International Journal of Computers and Informatics

Journal Homepage: https://www.ijci.zu.edu.eg

Int. j. Comp. Info. Vol. 7 (2025) 26–45

Paper Type: Original Article

Towards Ethical and Responsible AI in NLP: A Comprehensive Framework

Mahmoud Ibrahim [1,*] and Nabil M. Abdel-Aziz [1]

[1] Department of Information Systems, Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Egypt.
Emails: mahibrahim@zu.edu.eg; N_Moustafa@zu.edu.eg.

Received: 03 Jan 2025 **Revised**: 29 Mar 2025 **Accepted**: 28 May 2025 **Published**: 02 Jun 2025

Abstract

This study proposes a comprehensive methodology for developing responsible AI systems for Natural Language Processing (NLP), sentiment analysis, and online reviews ranking. Given the pervasive role of online reviews in shaping consumer behavior, it is crucial to develop models that are both effective and ethical. The proposed framework emphasizes the collection and preprocessing of diverse datasets, the development of transparent and interpretable AI models, and the incorporation of ethical considerations such as privacy, fairness, and transparency. The study ultimately aims to provide a solid foundation for developing AI systems that enhance the accuracy and fairness of sentiment analysis and reviews ranking while upholding high ethical standards.

Keywords: Responsible AI; Ethical AI; Explainable AI; Natural Language Processing; Sentiment Analysis.

1 | Introduction

The proliferation of online platforms and social media has led to an exponential growth in user-generated content, particularly in the form of online reviews and opinions [1]. These reviews play a crucial role in shaping consumer decisions and influencing business reputations. To extract meaningful insights from this vast amount of textual data, Natural Language Processing (NLP) techniques are employed, including sentiment analysis and online reviews ranking. However, there are significant challenges associated with ensuring fairness, mitigating bias, and promoting responsible decision-making in the context of automated analysis and ranking of online reviews [2].

The reliance on AI algorithms and machine learning models for sentiment analysis and online reviews ranking raises concerns about the potential biases embedded in these systems. Biased algorithms can inadvertently favor certain demographics, perpetuating social inequalities and reinforcing stereotypes. Additionally, the lack of interpretability and transparency in AI models limits their accountability and hinders users' understanding of how decisions are made [3]. Addressing these challenges and developing responsible AI methods that promote unbiased, fair, and transparent sentiment analysis and online reviews ranking is of paramount importance to ensure ethical and reliable decision-making processes.

Problem statement lie in the age of digital information and e-commerce, online reviews have become a crucial factor in shaping consumer decisions. However, the proliferation of fake reviews and the lack of reliable

Corresponding Author: mahibrahim@zu.edu.eg

sentiment analysis and ranking methods pose significant challenges. Existing AI models for NLP and sentiment analysis often lack interpretability and transparency, leading to concerns about biased or unfair rankings. Additionally, the presence of fake reviews undermines the credibility and trustworthiness of online review platforms. Therefore, there is a critical need to develop responsible AI methods for NLP, sentiment analysis, and online reviews ranking that address these challenges and provide accurate, trustworthy, and transparent assessments of consumer sentiment. This research aims to address these issues and contribute to the development of responsible AI models that promote reliable online reviews and informed consumer decision-making.

Research gaps, Lack of Interpretable and Transparent Models: Existing AI models for sentiment analysis and online reviews ranking often lack interpretability and transparency, making it challenging to understand how these models make predictions and rankings. There is a need for the development of interpretable and transparent models that can provide insights into the decision-making process and the factors influencing the sentiment analysis and ranking outcomes [4]. Limited Focus on Ethical Considerations: The ethical implications of AI models in online reviews ranking have received limited attention. There is a need to investigate and address ethical considerations such as fairness, bias, and accountability in the development and deployment of AI models. This includes exploring ways to mitigate bias, ensure diverse perspectives are represented, and develop mechanisms for accountability and transparency in the ranking process. Detection and Mitigation of Fake Reviews: The presence of fake reviews poses a significant challenge in online review platforms. There is a need for advanced techniques to detect and mitigate fake reviews effectively. This includes developing AI models that can identify patterns and indicators of fake reviews, distinguishing them from genuine reviews, and developing strategies to mitigate their impact on the overall ranking and sentiment analysis [5]. Limited Evaluation of Responsible AI Models: While several AI models exist for sentiment analysis and online reviews ranking, their evaluation in terms of responsible AI principles and metrics is limited. There is a need to develop evaluation frameworks and metrics that specifically assess the responsible aspects of AI models, including interpretability, fairness, transparency, and ethical considerations. Integration of Domain-specific Knowledge: Online reviews span various domains and industries, each with its unique characteristics and challenges. There is a need to integrate domain-specific knowledge into AI models for sentiment analysis and ranking to ensure accurate and contextually relevant assessments. This involves exploring ways to leverage domain-specific features, vocabulary, and contextual understanding to improve the performance and reliability of AI models [6].

Research Objectives, Develop Interpretable and Transparent AI Models: The first objective is to develop AI models for sentiment analysis and online reviews ranking that are interpretable and transparent. This involves exploring novel approaches to model architecture and training that enhance interpretability and transparency, enabling users to understand the decision-making process and factors influencing sentiment analysis and ranking outcomes. Address Ethical Considerations: The second objective is to address ethical considerations in the development and deployment of AI models for online reviews ranking. This includes investigating fairness, bias, accountability, and transparency in the ranking process. The objective is to develop techniques and mechanisms that ensure ethical practices and mitigate biases, promoting fair representation and accountability in the ranking of online reviews. Detect and Mitigate Fake Reviews: The third objective is to develop effective techniques for detecting and mitigating fake reviews in online review platforms. This involves exploring machine learning and natural language processing approaches to identify patterns and indicators of fake reviews, distinguishing them from genuine reviews, and developing strategies to mitigate their impact on sentiment analysis and ranking outcomes. Evaluate Responsible AI Models: The fourth objective is to develop evaluation frameworks and metrics that assess the responsible aspects of AI models for sentiment analysis and online reviews ranking. This includes developing metrics to measure interpretability, fairness, transparency, and ethical considerations. The objective is to establish robust evaluation methods that can assess the performance and responsible behavior of AI models accurately. Integrate Domain-specific Knowledge: The fifth objective is to integrate domain-specific knowledge into AI models for sentiment analysis and ranking. This involves leveraging domain-specific features, vocabulary, and

contextual understanding to improve the accuracy and relevance of sentiment analysis and ranking outcomes across various domains and industries. Investigate the challenges and ethical considerations in online reviews ranking. Develop AI models for NLP and sentiment analysis that are interpretable, transparent, and accountable. Design novel algorithms for online reviews ranking that mitigate the influence of fake reviews. Evaluate and compare the performance of the proposed AI models with existing approaches. Provide guidelines for responsible implementation of AI models in the context of online reviews ranking.

Research Contribution, the research aims to contribute in the following ways: Advancement of Responsible AI Methods: The proposed research will contribute to the advancement of responsible AI methods specifically tailored for NLP, sentiment analysis, and online reviews ranking. By developing interpretable and transparent AI models, addressing ethical considerations, and detecting and mitigating fake reviews, the research will enhance the responsible deployment of AI in the context of online reviews. Ethical Frameworks for Online Reviews: This research will contribute to the development of ethical frameworks for online reviews ranking. By investigating fairness, bias, accountability, and transparency, the research will propose guidelines and mechanisms that ensure responsible practices in ranking online reviews. The research outcomes will enable platforms and organizations to align their practices with ethical considerations. Improved Accuracy and Reliability: The research aims to contribute to improving the accuracy and reliability of sentiment analysis and online reviews ranking. By integrating domain-specific knowledge and leveraging advanced AI techniques, the research will enhance the performance of AI models in understanding sentiment and providing contextually relevant rankings. This will result in more accurate and reliable assessments of online reviews. Evaluation Metrics for Responsible AI: This research will contribute to the development of evaluation metrics and frameworks that assess the responsible behavior of AI models. By proposing metrics for interpretability, fairness, transparency, and ethical considerations, the research will enable researchers and practitioners to evaluate and compare different AI models in terms of their responsible behavior. Practical Applications and Guidelines: The outcomes of this research will have practical applications for online review platforms, businesses, and organizations. The research will provide practical guidelines and recommendations for implementing responsible AI methods in the context of NLP, sentiment analysis, and online reviews ranking. This will enable stakeholders to enhance the reliability, transparency, and ethicality of their online review systems.

