Indian Media Dataset with Manipulated Titles and Facial Expressions

Shubhi Agarwal

April 2024

1 Datasheets for datasets

Indian Media Dataset with Manipulated Titles and Facial Expressions "document [the dataset] motivation, composition, collection process, recommended uses, and so on. This dataset has immense potential to drive advancements in sentiment analysis, natural language processing, and computer vision. By offering a curated collection of news articles alongside manipulated titles and sentiment labels, it facilitates research into bias mitigation, model robustness, and ethical AI. With its integration of textual and visual data, the dataset enables interdisciplinary studies, empowering researchers, practitioners, and policymakers to develop more transparent and accountable DeepFake Detection systems. "

The primary motivation behind creating this dataset was to detect media manipulation within the Indian media landscape using both textual and visual information and detecting deepfake using multiple modalities. Inspired by the paper *'Detecting and grounding multi-modal media manipulation and beyond'* by Shao et al [1], my aim is to implement their methodology in the context of Indian media. By providing a diverse range of news articles alongside manipulated titles and facial expressions, this dataset serves as a valuable resource for studying and combatting media manipulation in the digital age.

2 Template

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 with the overarching goal of detecting **media manipulation within the Indian media landscape**, leveraging both **textual and visual data**. This endeavor was motivated by the pressing need to address the proliferation of misin-

formation, propaganda, and fake news in online platforms. The specific task at hand was to develop and implement a methodology for detecting and grounding multi-modal media manipulation, as outlined in the paper *'Detecting and grounding multi-modal media manipulation and beyond'* by Shao et al [1], within the context of Indian media. The dataset aimed to fill a critical gap in existing research by providing a comprehensive collection of news articles paired with manipulated titles and facial expressions, enabling researchers and practitioners to study and combat media manipulation more effectively.

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

This dataset is created by Shubhi Agarwal, an undergraduate student of Indian Institute of Technology, Guwahati.

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.

The dataset creation was self funded.

Any other comments? The dataset is created by scraping Indian Express website and using AI tools like Media.io

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 within the dataset represent multimedia content extracted from Indian media sources, primarily consisting of news articles and associated images.

1. Text Dataset: Each instance in the text dataset represents a news article and contains detailed information about the article, such as its ID, title, description, publication date, URL of the article and the image, headline, article length, sentiment label, and manipulated headline. These instances can be considered as documents or textual data entries.

2. Image Dataset: Instances in the image dataset represent images associated with the news articles. This includes both original images and manipulated images based on various facial attributes like sadness and happiness. Each instance can be viewed as a photo or image data entry.

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

There are 20,000 instances in the text dataset and 20,000 corresponding image URL. However, out of the 20,000 images currently 75 images have original images downloaded along with their corresponding attribute manipulated images making 150 manipulated images and 75 original images. The rest of the manipulated images will also be uploaded later.

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).

As the dataset only includes news articles from the Indian Express website and covers the years 2019-2020, it is not representative of the entire Indian media landscape. The sample is limited both in terms of its geographic coverage (focused on content from a single news website) and temporal scope (limited to two years). The larger dataset would include articles from different websites and over many years.

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

For the text dataset:
Raw Data:

- Unprocessed text, including the title, description, headline, and full content of the news article.
- Metadata such as the article ID, publication date, URL, and article length.

Features: Extracted features from the raw text, such as sentiment labels (positive, negative, neutral) and manipulated headlines.

For the image dataset:

Raw Data:Original images associated with each news article. Manipulated images based on different facial attributes like sadness and happiness.

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

Every article is given a label named article id. Image from a particular article is saved as (article id).jpg. Also the manipulated images are stored in different folders with (face atribute) as the folder name and the name of the manipulated images are same as the article id. Each image is labelled as origianl or manipulted (sad , happy)

Also, each article is also given a label of either positive, negative or neutral based on the semantics of the content of the article.

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.

In some articles, the images are missing as they have been taken out from the articles.

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, the relationships between different instances are not explicit.

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.

Recommended data splits:

1. Training Set: Used for model training. Size: 70-80% of the dataset.

Rationale: Sufficient data for model learning without overfitting.

