on the same 300-entry Twitter test set. To ensure deterministic and reproducible outputs, the LLM temperature was fixed at 0 across all experiments.

The experimental results of the comparative methods, ablation studies, and model comparisons are presented in Table A1, with the best-performing results highlighted in bold.

Table A1. Experimental results of sentiment classification on the AGI and HIC test datasets.

Method	AGI Test Result			HIC Test Result		
	Macro-Pre.	**Macro-Rec.**	**Macro-F1**	**Macro-Pre.**	**Macro-Rec.**	**Macro-F1**
Zero-Shot	0.8461	0.8417	0.8425	0.8638	0.8625	0.8629
No-CC	0.8853	0.8833	0.8837	0.8935	0.8917	0.8920
NO-IR	0.8885	0.8875	0.8877	0.8978	0.8958	0.8961
$\lambda_t = 1$	0.9077	0.9042	0.9048	0.9232	0.9208	0.9214
$\lambda_t = 2$	0.9119	0.9102	0.9110	0.9297	0.9281	0.9289
$\lambda_t = 3$	**0.9131**	**0.9125**	**0.9126**	**0.9314**	**0.9292**	**0.9297**
Flan-T5 (XXL)	0.8723	0.8691	0.8707	0.8880	0.8845	0.8862
GPT-o3-mini	0.8826	0.8794	0.8810	0.8955	0.8920	0.8937
Qwen3	0.8894	0.8861	0.8877	0.9097	0.9060	0.9078

Note: Macro-Pre. and Macro-Rec. represent macro-precision and macro-recall, respectively.

Based on the experimental results, several key observations are summarized below:

(a) Among all evaluated LLMs, GPT-4.1 achieved the highest performance and the closest alignment with human annotations. Consequently, we selected GPT-4.1 as the final annotation model for our study.

(b) The sentence classification method we ultimately selected—where $\lambda_t = 3$, incorporating ICL demonstrations and classification criteria optimized through three rounds of development set iterations—achieved the best performance on the test set. This demonstrates the effectiveness of our ICL-based strategy combined with iterative development set refinement.

(c) Increasing the number of development sets generally improves performance. However, the performance gain from $\lambda_t = 2$ to $\lambda_t = 3$ is relatively modest. Therefore, we chose $\lambda_t = 3$ as a practical trade-off between performance and efficiency.

(d) The Zero-Shot approach performed significantly worse than all other methods, highlighting the importance of including ICL demonstrations for improving sentiment classification accuracy.

The macro-F1 score of approximately 90% indicates that the sentiment classification model performs reliably. A natural question arises: does this level of accuracy significantly impact subsequent opinion dynamics modeling? Our argument is that as long as the classifier maintains **internal consistency**, i.e., it applies its classification criteria consistently across all data samples, moderate deviations from human judgment (e.g., in labeling highly

polarized opinions) will not undermine the validity of dynamic trend analysis. This is because public opinion modeling primarily focuses on **relative changes and category proportions over time**, rather than the accuracy of individual labels.

Nonetheless, improving the alignment between model predictions and human annotations remains important, as it enhances the interpretability of opinion distributions in downstream analyses, as discussed in Section 4.4.

Appendix A.3. Methodology for Subjectivity Judgment of Tweets

The LLM used for subjectivity judgment of tweets is GPT-4.1. The prompt used for classification is as follows:

> *Analyze the given TWEET and classify it into one of the following categories: objective (primarily describing facts or events), subjective (primarily expressing personal emotions or opinions), or both. Provide a rationale for your classification. Output the result in the following format: '[type]: reason.'*
>
> *TWEET:*

No example-based guidance was provided in the prompt; instead, a zero-shot approach was employed. Despite the absence of in-depth semantic analysis, this method achieves an accuracy exceeding 95%.

Appendix B. Word Clouds for Different Sentiments

In this appendix section, we present the word cloud figures for the remaining sentiments in both datasets.



(**a**) (**b**) (**c**)

Figure A1. Word clouds generated from tweets in the AGI dataset with "Mildly Negative," "Neutral," and "Mildly Positive" sentiments: (**a**) tweets labeled as "Mildly Negative"; (**b**) tweets labeled as "Neutral"; (**c**) tweets labeled as "Mildly Positive." The size of each word reflects its relative frequency within the corresponding sentiment class.



(**a**) (**b**) (**c**)

Figure A2. Word clouds generated from tweets in the HIC dataset with "Mildly Positive," "Neutral," and "Mildly Negative" sentiments: (**a**) tweets labeled as "Mildly Positive"; (**b**) tweets labeled as "Neutral"; (**c**) tweets labeled as "Mildly Negative." The size of each word reflects its relative frequency within the corresponding sentiment class.