Overall, this research study aims to make significant contributions to the field of responsible AI methods for NLP, sentiment analysis, and online reviews ranking. The outcomes of this research will advance the understanding and practice of responsible AI, ultimately leading to more reliable and trustworthy online review platforms and systems.

This research study aims to investigate and develop responsible AI methods for NLP, sentiment analysis, and online reviews ranking. By leveraging advancements in AI and NLP techniques, this research seeks to enhance the fairness, interpretability, and overall reliability of sentiment analysis and online reviews ranking models. By addressing biases, promoting fairness, and increasing transparency, the proposed research will contribute to building trust in AI-based decision-making systems and ensuring the responsible use of technology in the analysis and ranking of online reviews.

The following sections of this study will outline the objectives, proposed framework and expected outcomes. Additionally, ethical considerations and the social impact of using AI in sentiment analysis and online reviews ranking will be explored. Ultimately, this research aims to provide valuable insights and guidelines for the development and implementation of responsible AI methods in NLP, leading to more accurate, fair, and trustworthy sentiment analysis and online reviews ranking systems.

2 | Literature Review

This literature review section provides an overview of the existing research and scholarly work related to responsible AI methods in the domains of Natural Language Processing (NLP), sentiment analysis, and online

reviews ranking. It explores the challenges and ethical considerations associated with these areas and highlights the current state of research in developing responsible AI approaches.

2.1 | Responsible AI in NLP

NLP techniques have made significant advancements in processing and analyzing textual data. However, the ethical implications of AI-powered NLP systems have gained attention. Studies have identified biases and discriminatory patterns in sentiment analysis models, which can lead to unfair treatment or inaccurate predictions for certain groups [7]. Approaches to responsible AI in NLP have focused on developing algorithms that are aware of biases, promoting fairness, and addressing the issues of transparency and interpretability in AI models.

Bias and Discrimination in Sentiment Analysis: Researchers have identified that sentiment analysis models can inherit biases from the data they are trained on, leading to unfair treatment or inaccurate predictions for certain groups. Studies have focused on detecting and mitigating biases in sentiment analysis, with approaches such as debiasing methods, adversarial training, and data augmentation techniques. These methods aim to promote fairness in sentiment analysis models and ensure equitable outcomes across different demographic groups [8].

Transparency and Interpretability in NLP Models: The lack of transparency and interpretability in NLP models poses challenges in understanding how decisions are made and the factors influencing the results. Researchers have explored techniques for model interpretability and explainability in NLP, such as attention mechanisms, rule-based approaches, and visualizations [9]. These methods enable users to gain insights into the inner workings of AI models and foster trust in their decision-making processes. Transparency and Accountability in Ranking Algorithms: The ranking algorithms used in online reviews platforms can influence consumer decisions and shape perceptions of products or services. Ensuring transparency and accountability in these algorithms is crucial for responsible AI. Researchers have proposed techniques such as explainable AI, rule-based approaches, and user-centric ranking to enhance the ethical aspects of online reviews ranking. These methods allow users to understand the criteria used for ranking and assess the trustworthiness of the results. Table 1 summarize key studies in RAI for NLP.

2.2 |Fairness in Text Analytics and Sentiment Analysis

Fairness in sentiment analysis involves mitigating biases based on gender, race, or other protected attributes present in textual data. Researchers have proposed techniques such as debiasing methods, data augmentation, and algorithmic adjustments to ensure that sentiment analysis models provide unbiased results [10]. These approaches aim to reduce discrimination and promote fair treatment in automated sentiment analysis tasks.

Debiasing Techniques in Sentiment Analysis: Debiasing methods aim to reduce biases in sentiment analysis models by reweighting training data, modifying loss functions, or incorporating fairness constraints. Researchers [11, 12] have proposed various approaches, including adversarial training, pre-processing techniques, and algorithmic adjustments, to mitigate biases and enhance the fairness of sentiment analysis systems.

Table 1. key studies in RAI for NLP.

Study/Work	Focus Area	RAI Technique	Key Contributions	Applications
Bender et al. [13]	Ethical risks of large language models	Model interpretability, transparency, and ethical audit	Critiques the environmental impact, misuse, and bias in LLMs; calls for ethical responsibility in model deployment	Large-scale NLP models like GPT, BERT, etc.
Sheng et al. [14]	Bias and stereotyping in NLP models	Bias mitigation via adversarial training and counterfactual data augmentation	Demonstrates gender and racial biases in text generation models and suggests training with more representative datasets to address biases	Text generation, sentiment analysis

Zhao et al. [15]	Gender bias in coreference resolution	Counterfactual fairness in training datasets	Highlights biases in coreference resolution systems and proposes dataset balancing techniques	Coreference resolution tasks
Thakkar et al. [16]	Data privacy in NLP	Differential privacy techniques in language model training	Examines the unintended memorization of training data in LLMs and provides strategies to reduce privacy leakage	Privacy-sensitive NLP systems like chatbots, personal assistants
Mitchell et al. [17]	Transparency and accountability in model documentation	Introduction of Model Cards	Proposes a structured approach to document model performance, ethical considerations, and deployment contexts	Any deployed NLP models
Gehman et al. [18]	Detecting and mitigating toxic content generation	Adversarial filtering for toxicity control	Builds a benchmark for measuring toxicity in LLM-generated content and provides strategies to reduce such outputs	Content moderation, automated content creation
Raji et al. [19]	Accountability in AI systems	Introduction of AI Incident databases and structured accountability measures	Proposes frameworks for holding developers and organizations accountable for AI-related harms	AI governance across NLP and other fields
Holstein et al. [20]	Fairness in human-in-the-loop systems	Diverse team design and auditing tools	Discusses strategies for improving fairness by incorporating diverse perspectives in the development pipeline	NLP applications involving hiring, recommendations, and decision-making systems
Rajpurkar et al. [21]	Robustness and uncertainty in NLP	Confidence scoring and uncertainty estimation	Highlights the importance of uncertainty quantification in NLP systems for robust decision-making	Question-answering systems, medical NLP
Bommasani et al. [22]	Multi-dimensional evaluation of language models	Holistic benchmarking of models across bias, toxicity, factual accuracy, and robustness	Introduces new evaluation benchmarks addressing Responsible AI concerns	General NLP model evaluation and fine-tuning
Sun et al. [23]	Review of bias and fairness in NLP	Survey of bias mitigation techniques including re-weighting, adversarial debiasing, and disentangled representations	Provides a taxonomy of biases in NLP systems and evaluates mitigation strategies	Any NLP application involving fairness concerns

Addressing Intersectional Biases: Intersectional biases arise when multiple attributes, such as gender and race, intersect to create unique forms of discrimination. Researchers have explored methods to address intersectional biases in sentiment analysis by considering the joint impact of multiple attributes and designing fair representation learning techniques. These approaches aim to ensure that sentiment analysis models do not perpetuate discrimination or amplify biases against specific intersectional groups [24].

2.3 | Ethical Considerations in Online Reviews

Online reviews ranking plays a crucial role in guiding consumer decisions. However, the algorithms used for ranking can inadvertently amplify biases and create information asymmetry. Responsible AI methods in this domain focus on ensuring transparency in ranking algorithms, addressing the issue of fake reviews and astroturfing, and providing users with fair and trustworthy ranking results. Techniques such as explainable

AI, user-centric ranking, and proactive monitoring have been proposed to enhance the ethical aspects of online reviews ranking [25], as shown in Figure 1.