2. Validation Set: Used for hyperparameter tuning and model performance assessment during training. Size: 10-15% of the dataset. Rationale: Helps optimize model performance and prevent overfitting.

3. Testing Set: Used for final model evaluation. Size: 10-15% of the dataset. Rationale: Provides an unbiased estimate of model performance on unseen data, ensuring generalization ability.

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

Yes, there are some image redundancies as in some instances the images used are same but the article is different.

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 relies on the Indian Express website for news articles and associated images. While the dataset includes scraped image URLs and article information stored in CSV format, there is a risk of image unavailability if the external website removes them. However, the article information remains accessible even if the images are no longer available.

Official archival versions of the complete dataset, including external resources, are not available. Therefore, it is advisable to create backups of the dataset to mitigate the risk of data loss or unavailability.Though the dataset here provides downloaded images as well but it do not contain the images that have already been taken down.

As of April 2024, there were no fees associated with accessing or using the dataset. However, users should stay informed about any potential changes in fees or licensing terms that may arise in the future.

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 do not contain such 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.

The dataset contains news articles and associated images, which may include content that could potentially be sensitive or distressing to some individuals. While efforts are made to ensure that the dataset adheres to ethical guidelines and standards, it is possible that certain news topics or images may

contain content that could be considered offensive, insulting, threatening, or anxiety-inducing.

News articles may cover a wide range of topics, including but not limited to politics, social issues, crime, and disasters, which could contain sensitive or distressing subject matter. Additionally, images associated with news articles may depict scenes or events that could evoke strong emotional responses.

It is important for users of the dataset to exercise discretion and sensitivity when accessing and analyzing the content, particularly when dealing with potentially sensitive topics. Researchers and practitioners should consider the potential impact of the dataset on themselves and others and take appropriate precautions to mitigate any adverse effects.

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

Yes, the dataset does relate to people. As it comprises news articles from the Indian Express website, it likely includes content that involves individuals, such as politicians, public figures, activists, and members of the general population. News articles often cover stories about people, including their actions, achievements, opinions, and experiences, across a wide range of contexts and topics. Additionally, the associated images may feature individuals captured in various settings, events, or activities depicted in the news articles. However, it's important to note that the dataset does not directly identify or provide personal information about specific 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 may potentially identify subpopulations based on various demographic characteristics such as age, gender, ethnicity, occupation, or other factors. However, without specific information on how these subpopulations are identified or labeled within the dataset, it is challenging to provide a detailed description of their distributions.

In news articles, subpopulations may be implicitly identified through the content of the articles themselves. For example, articles may focus on specific demographic groups or discuss issues that predominantly affect certain populations. Similarly, images associated with the articles may depict individuals representing various demographic characteristics.

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.

It is possible that the dataset contains information that could potentially identify individuals, either directly or indirectly, particularly if the news articles or associated images contain personal information, such as names, addresses, photographs, or other identifiable characteristics.

Direct identification may occur if the dataset includes explicit personal identifiers, such as names or unique identifiers, that directly link individuals to specific articles or images.

Indirect identification may occur through a combination of data points

or contextual information present in the dataset. For example, even if names are not explicitly mentioned, details provided in news articles or images, such as occupation, location, affiliations, or descriptions of events or activities, could potentially allow individuals to be identified when combined with external knowledge or other datasets.

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.

The dataset may contain data that could be considered sensitive, depending on the content of the news articles and associated images. Some examples of sensitive data that may be present in the dataset include:

1. Racial or Ethnic Origins: News articles or images may discuss or depict individuals or groups based on their racial or ethnic backgrounds.

2. Sexual Orientations: Articles or images may address issues related to sexual orientation, LGBTQ+ communities, or individuals' identities.

3. Religious Beliefs: Content may reference religious beliefs, practices, or events related to specific religious groups or individuals.

4. Political Opinions: Articles may discuss political figures, parties, ideologies, or events, potentially revealing individuals' political affiliations or opinions.

5. Health Data: News articles may cover health-related topics, including medical conditions, treatments, or public health issues.

6. Locations: Information about specific geographic locations or events may be included, potentially revealing individuals' whereabouts or activities especially for public figures.