References

1. Sagarra, O.; Gutiérrez-Roig, M.; Bonhoure, I.; Perelló, J. Citizen science practices for computational social science research: The conceptualization of pop-up experiments. *Front. Phys.* **2016**, *3*, 93. [CrossRef]
2. Okawa, M.; Iwata, T. Predicting opinion dynamics via sociologically-informed neural networks. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 14–18 August 2022; pp. 1306–1316.
3. Biondi, E.; Boldrini, C.; Passarella, A.; Conti, M. Dynamics of opinion polarization. *IEEE Trans. Syst. Man Cybern. Syst.* **2023**, *53*, 5381–5392.
4. Loomba, S.; De Figueiredo, A.; Piatek, S.J.; De Graaf, K.; Larson, H.J. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat. Hum. Behav.* **2021**, *5*, 337–348. [PubMed]
5. Pennycook, G.; Epstein, Z.; Mosleh, M.; Arechar, A.A.; Eckles, D.; Rand, D.G. Shifting attention to accuracy can reduce misinformation online. *Nature* **2021**, *592*, 590–595. [CrossRef]
6. Galam, S. Unanimity, coexistence, and rigidity: Three sides of polarization. *Entropy* **2023**, *25*, 622. [CrossRef]
7. Li, K.; Liang, H.; Kou, G.; Dong, Y. Opinion dynamics model based on the cognitive dissonance: An agent-based simulation. *Inf. Fusion* **2020**, *56*, 1–14. [CrossRef]
8. Hassani, H.; Razavi-Far, R.; Saif, M.; Chiclana, F.; Krejcar, O.; Herrera-Viedma, E. Classical dynamic consensus and opinion dynamics models: A survey of recent trends and methodologies. *Inf. Fusion* **2022**, *88*, 22–40. [CrossRef]
9. Li, Y.; Xu, Z. A bibliometric analysis and basic model introduction of opinion dynamics. *Appl. Intell.* **2023**, *53*, 16540–16559. [CrossRef]
10. Chuang, P.M.; Shirai, K.; Kertkeidkachorn, N. Identification of opinion and ground in customer review using heterogeneous datasets. In Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024), Rome, Italy, 24–26 February 2024; Volume 2, pp. 68–78.
11. Chen, B.; Wang, X.; Zhang, W.; Chen, T.; Sun, C.; Wang, Z.; Wang, F.Y. Public opinion dynamics in cyberspace on Russia–Ukraine War: A case analysis with Chinese Weibo. *IEEE Trans. Comput. Soc. Syst.* **2022**, *9*, 948–958. [CrossRef]
12. Wu, Z.; Zhou, Q.; Dong, Y.; Xu, J.; Altalhi, A.H.; Herrera, F. Mixed opinion dynamics based on DeGroot model and Hegselmann–Krause model in social networks. *IEEE Trans. Syst. Man Cybern. Syst.* **2023**, *53*, 296–308.
13. Chen, K.; He, Z.; Burghardt, K.; Zhang, J.; Lerman, K. Isamasred: A public dataset tracking reddit discussions on Israel-Hamas conflict. In Proceedings of the International AAAI Conference on Web and Social Media, Athens, Greece, 27–30 May 2024; Volume 18, pp. 1900–1912.
14. Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.S.; Nemade, G.; Ravi, S. GoEmotions: A Dataset of Fine-Grained Emotions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), Online, 5–10 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 4040–4054.
15. Hettiarachchi, H.; Al-Turkey, D.; Adedoyin-Olowe, M.; Bhogal, J.; Gaber, M.M. Ted-s: Twitter event data in sports and politics with aggregated sentiments. *Data* **2022**, *7*, 90. [CrossRef]
16. Gupta, S.; Bouadjenek, M. R.; Robles-Kelly, A.; Lee, T.-K.; Nguyen, T.T.; Nazari, A.; Thiruvady, D. Covid19-twitter: A Twitter-based dataset for discourse analysis in sentence-level sentiment classification. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, Boise, ID, USA, 21–25 October 2024; pp. 5370–5374.
17. Braha, D.; De Aguiar, M.A.M. Voting contagion: Modeling and analysis of a century of US presidential elections. *PLoS ONE* **2017**, *12*, e0177970.
18. Redner, S. Reality-inspired voter models: A mini-review. *Comptes Rendus Phys.* **2019**, *20*, 275–292. [CrossRef]
19. He, M.; Zhang, X.J. Affinity, value homophily, and opinion dynamics: The coevolution between affinity and opinion. *PLoS ONE* **2023**, *18*, e0294757.
20. Fernández-Gracia, J.; Suchecki, K.; Ramasco, J.J.; San Miguel, M.; Eguíluz, V.M. Is the voter model a model for voters? *Phys. Rev. Lett.* **2014**, *112*, 158701. [CrossRef]
21. Phan, H.V.; Nguyen, N.H.; Nguyen, H.T.; Hegde, S. Policy uncertainty and firm cash holdings. *J. Bus. Res.* **2019**, *95*, 71–82. [CrossRef]
22. Balietti, S.; Getoor, L.; Goldstein, D.G.; Watts, D.J. Reducing opinion polarization: Effects of exposure to similar people with differing political views. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2112552118.
23. Sharma, N.A.; Ali, A.S.; Kabir, M.A. A review of sentiment analysis: Tasks, applications, and deep learning techniques. *Int. J. Data Sci. Anal.* **2025**, *19*, 351–388.
24. Wang, Z.; Hu, Z.; Ho, S.-B.; Cambria, E.; Tan, A.-H. MiMuSA—Mimicking human language understanding for fine-grained multi-class sentiment analysis. *Neural Comput. Appl.* **2023**, *35*, 15907–15921.
25. AlQahtani, A.S.M. Product sentiment analysis for amazon reviews. *Int. J. Comput. Sci. Inf. Technol.* **2021**, *13*, 15–30. [CrossRef]
26. Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C.D.; Ng, A.Y.; Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1631–1642.