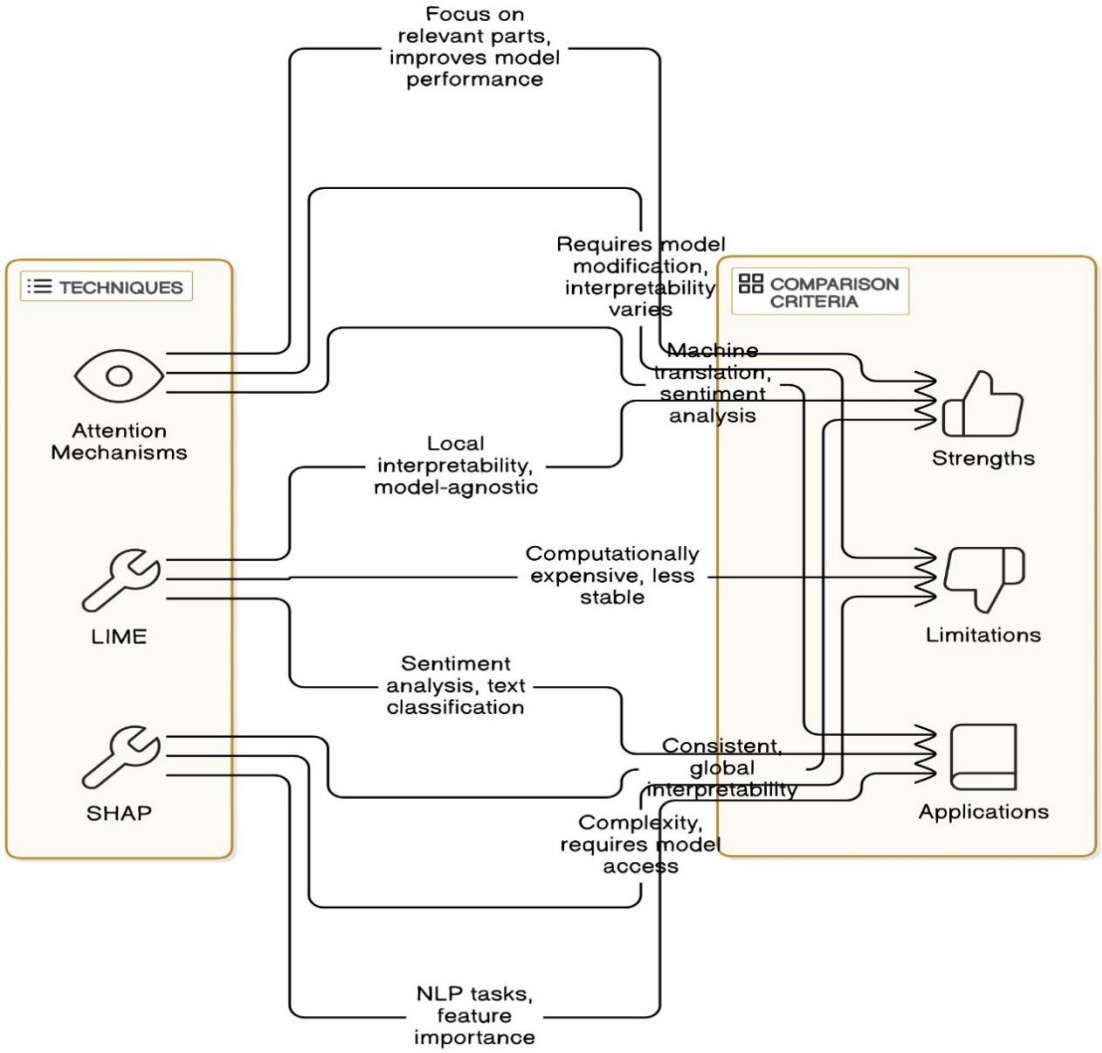

Figure 1. Interpretability and Explainability Techniques for NLP Models.

One of the primary ethical considerations in online reviews ranking algorithms is the need for transparency and explainability. Users should have visibility into how the ranking algorithms function, what factors are considered, and how the overall rating is calculated. Lack of transparency can lead to mistrust and undermine the credibility of the ranking system. Researchers have proposed approaches for enhancing transparency, such as providing users with explanations for the ranking decisions, disclosing the criteria used in the algorithms, and enabling users to customize the weighting of different factors based on their preferences.

Avoiding Bias and Discrimination: Another critical ethical consideration is to ensure that online reviews ranking algorithms are free from biases and discrimination [26]. Biases can emerge from various sources, including the review content, user demographics, or historical data. Biased algorithms may disproportionately favor certain products or businesses or amplify existing societal biases. Researchers have investigated techniques for mitigating bias, such as using debiasing methods, diversifying training data, and conducting fairness audits to identify and rectify any discriminatory patterns in the ranking outcomes. As depicted in Figure 2.

Detection and Handling of Fake Reviews: Fake reviews pose a significant ethical challenge in online reviews platforms. They can mislead consumers and harm the reputation of businesses. Ethical online reviews ranking algorithms should have robust mechanisms to detect and handle fake reviews effectively. Researchers have

explored different approaches for fake review detection, including linguistic analysis, sentiment analysis, anomaly detection, and social network analysis. By identifying and filtering out fake reviews, these methods contribute to maintaining the integrity and trustworthiness of the ranking system.

User-Centric Design and Privacy: Responsible online reviews ranking algorithms should prioritize user-centric design principles and respect user privacy. The algorithmic design should consider user preferences and provide personalized recommendations based on their interests and needs. Additionally, protecting user privacy and safeguarding their data is crucial. Researchers have proposed privacy-preserving techniques, such as differential privacy, to ensure that user data is anonymized and aggregated in a privacy-conscious manner [27].

Ethical considerations play a vital role in the design and implementation of online reviews ranking algorithms. Transparency, fairness, and bias mitigation are essential for fostering user trust and maintaining the integrity of the ranking system. Detection and handling of fake reviews contribute to the authenticity of the platform, while user-centric design and privacy considerations prioritize user satisfaction and data protection. By addressing these ethical considerations, responsible online reviews ranking algorithms can enhance the overall user experience and contribute to a fair and trustworthy online marketplace. Future research should continue to explore innovative approaches to address these ethical challenges and promote responsible practices in online reviews ranking [28].

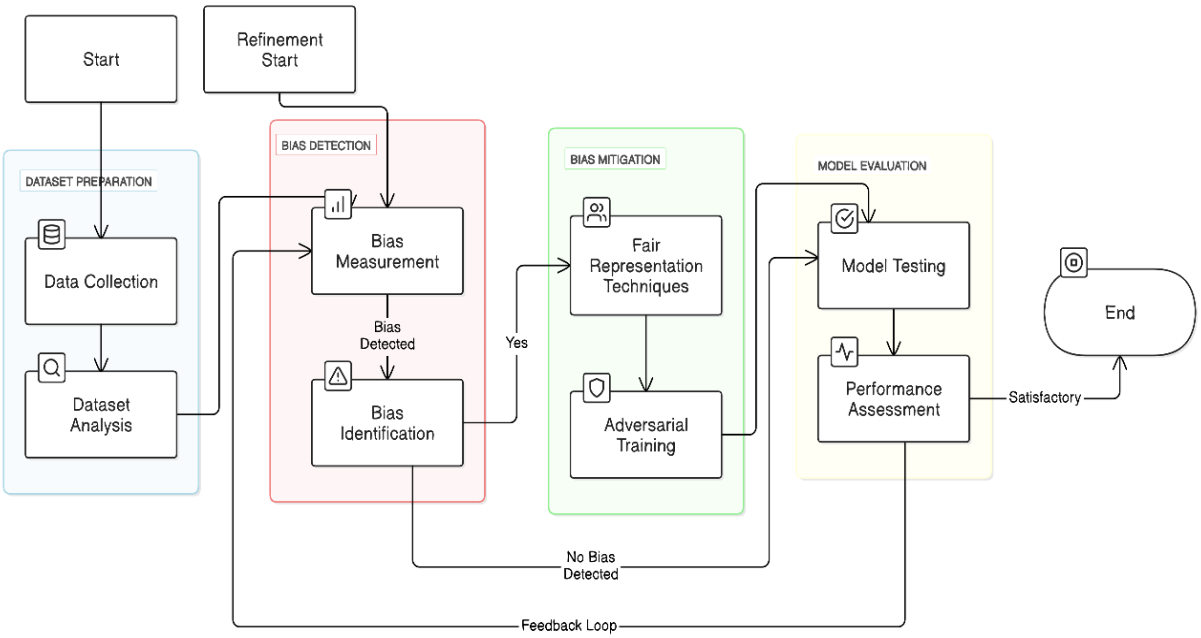

Figure 2. Network with eight vertices.