7. Criminal History: News coverage may include reports on criminal activities, legal proceedings, or individuals involved in criminal cases.

Given the diverse range of topics covered in news articles, it is possible that the dataset contains sensitive information across various dimensions. Researchers and users should handle this data with care, ensuring compliance with privacy regulations and ethical guidelines, and take appropriate measures to protect individuals' privacy and confidentiality.

Any other comments? For future users accessing the dataset, it's crucial to ensure that it aligns with their research objectives and ethical standards. The dataset should encompass diverse topics and perspectives, accompanied by rigorous quality assurance measures to verify data accuracy and completeness. Comprehensive documentation, including metadata and usage guidelines, facilitates understanding and reproducibility. Ethical considerations, such as privacy rights and data protection, must be prioritized throughout the dataset's lifecycle. Accessibility to a broad user base and a commitment to continuous improvement further enhance the dataset's utility and impact. By adhering to these principles, users can leverage the dataset effectively for research and analysis while upholding ethical stan-

dards and promoting transparency.

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 was acquired through web scraping of the Indian Express website, leveraging publicly available information from a trusted source. Given the reliability of the source, extensive verification of the data was not necessary, except for potential issues encountered during the scraping process. However, manual verification was conducted to ensure the effectiveness of the scraping, semantic attribute generation, and face attribute manipulation processes. This validation step helped confirm the accuracy and quality of the acquired data, ensuring that it aligns with the intended objectives of the dataset creation.

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 associated with each instance was acquired through a multi-step process.

- Initially, news data was scraped from the Indian Express website, including titles and descriptions, using a web scraping script. The URLs for the images associated with each news article were extracted from a publicaly available dataset of news article [2]. Subsequently, the images were downloaded from these URLs and stored in a local folder.

- Following the image acquisition, further processing was conducted to manipulate facial expressions using AI tools such as Media AI [**?**]. This involved manually filtering images and applying facial attribute manipulation techniques to change expressions.

- Additionally, sentiment labels (Positive, Negative, Neutral) were assigned to each news article based on its description using the Flair library. This sentiment analysis provided insights into the emotional tone conveyed by the article.

- Furthermore, fake or manipulated titles were generated for each news article using the sentiment labels obtained from the sentiment analysis. Specifically, titles with positive or neutral sentiments were transformed to negative sentiments, and titles with negative sentiments were transformed to positive sentiments.

- The data acquisition process primarily involved directly observable data sources, such as raw text from news articles and image URLs extracted from webpages. The sentiment labels assigned to the articles were in-

directly inferred from the article descriptions using sentiment analysis techniques.

Furthermore, a comprehensive error handling mechanism was employed to address any issues encountered during the data acquisition, manipulation, or sentiment analysis stages. This included error detection, logging, and resolution strategies to mitigate potential errors and ensure the integrity of the dataset. Additionally, a sample of the manipulated images and generated titles underwent manual inspection to validate the effectiveness of the facial attribute manipulation and sentiment transformation processes. Continuous monitoring and refinement of the data processing pipeline were also carried out iteratively to improve data quality over time and enhance the reliability of the dataset for future use.

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

The sampling strategy employed in this scenario was deterministic, as all available data within the defined parameters (time interval: 2019-20 and website: Indian Express) were included in the dataset.

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 was done by Shubhi Agarwal of IIT Guwahati as part of an individual project. Therefore, no compensation was given to anyone.

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 collection was done in April 2024. The timeframe does not match with instances of the dataset which is 2019-2020.

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.

No, there were no ethical review processes conducted for this dataset. As it was developed as a course project by the university, it did not undergo review by an institutional review board or similar ethics review body. Therefore, there are no formal documentation or outcomes related to ethical review processes available for this dataset.

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

Yes, the dataset relates to people indirectly through the news articles and associated images. The news articles likely contain information about individuals, events, or activities involving people, such as politicians, celebrities, or other public figures. Additionally, the images may include photographs of individuals relevant to the news stories. However, it's important to note that the dataset does not directly identify or provide personal information about specific individuals.

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 obtained via third-party sources, specifically through web scraping of the Indian Express website. The news articles and associated images were publicly available on the website, and the data was collected from these sources without direct interaction with the individuals mentioned in the articles or depicted in the images. Therefore, no data was collected directly from the individuals in question.

