1 NBMSG Feedback Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to simulate volunteer feedback in a sports event context, filling a gap in available ABSA (Aspect-Based Sentiment Analysis) datasets, particularly in sports surveys. It aims to provide a conservative evaluation of LLMs (Large Language Models) by mitigating the memorization effect and exploring optimal context windows for dataset annotation. The dataset also contributes to understanding LLM performance in generating and annotating data with diverse content, specifically focusing on implicit aspects.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

To be revealed later

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

To be revealed later

Any other comments?

The dataset and associated prompts are publicly available, ensuring transparency and enabling further research in sentiment analysis and LLM evaluation.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances in the dataset represent documents containing simulated feedback from volunteers participating in a sports event. Each document reflects a combination of implicit and explicit aspects related to the event, capturing a range of sentiments.

How many instances are there in total (of each type, if appropriate)?

The dataset comprises 480 documents.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a curated sample specifically designed for this study, rather than a representative sample from a larger set of volunteer feedback instances. It was generated to explore specific aspects of ABSA in the con-

text of sports events, thus not aiming to represent a broader population.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance consists of unprocessed text data, representing the feedback provided by simulated volunteers, along with annotations of aspects and their corresponding sentiments.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance is labeled with aspects related to the event and their corresponding sentiment polarities (positive, negative, neutral).

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No information is missing from the instances; however, 12.5% of the annotations were adjusted post-generation to enhance accuracy.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

No explicit relationships between individual instances are defined within the dataset.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

Yes, the dataset is split into 80% for testing and 20% for fine-tuning. This split was chosen due to the dataset's small size and to ensure that the majority of the data is used for evaluating model performance.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

There are no known errors or redundancies in the dataset. However, the dataset includes a controlled amount of noise, particularly in the form of implicit aspects, to challenge model performance.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained and does not rely on any external resources.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the

content of individuals non-public communications)? If so, please provide a description.

No, the dataset does not contain any confidential data.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No, the dataset does not contain any data that could be considered offensive, insulting, or anxiety-inducing.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset simulates feedback from volunteers, although it does not represent real individuals.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

The dataset identifies subpopulations based on the following characteristics:

- **Age:** The prompts specify different age ranges for the survey respondents, including 18-25, 25-35, 35-50, and 50+.
- **Education Level:** The prompts also vary the education level of the respondents, including high school, college diploma, and university degree.
- **Minority Status:** The core context of the survey is that the respondents are volunteers at a sports event for minorities, implying that they belong to various minority groups. However, the specific minority groups are not explicitly identified in the prompts.

The distribution of these subpopulations is not explicitly stated in the dataset, but it can be inferred that the prompts aim to generate a diverse range of responses, suggesting a relatively balanced distribution across the different age groups, education levels, and occupations.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No, it is not possible to identify individuals from the dataset as it does not represent real persons.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No, the dataset does not contain any sensitive data.

Any other comments?

None.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data

(e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data associated with each instance was generated by large language models, specifically GPT-4 and Gemini 1.0 Ultra, based on designed prompts. These prompts were used to simulate survey responses from volunteers at a sports event. Validation was conducted through a combination of model cross-evaluation, human review by undergraduate volunteers, and expert oversight to ensure the quality and accuracy of the generated annotations.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data was collected using chat interfaces of GPT-4 and Gemini 1.0 Ultra. These LLMs generated responses to prompts designed to simulate survey data. The generated data was validated through multiple stages, including model-based evaluations, human annotation, and expert review to refine and ensure the quality of the dataset.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The dataset is not a sample from a larger set. It was specifically generated for the purpose of this study using the described prompts and LLMs.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data collection process involved the use of LLMs for initial data generation, followed by human annotation and review. Three undergraduate student volunteers participated in the review process. These students were not monetarily compensated but participated as part of an academic exercise.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data was generated and annotated over a period of one month, during March 2024. This timeframe matches the creation timeframe of the data instances, as the dataset was specifically created for this study within that period.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Yes, an ethical review was conducted by an expert in ethical AI practices. The review ensured that the dataset adhered to ethical standards, particularly concerning the use of LLMs for data generation and the potential biases associated with this approach. The outcomes of the review affirmed

that the dataset posed no foreseeable harm if used for sentiment analysis research.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset simulates responses from volunteers at a sports event, which are meant to represent human feedback in a survey context.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data was not collected from individuals directly. Instead, it was artificially generated by large language models based on prompts designed to simulate survey responses.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Not applicable, as the data was artificially generated by LLMs and did not involve real individuals.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Not applicable, as the data was artificially generated by LLMs and did not involve real individuals.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Not applicable, as the data was artificially generated by LLMs and did not involve real individuals.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Not applicable.