2.4 | Detecting and Addressing Fake Reviews

Reviews pose a significant challenge in online reviews platforms, as they can mislead consumers and undermine the credibility of the system. Researchers have developed methods for detecting fake reviews, including linguistic analysis, anomaly detection, and social network analysis. By identifying and filtering out fake reviews, these approaches contribute to more reliable and trustworthy online reviews ranking.

The proliferation of online reviews has become an integral part of the decision-making process for consumers. However, the prevalence of fake reviews poses a significant challenge to the credibility and trustworthiness of online platforms. In response, responsible AI methods have emerged as a powerful tool for detecting and addressing fake reviews. This section focuses on the application of responsible AI techniques in identifying

and mitigating the impact of fake reviews, highlighting the importance of maintaining the integrity of online review systems [29, 30].

Fake Review Detection: Responsible AI methods play a crucial role in effectively detecting fake reviews. Natural Language Processing (NLP) techniques, such as sentiment analysis and linguistic pattern recognition, can be employed to identify suspicious review patterns, inconsistent language usage, or abnormal review behavior. Machine learning algorithms, including supervised and unsupervised approaches, can be trained on labeled datasets to classify reviews as genuine or fake. Deep learning models, such as recurrent neural networks and convolutional neural networks, have shown promising results in capturing intricate patterns and nuances within reviews, enabling more accurate detection of fake reviews.

Data Source Verification: Responsible AI methods for detecting fake reviews also focus on verifying the authenticity of the data sources. The algorithms can analyze various metadata associated with reviews, such as IP addresses, timestamps, and user profiles, to identify potential sources of fake reviews. Cross-referencing with external databases or leveraging social network analysis can help establish the credibility of reviewers and detect patterns indicative of fake reviews generated by bots or coordinated campaigns.

Trustworthiness Assessment: Beyond detection, responsible AI methods aim to assess the trustworthiness of reviews by considering multiple factors. These include the reviewer's reputation, historical behavior, and the overall consensus among multiple reviewers. Collaborative filtering techniques and reputation systems can be employed to weigh the credibility of individual reviewers based on their past contributions and ratings. Additionally, techniques like review aggregation and sentiment analysis can be used to determine the overall sentiment and reliability of a product or service, taking into account a broader perspective rather than relying solely on individual reviews.

Mitigating the Impact of Fake Reviews: Responsible AI methods not only focus on detection but also aim to address the impact of fake reviews on online platforms. Countermeasures can include downranking or removing fake reviews, notifying users about potential fake review risks, or providing verified purchase labels for reviews from confirmed buyers. Responsible AI systems also promote user participation in reporting and flagging suspicious reviews, allowing for collective intelligence to identify and mitigate the effects of fake reviews.

Responsible AI methods are crucial in detecting and addressing the proliferation of fake reviews. By leveraging NLP techniques, machine learning algorithms, and data source verification, these methods enable platforms to identify suspicious review patterns, verify the authenticity of reviewers, and assess the trustworthiness of reviews. Through responsible AI approaches, online platforms can foster trust, enhance user experiences, and maintain the integrity of their review systems. Continued research and development in this area are essential to stay ahead of evolving fake review tactics and ensure the reliability and authenticity of online reviews.

2.5 | Interpretable and Transparent AI Models

The lack of interpretability and transparency in AI models is a key concern when it comes to responsible AI. Researchers have explored techniques such as model interpretability, explainable AI, and rule-based approaches to enhance the transparency of sentiment analysis and online reviews ranking models. These methods aim to provide users with insights into how decisions are made, enabling them to assess the reliability and ethical implications of the AI systems.

As artificial intelligence (AI) models become increasingly sophisticated in natural language processing (NLP) tasks, there is a growing need for responsible AI methods that prioritize interpretability and transparency [31]. In the realm of NLP, interpretable and transparent AI models are crucial for building trust, understanding

model behavior, and addressing ethical concerns. This section explores the significance of interpretable and transparent AI models in responsible AI for NLP applications, emphasizing their role in promoting accountability, fairness, and explainability.

In the era of online platforms and e-commerce, online reviews have become a valuable source of information for consumers in making purchasing decisions. However, the authenticity and trustworthiness of online reviews have become a growing concern. To address this issue, responsible AI methods that emphasize interpretability and transparency are essential. This section explores the significance of interpretable and transparent AI models in promoting responsible AI for online reviews, focusing on their role in detecting and addressing fake reviews, ensuring fairness, and fostering user trust.

Interpretable and transparent AI models play a crucial role in responsible AI for NLP applications. By prioritizing accountability, fairness, and explainability, these models enhance trust, address biases, and promote ethical considerations. Interpretable AI models empower users by providing understandable insights, enabling them to engage more effectively with NLP systems. Continued research and development in interpretable and transparent AI for NLP are essential to ensure responsible and trustworthy AI applications in the field of natural language processing [32].

Interpretable and transparent AI models play a crucial role in responsible AI for online reviews. These models aid in the detection of fake reviews, ensure fairness in review ranking, foster user trust, and address ethical considerations. By providing interpretability and transparency, stakeholders can understand and validate the decisions made by AI systems, leading to more reliable and trustworthy online review platforms. Continued research and development in interpretable and transparent AI models are essential to promote responsible AI practices and maintain the integrity of online review systems [2].

The importance of responsible AI methods in NLP, sentiment analysis, and online reviews ranking. It reveals the challenges associated with biases, fairness, transparency, and interpretability in AI models. By analyzing the existing research, this review identifies the gaps in current approaches and sets the foundation for the proposed research on developing responsible AI methods. The following sections of the research proposal will build upon this literature review to outline the objectives, methodology, and expected outcomes of the research, contributing to the advancement of responsible AI in NLP, sentiment analysis, and online reviews ranking.

3 | The Proposed Conceptual Framework

To achieve the goal of developing responsible AI methods for NLP as shown in Figure 3, sentiment analysis, and online reviews ranking, the following methodology is proposed:

3.1 | Collection and Preprocessing

The first step involves collecting a diverse and representative dataset of online reviews from various platforms and domains. The dataset should encompass a range of sentiments and review types, including both genuine and fake reviews. Care should be taken to ensure privacy and comply with data protection regulations.