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.

As the data was collected via web scraping of publicly available information from the Indian Express website, there was no direct notification provided to the individuals mentioned in the news articles or depicted in the images. Since the data was obtained from a website accessible to the public, individuals would not have been specifically notified about the data collection process. Therefore, there are no screenshots, notifications, or language of notification to reproduce in this context.

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.

As the data was collected via web scraping of publicly available information from the Indian Express website, there was no direct notification provided to the individuals mentioned in the news articles or depicted in the images. Since the data was obtained from a website accessible to the public, individuals would not have been specifically notified about the data collection process. Therefore, there are no screenshots, notifications, or language of notification to reproduce in this context.

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 in this scenario

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.

No, an analysis of the potential impact of the dataset and its use on data subjects, such as a data protection impact analysis (DPIA), has not been conducted. Since the dataset primarily consists of publicly available news articles and associated images from the Indian Express website, and does not contain personally identifiable information or sensitive data about specific individuals, the need for a formal

DPIA was not identified. Therefore, there are no specific documentation or outcomes related to such an analysis available for this dataset.

Any other comments?

Future users of this dataset should exercise caution regarding data integrity, adhere to ethical guidelines, prioritize reproducibility, ensure proper citation and attribution, and uphold data security standards. While efforts were made to maintain accuracy during collection and processing, users should validate the data for their specific needs. Transparency in methodologies and proper acknowledgment of sources are crucial for academic integrity. Additionally, users should handle the data securely to prevent unauthorized access or misuse. By approaching the dataset with diligence and ethical awareness, users can maximize its utility while upholding standards of integrity and responsibility.

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, preprocessing, cleaning, and labeling of the data were performed as part of the dataset preparation process.

- **Text Preprocessing:** Tokenization: Breaking down the text into individual tokens (words or subwords) for analysis. Removal of special characters and punctuation: Stripping out non-alphanumeric characters and punctuation marks from the text and making all the characters lowercase.
- **Image Preprocessing:** All the images were resized to a common dimension. Additionally fake images were created based on face attribute manipulation.
- **Labelling:** Sentiment labels were assigned to the news articles based on the sentiment analysis of their descriptions. The sentiment analysis task classified each article into one of three categories: Positive, Negative, or Neutral. Additionally, manipulated titles were generated for a subset of articles by transforming the semantics of the original titles to create variations with opposite sentiment polarity

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 dataset is stored in Indian Media Dataset with Manipulated Titles and Facial Expressions

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

Yes, Media.AI was used to create manipulated images.

Uses

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

No, the dataset has not been used for any task till now.

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.

Not applicable

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

In addition to sentiment analysis and facial attribute manipulation, the dataset could be utilized for various other tasks, including:

1. **Fake news detection :** Developing models to distinguish between authentic and fabricated news articles based on textual and visual features.

2. **Text Summerization :** Generating concise summaries of news articles based on their titles and descriptions.

3. **Topic modeling:** Identifying prevalent topics or themes within the news articles using techniques like Latent Dirichlet Allocation (LDA) or Non-Negative Matrix Factorization (NMF).

4. **Image captioning:** Generating descriptive captions for news images using natural language processing (NLP) techniques.

5. **Multi-modal analysis:** Exploring correlations between textual content and image attributes to gain deeper insights into news articles' content and sentiment.

6. **Media studies:** Investigating trends and patterns in news content, sentiment, and manipulation techniques to understand their impact on public perception and media consumption habits.

7. **Ethical AI and misinformation:** Examining the ethical implications of AI-based manipulation techniques on news content and exploring strategies to mitigate misinformation and manipulation in digital media.

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's composition, collection, and preprocessing methods carry implications that users should heed to prevent unfair treatment or undesirable outcomes. Firstly, potential biases inherent in the data collection process, reflecting editorial stances or underrepresented perspectives, could inadvertently reinforce existing biases. Secondly, the inclusion of manipulated content, such as titles and facial expressions, poses risks of misinformation or deception if not properly identified and addressed. These factors demand careful consideration to en-

sure ethical use and mitigate potential harm.