27. Prabhakar, E.; Santhosh, M.; Krishnan, A.H.; Kumar, T.; Sudhakar, R. Sentiment analysis of US airline twitter data using new adaboost approach. *Int. J. Eng. Res. Technol.* **2019**, *7*, 1–6.
28. Maas, A.; Daly, R.E.; Pham, P.T.; Huang, D.; Ng, A.Y.; Potts, C. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 142–150.
29. Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 23–24 August 2014; pp. 27–35.
30. Saias, J. Sentiue: Target and Aspect based Sentiment Analysis in SemEval-2015 Task 12. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, CO, USA, 4–5 June 2015; Association for Computational Linguistics: Stroudsburg, PA, USA, 2015; pp. 767–771.
31. Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; AL-Smadi, M.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; De Clercq, O.; et al. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, CA, USA, 16–17 June 2016; pp. 19–30.
32. Haque, T.U.; Saber, N.N.; Shah, F.M. Sentiment analysis on large scale Amazon product reviews. In Proceedings of the 2018 IEEE International Conference on Innovative Research and Development (ICIRD), Kuala Lumpur, Malaysia, 12–14 July 2018; pp. 1–6.
33. Husein, A.M.; Livando, N.; Andika, A.; Chandra, W.; Phan, G. Sentiment analysis of hotel reviews on Tripadvisor with LSTM and ELECTRA. *Sink. J. Penelit. Tek. Inf.* **2023**, *7*, 733–740.
34. Munikar, M.; Shakya, S.; Shrestha, A. Fine-grained sentiment classification using BERT. In Proceedings of the 2019 Artificial Intelligence for Transforming Business and Society (AITB), Kathmandu, Nepal, 12–14 December 2019; Volume 1, pp. 1–5.
35. Onyibe, C.; Habash, N. OMAM at SemEval-2017 Task 4: English sentiment analysis with conditional random fields. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 670–674.
36. Zhang, W.; Deng, Y.; Liu, B.; Pan, S.J.; Bing, L. Sentiment Analysis in the Era of Large Language Models: A Reality Check. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2024), Mexico City, Mexico, 16–21 June 2024; pp. 3881–3906.
37. Chuang, Y.-S.; Goyal, A.; Harlalka, N.; Suresh, S.; Hawkins, R.; Yang, S.; Shah, D.; Hu, J.; Rogers, T. Simulating Opinion Dynamics with Networks of LLM-based Agents. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, 9–14 June 2024; pp. 3326–3346.
38. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019), Hong Kong, China, 3–7 November 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3980–3990.
39. Sun, Z.; Wang, D.; Li, Z. Quantification and Evolution of Online Public Opinion Heat Considering Interactive Behavior and Emotional Conflict. *Entropy* **2025**, *27*, 701. [CrossRef]
40. Volchenkov, D.; Putkaradze, V. Mathematical Theory of Social Conformity I: Belief Dynamics, Propaganda Limits, and Learning Times in Networked Societies. *Mathematics* **2025**, *13*, 1625. [CrossRef]
41. Qian, X.; Han, W.; Yang, J. From the DeGroot Model to the DeGroot-Non-Consensus Model: The Jump States and the Frozen Fragment States. *Mathematics* **2024**, *12*, 228.
42. Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; Cherry, C. Semeval-2016 Task 6: Detecting Stance in Tweets. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, CA, USA, 16–17 June 2016; pp. 31–41.
43. Pick, R. K.; Kozhukhov, V.; Vilenchik, D.; Tsur, O. Stem: Unsupervised Structural Embedding for Stance Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 22–28 February 2022; Volume 36, pp. 11174–11182.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.