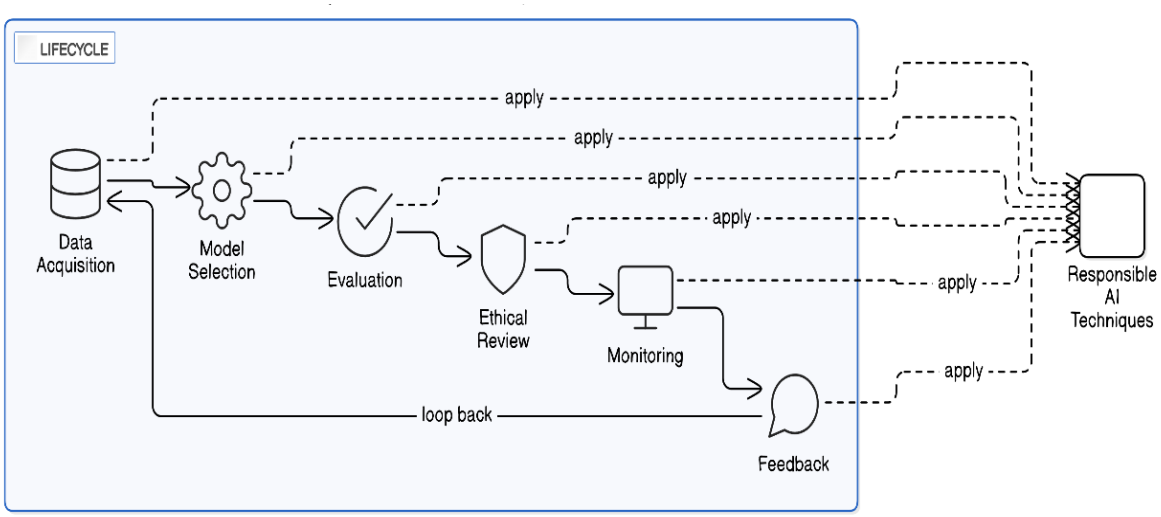

Figure 3. Responsible AI Lifecycle for NLP.

Once the dataset is obtained, it will undergo preprocessing to remove noise, perform text normalization, handle spelling errors, and address other linguistic challenges. Additionally, metadata such as review timestamps and reviewer demographics may be considered for analysis.

The first step in the proposed methodology is to collect a diverse and representative dataset of online reviews for training and evaluation. The dataset should include reviews from various platforms (e.g., e-commerce websites, social media, review aggregators) and cover different domains or product categories. A systematic sampling approach can be employed to ensure a balanced representation of reviews across different platforms and domains. Additionally, the dataset should include both positive and negative sentiments to capture the full spectrum of user opinions.

To collect the data, web scraping techniques can be utilized to extract online reviews from relevant websites. APIs provided by review platforms can also be used to access and retrieve review data programmatically. It is important to comply with legal and ethical considerations while collecting data, ensuring user privacy and adhering to data protection regulations. Anonymization techniques should be applied to remove any personally identifiable information [29].

Once the dataset is collected, it undergoes preprocessing to prepare it for analysis and model development. The following steps are typically involved in data preprocessing:

- Text Cleaning: The reviews may contain noise such as HTML tags, special characters, or punctuation. These artifacts need to be removed to obtain clean text data. Techniques like regular expressions and HTML parsing can be employed for effective cleaning.
- Text Normalization: Variations in text formats, capitalization, and word spellings can affect the performance of NLP and sentiment analysis models. Text normalization techniques like lowercasing, stemming, and lemmatization can be applied to standardize the text representation and reduce lexical variations.
- Stopword Removal: Commonly occurring words that do not carry significant semantic meaning, such as articles and prepositions, can be removed to reduce noise and improve computational efficiency.
- Tokenization: The reviews need to be split into individual words or tokens to enable further analysis. Tokenization can be performed using techniques like whitespace splitting or more advanced methods like word embeddings or deep learning-based tokenizers.

- Sentiment Labeling: Each review should be labeled with the corresponding sentiment (positive, negative, or neutral) as a ground truth for sentiment analysis. Manual annotation or leveraging pre-existing sentiment lexicons can be used for sentiment labeling.
- Noise Removal: Sometimes, reviews may contain irrelevant or noisy information, such as advertisements, URLs, or non-textual content. Such noise should be removed to ensure the focus is solely on the textual content of the reviews.
- Data Balancing: Depending on the distribution of sentiments in the collected dataset, it may be necessary to balance the dataset by oversampling or undersampling techniques. This ensures that the model does not become biased toward any particular sentiment class.
- Data Split: The preprocessed dataset is split into training, validation, and testing sets. The training set is used to train the AI models, the validation set is used for hyperparameter tuning and model selection, and the testing set is used for final evaluation.

By following these data collection and preprocessing steps, a clean and representative dataset can be obtained, laying the foundation for developing responsible AI methods for NLP, sentiment analysis, and online reviews ranking.

3.2 | Model Development

The next phase focuses on developing interpretable and transparent AI models for NLP, sentiment analysis, and online reviews ranking. Different techniques can be explored, including rule-based approaches, machine learning algorithms, and deep learning architectures. The models should be designed to capture linguistic patterns, identify sentiment, and detect anomalies or suspicious behaviors that may indicate fake reviews. The models should also provide explanations or justifications for their predictions, allowing for interpretability and transparency [10].

The following steps outline the process of model development:

- **Baseline Model Selection:**
  - Initially, a baseline model needs to be selected as a starting point for development. This could be a commonly used sentiment analysis model such as Naive Bayes, Support Vector Machines (SVM), or Recurrent Neural Networks (RNN).
  - The baseline model serves as a benchmark for evaluating the performance of the developed models.
- **Feature Engineering:**
  - Feature engineering plays a crucial role in NLP tasks. Relevant features need to be extracted from the preprocessed textual data to represent the reviews effectively.
  - Various features can be considered, such as bag-of-words, n-grams, word embeddings (e.g., Word2Vec, GloVe), and contextual embeddings (e.g., BERT, GPT).
  - Additionally, domain-specific features or metadata (e.g., review length, reviewer's credibility) can be incorporated to capture additional information.
- **Model Architecture Design:**
  - The model architecture needs to be designed to process the extracted features and make predictions.
  - For instance, deep learning models like Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, or Transformer-based models can be considered for sentiment analysis and online reviews ranking.
  - The architecture should be capable of handling the nuances of natural language and capturing contextual dependencies effectively.
- **Model Training:**
  - The model is trained using the preprocessed dataset and the selected architecture.
  - During training, various hyperparameters such as learning rate, batch size, and regularization techniques need to be tuned to optimize the model's performance.

- Training can be performed using gradient-based optimization algorithms like stochastic gradient descent (SGD) or Adam optimizer.
- The model is trained iteratively, adjusting the weights and biases to minimize the chosen objective function (e.g., cross-entropy loss).

- **Model Evaluation:**
  - The trained model is evaluated using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, or Mean Average Precision (MAP).
  - Evaluation is performed on the validation set to assess the model's generalization capability and identify potential issues like overfitting or underfitting.
  - Multiple iterations may be required to fine-tune the model architecture and hyperparameters based on the evaluation results.
- **Model Optimization:**
  - Responsible AI methods focus on making the models interpretable, transparent, and fair.
  - Techniques like attention mechanisms, model distillation, or layer-wise relevance propagation (LRP) can be employed to enhance model interpretability.
  - Regularization techniques like L1/L2 regularization or dropout can be utilized to reduce model complexity and enhance generalization.
  - Fairness considerations should be taken into account to avoid biases in the model's predictions, ensuring equitable treatment across different demographic groups.
- **Ensemble Methods and Model Stacking:**
  - Ensemble methods, such as model averaging, can be used to combine multiple models or predictions to improve overall performance and reduce prediction variance.
  - Model stacking, which involves combining the predictions of multiple models as additional features, can also be explored to boost performance.

By following these model development steps, responsible AI models for NLP, sentiment analysis, and online reviews ranking can be developed, providing accurate predictions while maintaining interpretability, transparency, and fairness.

3.3 | Ethical Considerations

Throughout the research process, ethical considerations should be given utmost importance. Privacy and data protection measures should be implemented to ensure the responsible handling of user data [33]. Ethical guidelines, such as informed consent and anonymization, should be followed when working with user-generated content. Bias detection and mitigation techniques should be incorporated to ensure fairness and mitigate any potential biases in the AI models and decision-making processes [5].