To navigate these challenges, users must conduct thorough ethical reviews and risk assessments before utilizing the dataset. Implementing rigorous validation procedures to detect and mitigate biases and manipulations is essential. Additionally, transparent documentation of the dataset's limitations and biases in resulting analyses or applications is crucial. Engaging with diverse stakeholders, including affected communities, can provide valuable insights and perspectives to inform ethical decision-making. By adhering to these guidelines and prioritizing fairness, transparency, and accountability, users can mitigate risks and promote responsible use of the dataset.

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

While the dataset offers valuable insights into media manipulation, sentiment analysis, and natural language processing, there are certain tasks for which it may not be suitable:

1. **Individual profiling:** The dataset should not be used for individual profiling or targeting based on sensitive attributes such as ethnicity, religion, or political affiliation. Doing so could perpetuate discrimination or amplify biases present in the data, leading to unfair treatment or harm to individuals or groups.

2. **Unethical manipulation:** Users should refrain from using the dataset to create or disseminate manipulated content with malicious intent, such as spreading misinformation or inciting violence. Engaging in such activities would undermine trust in AI and media integrity, potentially causing significant societal harm.

3. **Biased algorithm development:** Care should be taken when using the dataset to train machine learning algorithms, as biases present in the data could be inadvertently learned and perpetuated by the models. Developers should employ bias mitigation techniques and ensure that their algorithms prioritize fairness, transparency, and accountability.

Overall, users should exercise caution and adhere to ethical guidelines when utilizing the dataset, avoiding tasks that could result in unfair treatment, harm, or unethical behavior.

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 distributed to third parties outside of the entity through GitHub. While the dataset was initially created for a university course project, it is available on a GitHub repository, making it accessible to a wider audience for educational and research purposes. Users can access and download the dataset from the repository, facilitating its distribution beyond the original entity.

How will the dataset will be distributed (e.g., tarball on website,

API, GitHub) Does the dataset have a digital object identifier (DOI)?

Given the dataset's origin as an individual initiative for a university course project, there are currently no formal plans for widespread distribution to third parties outside of the originating entity. However, the dataset is publicly accessible via a GitHub repository Indian Media Dataset with Manipulated Titles and Facial Expressions, providing interested parties with the opportunity to access and utilize the data. Despite the absence of a Digital Object Identifier (DOI), the GitHub repository serves as a central location for accessing the dataset and any associated documentation. While widespread distribution may not be immediate, the dataset's availability on GitHub enables potential future sharing or collaboration opportunities as needed.

When will the dataset be distributed?

The dataset is already available on GitHub Indian Media Dataset with Manipulated Titles and Facial Expressions,

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 is distributed under CC0-1.0 license.

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.

As of now, there are no IP-based or other restrictions imposed by third parties on the data associated with the instances in the dataset. The dataset was collected and processed from publicly available sources, such as the Indian Express website, without encountering any explicit licensing terms or restrictions. Therefore, users can access and utilize the dataset freely for educational and research purposes without encountering any fees or restrictions from third parties.

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

There are no export controls or regulatory restrictions known to apply to the dataset or to individual instances within it. The dataset was collected from publicly available sources and does not contain sensitive or restricted information that would be subject to such controls. Therefore, users should be able to freely access, use, and distribute the dataset without encountering any regulatory restrictions.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

No one will be maintaining it.

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

The owner can be contacted via Email: shubhiagarwal2494@gmail.com

Is there an erratum? If so, please provide a link or other access point.
No.

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)?
The data will be updated by the owner and the updates will be communicated through GitHub where it is publicly accessible.

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.
No

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.
Not Applicable.

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, others are welcome to contribute to the dataset through the GitHub repository where it is hosted. The repository provides a platform for collaboration and version control, allowing users to suggest improvements, augmentations, or extensions to the dataset. Contributions can take the form of additional data samples, improved preprocessing techniques, or enhancements to existing features.

References

[1] R. Shao, T. Wu, J. Wu, L. Nie, and Z. Liu, "Detecting and grounding multi-modal media manipulation and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[2] [Online]. Available: https://www.kaggle.com/datasets/pulkitkomal/news-article-data-set-from-indian-express