Any other comments?

No additional comments.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data underwent several preprocessing steps. This included correcting misclassifications of aspect-based sentiment analysis tags, adjusting aspect polarities, and ensuring consistency in the labeling process.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Yes, the raw data produced by the LLMs for feedback generation is available on the associated GitHub repository. This repository includes both the raw text generated by the models and the prompts used for data generation.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

The scripts for data cleaning are available upon reasonable request.

Any other comments?

No additional comments.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

The dataset has not yet been used for any published tasks. It was specifically created for the purpose of this study and is intended for future use in research on aspect-based sentiment analysis (ABSA) and related natural language processing (NLP) tasks.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

There is currently no repository linking to papers or systems that use the dataset, as it has not yet been utilized in any published work. However, the dataset and associated scripts are available in a GitHub repository, which will be updated with references to any future work using the dataset. [Insert GitHub link here]

What (other) tasks could the dataset be used for?

The dataset could be used for a variety of tasks, including but not limited to:

- Sentiment analysis, particularly aspect-based sentiment analysis (ABSA).
- Text classification, focusing on detecting opinions or sentiments in feedback data.
- Natural language generation tasks, for evaluating the quality of text generation by language models.
- Survey data analysis, specifically in the context of event feedback.
- Training and fine-tuning language models on feedback and sentiment-related tasks.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

The dataset was generated using large language models (LLMs), which can carry inherent biases from the training data on which the models were built. Users should be aware that the simulated feedback might reflect these biases, which could impact tasks involving sensitive topics or demographic-specific analysis. To mitigate potential harms, users should critically evaluate

the dataset's outputs and consider additional steps, such as bias detection and correction, before deploying models trained on this data in real-world applications.

Are there tasks for which the dataset should not be used? If so, please provide a description.

The dataset should not be used for tasks that require real-world demographic analysis or tasks where accurate representation of actual human populations is critical, as the data is artificially generated and does not represent real human feedback. It is also not suitable for applications that could result in decisions affecting individuals' lives, such as hiring, lending, or legal judgments, due to the synthetic nature of the data.

Any other comments?

No additional comments.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset will be made publicly available to third parties for research purposes. It will be distributed through an open-access repository, allowing other researchers to use it in their work.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The dataset will be distributed via GitHub as a downloadable repository. As of now, the dataset does not have a digital object identifier (DOI), but one may be assigned in the future if the dataset becomes widely used.

When will the dataset be distributed?

The dataset will be distributed upon the publication of the associated research paper or at a time decided by the research team. Interested parties can access the dataset through the provided GitHub link.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset will be distributed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). This license allows others to share, use, and build upon the dataset, as long as appropriate credit is given.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No third parties have imposed IP-based or other restrictions on the data associated with the instances.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other ac-

cess point to, or otherwise reproduce, any supporting documentation.

No export controls or other regulatory restrictions apply to the dataset or to individual instances.

Any other comments?

No additional comments.

Maintenance

Who will be maintaining the dataset?

The dataset will be supported, hosted, and maintained by the research team that created it. This includes ongoing updates and addressing any issues that arise with the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The research team can be contacted via the email address provided in the associated GitHub repository or the contact information listed in the associated research paper.

Is there an erratum? If so, please provide a link or other access point.

As of now, there is no erratum associated with the dataset. Any future errata will be posted on the GitHub repository page.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Yes, the dataset will be updated as needed, primarily to correct labeling errors, add new instances, or delete problematic instances. Updates will be managed by the research team and communicated to users through the GitHub repository, where change logs will be maintained. Users can also subscribe to the repository to receive notifications of updates.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Not applicable, as the dataset is artificially generated and does not involve real individuals.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Older versions of the dataset will be archived and made available in the GitHub repository, allowing users to access previous versions if needed. The repository will include clear documentation on version updates and any significant changes made in newer versions.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Yes, the GitHub repository will be open for contributions from other researchers. Contributions will be reviewed and validated by the research

team before being integrated into the main dataset. Contributions that meet quality standards will be merged, and contributors will be credited. All contributions and updates will be communicated to the wider user community through the repository's update logs and announcements.

Any other comments?

No additional comments.