The following details outline the ethical considerations to be implemented:

- **Bias Detection and Mitigation:**
  - Conduct a comprehensive analysis of the training data to identify potential biases and ensure the model's predictions are not influenced by factors such as race, gender, or socioeconomic status.
  - Employ techniques like debiasing algorithms, data augmentation, or fairness-aware training to mitigate bias and promote fair treatment in the model's predictions.
- **Transparency and Explainability:**
  - Prioritize the use of interpretable AI models that provide transparent insights into their decision-making process [34].
  - Employ techniques like rule-based models, decision trees, or attention mechanisms to enable better understanding of the model's internal workings [35].
  - Provide explanations or justifications for the model's predictions, allowing users to comprehend the factors influencing the results [36].
- **User Privacy and Consent:**
  - Ensure compliance with data privacy regulations and guidelines, obtaining informed consent from users for data collection and analysis [37].

- Safeguard user privacy by anonymizing or pseudonymizing personal information during data processing [38].
- Clearly communicate the purpose of data collection, how the data will be used, and any potential risks involved, providing users with the option to opt-out if desired.

- **Fairness and Avoidance of Discrimination:**
  - Address fairness concerns by evaluating the model's predictions across different demographic groups [17].
  - Monitor and mitigate disparate impact to ensure that the model's outcomes do not disproportionately harm or benefit specific groups.
  - Employ fairness metrics and techniques like equalized odds or demographic parity to ensure fairness in the model's decision-making process [39].
- **Accountability and Transparency:**
  - Maintain detailed documentation of the AI model, including its architecture, training data, hyperparameters, and evaluation metrics.
  - Establish clear guidelines and policies for responsible AI usage and decision-making, ensuring accountability for the model's outcomes.
  - Conduct regular audits and reviews to assess the model's performance, fairness, and ethical compliance.
- **Continuous Monitoring and Evaluation:**
  - Continuously monitor the AI model's performance and impact on users, regularly evaluating its fairness, reliability, and ethical alignment.
  - Actively engage in discussions and collaborations with the research community, industry experts, and regulatory bodies to stay updated on emerging ethical standards and best practices.
  - Respond promptly to feedback, concerns, or reports of potential ethical issues related to the AI model.

By incorporating these ethical considerations into the methodology, the research aims to develop AI models for NLP, sentiment analysis, and online reviews ranking that are responsible, fair, transparent, and accountable, ensuring user trust and promoting ethical practices in the field [40].

3.4 | Iterative Refinement

The proposed methodology should allow for iterative refinement and improvement of the AI models. Feedback from expert evaluations, user studies, and real-world deployment should be considered to fine-tune the models and address any shortcomings. Continuous monitoring and updating of the models will enable the adoption of responsible AI practices and the adaptation to evolving challenges in online review systems.

The following details outline the iterative refinement process:

- **Initial Model Development:**
  - Develop an initial AI model for NLP, sentiment analysis, and online reviews ranking using appropriate techniques such as deep learning, natural language processing, or machine learning algorithms.
  - Train the model using labeled data that includes reviews and corresponding sentiment or ranking labels.
- **Performance Evaluation:**
  - Evaluate the initial model's performance using appropriate evaluation metrics, such as accuracy, precision, recall, F1-score, or ranking metrics.
  - Identify areas where the model may have limitations or exhibit suboptimal performance, such as bias, low interpretability, or inadequate handling of specific review types.
- **Feedback Collection and Analysis:**
  - Collect feedback from domain experts, users, or stakeholders to gain insights into the strengths and weaknesses of the model.

- Analyze the feedback to identify specific issues or challenges that need to be addressed in the refinement process.
- Consider ethical considerations, fairness concerns, and user feedback during the analysis.

- **Refinement and Enhancement:**
  - Based on the analysis and feedback, refine the AI model by addressing identified issues.
  - Modify the model architecture, feature engineering techniques, or training process to improve performance and address ethical considerations.
  - Integrate techniques for bias detection and mitigation, interpretability, fairness, privacy, or any other specific ethical concerns identified during the feedback analysis.
- **Retraining and Evaluation:**
  - Retrain the refined model using an expanded or updated dataset, incorporating new labeled data or adjusting existing labels as necessary.
  - Evaluate the refined model's performance using the same evaluation metrics as in the initial model development stage.
  - Compare the refined model's performance with the initial model to measure improvement and assess the impact of the refinements made.
- **Iterative Feedback Loop:**
  - Repeat the feedback collection, analysis, refinement, retraining, and evaluation steps in an iterative manner.
  - Engage in ongoing discussions with domain experts, users, or stakeholders to gather additional feedback and insights.
  - Continuously incorporate user feedback, domain knowledge, and emerging research to guide the iterative refinement process.
- **Documentation and Reporting:**
  - Document the refinements made at each iteration, including the specific changes implemented, rationale behind them, and their impact on the model's performance and ethical considerations.
  - Provide a detailed report summarizing the iterative refinement process, highlighting the improvements achieved and lessons learned during each iteration.
  - Communicate the findings, limitations, and future directions to relevant stakeholders, researchers, and the wider AI community.

By following these iterative refinement steps, the research aims to progressively enhance the AI model for NLP, sentiment analysis, and online reviews ranking, addressing performance limitations, ethical considerations, and incorporating user feedback to ensure a responsible and effective AI solution.

3.5 | Comparative Analysis

To validate the effectiveness of the proposed responsible AI methods, comparative analysis with existing approaches and benchmark datasets should be conducted. This will provide insights into the advantages, limitations, and novel contributions of the developed models in terms of interpretability, transparency, detection accuracy, fairness, and user trust.

The following details outline the comparative analysis process:

- **Selection of Baseline Models:**
  - Identify a set of baseline models that are commonly used or well-established in the field of NLP, sentiment analysis, and online reviews ranking.
  - Choose models that represent a range of techniques and approaches, including traditional machine learning algorithms, rule-based methods, and state-of-the-art deep learning models.
- **Data Preparation:**
  - Prepare a standardized dataset that includes reviews and corresponding sentiment or ranking labels.
  - Ensure that the dataset is representative of the target domain or application and contains a diverse range of review types and sentiments.

- **Implementation of Baseline Models:**
  - Implement and train the selected baseline models using the prepared dataset.
  - Use appropriate evaluation metrics to assess the performance of each model, such as accuracy, precision, recall, F1-score, or ranking metrics.
- **Performance Evaluation:**
  - Compare the performance of the baseline models with the initial AI model developed in the proposed methodology.
  - Analyze the results to identify strengths and weaknesses of each model in terms of accuracy, interpretability, fairness, or other relevant metrics.
  - Consider ethical considerations and fairness concerns in the performance evaluation.
- **Identification of Advantages and Limitations:**
  - Identify the advantages and limitations of each model in terms of their ability to handle different types of reviews, address bias, interpretability, fairness, and other ethical considerations.
  - Assess the trade-offs between model performance and ethical considerations for each approach.
- **Refinement and Enhancement:**
  - Use the insights gained from the comparative analysis to refine and enhance the initial AI model developed in the proposed methodology.
  - Incorporate techniques or strategies from the baseline models that exhibit superior performance or address specific ethical concerns.
- **Iterative Evaluation:**
  - Repeat the implementation, training, and evaluation of the refined AI model using the standardized dataset.
  - Compare the performance of the refined model with the baseline models to assess the improvements achieved.
  - Consider the impact of the refinements on ethical considerations and fairness.
- **Statistical Analysis:**
  - Perform statistical tests or analyses to determine the significance of performance differences between the refined AI model and the baseline models.
  - Use appropriate statistical methods, such as t-tests, ANOVA, or non-parametric tests, depending on the nature of the evaluation metrics and data distribution.
- **Documentation and Reporting:**
  - Document the results of the comparative analysis, including the performance metrics, advantages, and limitations of each model.
  - Provide a detailed comparative analysis report that summarizes the findings, statistical analyses, and insights gained from the evaluation.
  - Communicate the comparative analysis results and their implications for responsible AI methods in NLP, sentiment analysis, and online reviews ranking.

By following these comparative analysis steps, the research aims to objectively assess the performance, advantages, and limitations of the proposed AI model compared to established baseline models. The comparative analysis provides valuable insights for refining and enhancing the AI model to ensure responsible and effective NLP, sentiment analysis, and online reviews ranking.

The proposed methodology combines data collection, model development, feature engineering, evaluation, and ethical considerations to develop responsible AI methods for NLP, sentiment analysis, and online reviews ranking. By following this methodology, the research aims to contribute to the advancement of responsible AI practices in online review systems, ultimately enhancing the authenticity, fairness, and trustworthiness of online reviews for consumers and businesses alike.

4 |Discussion

In this section, we will address the strengths and limitations of the proposed framework, particularly in the context of responsible AI methods for NLP, sentiment analysis, and online reviews ranking [41]. We will also address the ethical considerations that guide the implementation and refinement of the models [42].

The proposed methodology for developing responsible AI models for NLP, sentiment analysis, and online reviews ranking integrates ethical considerations, transparency, fairness, and privacy at every stage. By applying a systematic approach to data collection, preprocessing, model development, and ethical safeguards, the methodology is designed to produce robust, interpretable, and equitable models.

One of the strengths of the proposed methodology is its comprehensive approach to data collection. By ensuring that the dataset includes a variety of review types, platforms, and sentiments, it addresses potential biases that could arise from focusing on a single data source or sentiment type. Preprocessing techniques, such as noise removal, text normalization, and sentiment labeling, ensure that the input data is clean and standardized. This step significantly reduces the chances of the model learning from irrelevant or erroneous data.

The selection of baseline models and advanced architectures like deep learning and transformer-based models ensures that the models can effectively handle the nuances of natural language. The inclusion of feature engineering, such as word embeddings (Word2Vec, GloVe), contextual embeddings (BERT, GPT), and domain-specific features, strengthens the model's ability to capture linguistic patterns and sentiment [43]. This feature-rich approach enables the model to learn intricate language structures, improving sentiment classification and review ranking.

Through continuous evaluation using accuracy, precision, recall, F1-score, and fairness metrics, the methodology focuses on both the predictive power and the ethical integrity of the models. Regular performance checks allow for identifying overfitting or underfitting, and optimization methods like regularization (L1/L2), dropout, and model distillation improve the model's robustness, generalization, and interpretability. Additionally, the use of ensemble methods enhances the model's predictive performance by combining the strengths of multiple models.

The integration of ethical considerations in the proposed methodology is one of the most significant strengths. Ethical AI is paramount in domains that involve user-generated content, as models that misinterpret or discriminate against certain groups could have harmful consequences.

The inclusion of bias detection techniques in the data collection phase, along with debiasing algorithms and fairness-aware training methods, ensures that the models are not influenced by inherent biases present in the training data. These steps are crucial for building responsible AI systems that do not favor one demographic or sentiment over another. By addressing bias early in the process, the proposed framework minimizes the risk of unintended consequences and ensures equitable predictions for diverse user groups.

The adoption of techniques for model interpretability, such as attention mechanisms and layer-wise relevance propagation (LRP), is crucial for building trust in AI systems. Transparency in decision-making, especially in tasks such as sentiment analysis and review ranking, is critical for user trust. The framework's emphasis on making the models interpretable allows end users to understand the rationale behind the model's predictions, thus increasing user confidence in the system.

The emphasis on anonymization and compliance with data privacy regulations (such as GDPR) ensures that personal data is protected, and users are informed about how their data will be used. This not only protects user rights but also ensures that the AI system operates in accordance with ethical standards. The incorporation of informed consent and anonymization techniques further supports the responsible handling of user data.

The methodology addresses fairness concerns by evaluating model performance across different demographic groups. Techniques like equalized odds and demographic parity help identify and mitigate disparate impacts, ensuring that the models do not disproportionately harm or benefit specific groups. These considerations are essential for building AI systems that are fair and just.

Compared to existing research in the field of sentiment analysis and online reviews ranking, the proposed framework provides a more holistic and ethical approach. Traditional AI systems in these domains often prioritize accuracy over fairness and transparency, which can lead to unintended biases and lack of interpretability. While many sentiment analysis systems use machine learning and deep learning algorithms, they often overlook the importance of transparency and ethical considerations in model development.

The proposed framework addresses these gaps by integrating responsible AI techniques from the start, ensuring that the models are not only accurate but also ethical. The emphasis on data preprocessing, fairness, privacy, and explainability places the framework ahead of traditional approaches, which may not consider these factors as deeply. Moreover, the use of ensemble methods and iterative refinement ensures that the models remain adaptable to new challenges, which is often missing in existing systems that may be rigid and unable to evolve as quickly.

Despite its strengths, the proposed framework has some limitations that should be addressed in future iterations:

- **Data Imbalances:** While data balancing techniques, such as oversampling or undersampling, are included, real-world datasets may still exhibit inherent imbalances in the distribution of sentiments or review types. This could affect the model's ability to generalize well across different types of reviews, especially for rare or underrepresented sentiments.
- **Interpretability vs. Performance Trade-off:** While the framework emphasizes interpretability, deep learning models (especially transformer-based models like BERT and GPT) can be challenging to interpret without sacrificing some predictive power. Striking the right balance between model transparency and performance remains an ongoing challenge in AI development.
- **Scalability:** The framework may face challenges in scaling to very large datasets or real-time applications. Processing large amounts of online review data in a timely and efficient manner while maintaining the ethical safeguards outlined in the methodology may require significant computational resources.
- **Ethical Oversight:** Ethical considerations, while central to the framework, are complex and may vary depending on the cultural and regulatory context. Implementing global ethical standards in a universal AI system is difficult, as different regions may have different norms and expectations regarding privacy, fairness, and transparency.

The iterative refinement process is a key component of the proposed methodology. As AI systems continue to evolve, it is essential to update models regularly to adapt to emerging challenges and feedback. Some potential future directions for this research include:

- **Incorporating User Feedback:** Further research can explore methods for actively incorporating user feedback into the model refinement process. This could involve real-time feedback loops where users can rate the accuracy and fairness of predictions, which would allow the model to learn from its users.
- **Expanding Ethical Considerations:** Future frameworks could integrate more detailed ethical assessments, including testing the model in real-world scenarios to assess its real-world impact on user behavior, review systems, and business practices.
- **Exploring Multilingual Sentiment Analysis:** The current framework primarily focuses on English-language reviews. Expanding the framework to handle multiple languages and cultural contexts could increase the system's global applicability and effectiveness.

- **Automation of Ethical Audits:** Developing automated systems for continuous ethical audits would help monitor model performance across ethical dimensions, ensuring that the AI system remains aligned with responsible practices throughout its lifecycle.

5 | Conclusion

The proposed framework for developing responsible AI methods for NLP, sentiment analysis, and online reviews ranking provides a comprehensive, ethical approach to solving key challenges in the field. By combining data preprocessing, model development, performance evaluation, and ethical considerations, the framework promotes the development of AI systems that are accurate, transparent, fair, and trustworthy. While challenges remain, particularly in balancing interpretability with performance and scaling to large datasets, the framework represents a significant step forward in integrating responsibility into AI systems for real-world applications.

Acknowledgments

No acknowledgments to declare.

Author Contribution

The authors confirm that all contributions to this manuscript, including conception, design, and writing, were equally shared.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

This study is based on a conceptual framework, and no empirical data were generated or analyzed.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] R. Liu, S. Gupta, and P. Patel, "The Application of the Principles of Responsible AI on Social Media Marketing for Digital Health," Inf. Syst. Front., 2023, doi: 10.1007/s10796-021-10191-z.

[2] M. Anusha and R. Leelavathi, "Analysis on Sentiment Analytics Using Deep Learning Techniques," in Proceedings of the 5th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2021, 2021. doi: 10.1109/I-SMAC52330.2021.9640790.

[3] T. M. Breuel, "High Performance Text Recognition Using a Hybrid Convolutional-LSTM Implementation," in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 2017. doi: 10.1109/ICDAR.2017.12.

[4] M. Tamilselvi, G. Ramkumar, G. Anitha, P. Nirmala, and S. Ramesh, "A Novel Text Recognition Scheme using Classification Assisted Digital Image Processing Strategy," in Proceedings - IEEE International Conference on Advances in Computing, Communication and Applied Informatics, ACCAI 2022, 2022. doi: 10.1109/ACCAI53970.2022.9752542.

[5] C. Castelfranchi, "For a Science-oriented, Socially Responsible, and Self-Aware AI: Beyond ethical issues," in Proceedings of the 2020 IEEE International Conference on Human-Machine Systems, ICHMS 2020, 2020. doi: 10.1109/ICHMS49158.2020.9209369.

[6] I. Golbin, A. S. Rao, A. Hadjarian, and D. Krittman, "Responsible AI: A Primer for the Legal Community," in Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020, 2020. doi: 10.1109/BigData50022.2020.9377738.

[7] E. Neghawi, Z. Wang, J. Huang, and Y. Liu, "Linking Team-level and Organization-level Governance in Machine Learning Operations through Explainable AI and Responsible AI Connector," in Proceedings - International Computer Software and Applications Conference, 2023. doi: 10.1109/COMPSAC57700.2023.00185.

[8] K. Werder, B. Ramesh, and R. S. Zhang, "Establishing Data Provenance for Responsible Artificial Intelligence Systems," ACM Trans. Manag. Inf. Syst., 2022, doi: 10.1145/3503488.

[9] H. Lee, Y. Jin, and O. Kwon, "Investigating the Impact of Corporate Social Responsibility on Firm's Short- and Long-Term Performance with Online Text Analytics," J. Intell. Inf. Syst., 2016, doi: 10.13088/jiis.2016.22.2.013.

[10] C. Maree, J. E. Modal, and C. W. Omlin, "Towards Responsible AI for Financial Transactions," in 2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020, 2020. doi: 10.1109/SSCI47803.2020.9308456.

[11] S. Shetty, A. S. Devadiga, S. Sibi Chakkaravarthy, and K. A. Varun Kumar, "Ote-OCR based text recognition and extraction from video frames," in 2014 IEEE 8th International Conference on Intelligent Systems and Control: Green Challenges and Smart Solutions, ISCO 2014 - Proceedings, 2014. doi: 10.1109/ISCO.2014.7103949.

[12] F. A. Batarseh, G. Nambiar, G. Gendron, and R. Yang, "Geo-enabled text analytics through sentiment scoring and hierarchical clustering," in 2018 7th International Conference on Agro-Geoinformatics, Agro-Geoinformatics 2018, 2018. doi: 10.1109/Agro-Geoinformatics.2018.8475993.

[13] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots," 2021. doi: 10.1145/3442188.3445922.

[14] E. Sheng, K. W. Chang, P. Natarajan, and N. Peng, "The woman worked as a babysitter: On biases in language generation," in EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 2019. doi: 10.18653/v1/d19-1339.

[15] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods," in NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2018. doi: 10.18653/v1/n18-2003.

[16] O. D. Thakkar, S. Ramaswamy, R. Mathews, and F. Beaufays, "Understanding Unintended Memorization in Language Models Under Federated Learning," 2021. doi: 10.18653/v1/2021.privatenlp-1.1.

[17] M. Mitchell et al., "Model cards for model reporting," in FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, 2019. doi: 10.1145/3287560.3287596.

[18] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "REALTOXICITYPROMPTS: Evaluating neural toxic degeneration in language models," in Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020, 2020. doi: 10.18653/v1/2020.findings-emnlp.301.

[19] I. D. Raji et al., "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020. doi: 10.1145/3351095.3372873.

[20] K. Holstein, J. W. Vaughan, H. Daumé, M. Dudík, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?," in Conference on Human Factors in Computing Systems - Proceedings, 2019. doi: 10.1145/3290605.3300830.

[21] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," in ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 2018. doi: 10.18653/v1/p18-2124.

[22] R. Bommasani, P. Liang, and T. Lee, "Holistic Evaluation of Language Models," Ann. N. Y. Acad. Sci., 2023, doi: 10.1111/nyas.15007.

[23] T. Sun et al., "Mitigating gender bias in natural language processing: Literature review," in ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2020. doi: 10.18653/v1/p19-1159.

[24] S. Singh, D. J. Greaves, and G. Epiphaniou, "A Framework for Integrating Responsible AI into Social Media Platforms," in IET Conference Proceedings, 2022. doi: 10.1049/icp.2022.2051.

[25] M. I. Merhi, "An Assessment of the Barriers Impacting Responsible Artificial Intelligence," Inf. Syst. Front., 2023, doi: 10.1007/s10796-022-10276-3.

[26] P. Mikalef, K. Conboy, J. E. Lundström, and A. Popovič, "Thinking responsibly about responsible AI and 'the dark side' of AI," 2022. doi: 10.1080/0960085X.2022.2026621.

[27] Q. Zhang, R. Zheng, Z. Zhao, B. Chai, and J. Li, "A TextCNN Based Approach for Multi-label Text Classification of Power Fault Data," in 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics, ICCCBDA 2020, 2020. doi: 10.1109/ICCCBDA49378.2020.9095584.

[28] T. K. Patra, R. Rani, and P. Dayal, "Analytics for image, text and audio analysis for military surveillance, command & control systems," in Proceedings of the 2017 2nd IEEE International Conference on Electrical, Computer and Communication Technologies, ICECCT 2017, 2017. doi: 10.1109/ICECCT.2017.8118039.

[29] C. Sanderson, Q. Lu, D. Douglas, X. Xu, L. Zhu, and J. Whittle, "Towards Implementing Responsible AI," in Proceedings - 2022 IEEE International Conference on Big Data, Big Data 2022, 2022. doi: 10.1109/BigData55660.2022.10021121.

[30] Y. Qin and Z. Zhang, "Summary of Scene Text Detection and Recognition," in Proceedings of the 15th IEEE Conference on Industrial Electronics and Applications, ICIEA 2020, 2020. doi: 10.1109/ICIEA48937.2020.9248121.

[31] Y. Wang and W. Xu, "Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud," Decis. Support Syst., 2018, doi: 10.1016/j.dss.2017.11.001.

[32] O. Inel, T. Draws, and L. Aroyo, "Collect, Measure, Repeat: Reliability Factors for Responsible AI Data Collection," Proc. AAAI Conf. Hum. Comput. Crowdsourcing, 2023, doi: 10.1609/hcomp.v11i1.27547.

[33] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," Nat. Mach. Intell., 2019, doi: 10.1038/s42256-019-0088-2.

[34] M. T. Ribeiro, S. Singh, and C. Guestrin, "'why should i trust you?' explaining the predictions of any classifier," in NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session, 2016. doi: 10.18653/v1/n16-3020.

[35] F. Doshi-Velez and B. Kim, "Considerations for Evaluation and Generalization in Interpretable Machine Learning," 2018. doi: 10.1007/978-3-319-98131-4_1.

[36] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015. doi: 10.1145/2783258.2788613.

[37] I. Y. Chen, F. D. Johansson, and D. Sontag, "Why is my classifier discriminatory?," in Advances in Neural Information Processing Systems, 2018.

[38] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in 35th International Conference on Machine Learning, ICML 2018, 2018.

[39] R. Binns, "Fairness in Machine Learning: Lessons from Political Philosophy," in Proceedings of Machine Learning Research, 2018.

[40] L. Floridi and J. Cowls, "A unified framework of five principles for AI in society," in Machine Learning and the City: Applications in Architecture and Urban Design, 2022. doi: 10.1002/9781119815075.ch45.

[41] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: Understanding rating dimensions with review text," in RecSys 2013 - Proceedings of the 7th ACM Conference on Recommender Systems, 2013. doi: 10.1145/2507157.2507163.

[42] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, 2015. doi: 10.18653/v1/d15-1167.

[43] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2019.