Sentiment Analysis of AI Perception

1st Jesús Arturo Mandujano Granillo
School of Engineering and Science
Tecnologico de Monterrey
Monterrey, Mexico
a01250871@tec.mx

***Abstract*—This study examines public sentiment towards artificial intelligence (AI) by analyzing three decades of New York Times articles. Employing sentiment analysis, the research captures evolving perceptions, from optimism surrounding technological advancements to concerns over ethical implications. Initial data cleaning ensures a robust dataset, removing duplicates and standardizing dates. Sentiment scores are categorized from "Very Negative" to "Very Positive," and the dataset is balanced to mitigate bias. Exploratory data analysis reveals sentiment distribution, trends over time, and prevalent themes using visualizations such as word clouds. For in-depth analysis, combining TF-IDF and K-means clustering allow the identification of sentiment patterns, and transformer-based models like BERT and VADER classify the sentiment of texts. Model performance is evaluated using precision, recall, and F1 score, ensuring comprehensive sentiment classification. The study's findings aim to inform AI researchers by providing insights into public opinion, with the potential to guide responsible AI development and policy-making.**

***Index Terms*—artificial intelligence, sentiment analysis, text mining, data validation, statistic analysis, exploratory data analysis, transformer-based models, k-means clustering**

I. Introduction

Artificial intelligence has been seen as a critical influence across various domains. As AI technologies advance, they provoke a diverse array of public sentiments, encompassing both enthusiasm for technological possibilities and anxiety over potential repercussions. Sentiment analysis, utilizing data analysis techniques to interpret and classify emotions within text data, stands as an essential method for quantifying these diverse perspectives. By applying sentiment analysis to large-scale textual data from media, forums, and other public platforms, researchers can monitor and analyze public opinion, providing feedback to align AI development with concerns from the society.

A. Background and related work

The perception of AI has evolved significantly with technological advancements that have pushed the boundaries of what machines can emulate or accomplish. Recent years have seen AI systems generate realistic images, engage in complex human interactions via chatbots, among others. These capabilities have not only enhanced convenience and efficiency across several sectors but have also sparked ethical, economic, and social concerns. For instance, the ability of AI to replicate human attributes raises questions about privacy and creativity, while its role in automating jobs brings economic debates concerning employment and inequality. Researchers have addressed the shifting public perception of AI. Studies often employ sentiment analysis to assess the tone and context of discussions surrounding AI in various media outlets. These works reveal a nuanced landscape where excitement about AI's potential coexists with fear about its implications. Articles have increasingly focused on temporal analysis, showing how specific events, such as the introduction of AI in autonomous vehicles or revelations of surveillance capabilities, shift public sentiment discernibly.

B. Problem definition

The integration of artificialAI into everyday applications and critical infrastructures has accelerated, bringing significant benefits and challenges. Public perception of AI is pivotal as it influences policy-making, development strategies, and societal acceptance of technology. However, accurately gauging this perception is complex due to the volume and variability of public discourse on AI. This project addresses the need to systematically analyze public sentiment towards AI, leveraging sentiment analysis to decode the vast textual data from a prominent news source over an extended period. This approach aims to elucidate the multifaceted views of the public that range from optimistic endorsements to serious ethical concerns. It is hypothesized that public sentiment towards AI usage has experienced a notable shift over the past three decades, largely influenced by technological advancements and emerging applications of AI. These shifts are presumed to reflect a complex interplay of increased optimism about AI's benefits, tempered by escalating concerns over its ethical ramifications and socio-economic impacts. This hypothesis posits that significant AI developments or controversies significantly impact public sentiment, with these effects traceable through changes in media discourse.

- How has public sentiment towards artificial intelligence changed over the last 30 years in response to ongoing developments and debates?
- What specific aspects of AI applications are most frequently associated with positive and negative sentiments in public discourse?
- Does sentiment towards AI vary significantly across different sections of the New York Times, and what does this indicate about the contextual sensitivity of AI perception?

C. Proposed approach

This research uses a robust dataset extracted from Kaggle related to the "Public Perception of AI" based on the New York Times newspaper, encompassing articles published over a 30-year period. This dataset offers a longitudinal view of public discourse on artificial intelligence, capturing the nuances and shifts in sentiment across different societal spheres. The dataset is rich in text data, including individual paragraphs that mention AI, the date of publication, and categorical AI mood ratings, making it ideal for sentiment analysis over time. Initial data cleaning involves removing duplicates, handling missing data, and standardizing dates. AI mood scores will be categorized for clarity, and the dataset was balanced to negate bias in sentiment representation. Statistical and exploratory analyses will be conducted to understand sentiment distribution, temporal trends, and sectional emphasis on AI, as well as to identify prevalent keywords. The clean and balanced dataset will be analyzed using clustering for analyzing sentiment patterns and transformer-based models to classify paragraph-level sentiments, with their performance evaluated by standard metrics.

II. Related Work

The intersection of generative AI and advertising opens intriguing avenues for exploring generative AI technologies' impact on consumer perceptions. This section unveils a collection of experimental studies that critically examine use and public perception of AI-generated content within advertising campaigns. Focusing on scenarios ranging from charitable campaigns to broader marketing strategies, this discussion aims to bridge theoretical insights with practical implications, offering a nuanced understanding of generative AI's role in shaping future advertising landscapes and its reception among consumers.

The study made by Arango (2023) discusses the impact of artificial intelligence AI-generated content on consumer perceptions in the context of charitable advertising. Through three experimental studies, the research explores how potential donors react to AI-generated images of children's faces in charitable giving contexts. The findings highlight that knowledge of the images being AI-generated or not real negatively affects donation intentions, with this effect being mediated by reduced empathy and increased anticipatory guilt. However, when charities make their ethical reasons for using AI images clear, the negative impact lessens. The research also suggests that in extraordinary circumstances, consumers may find the use of AI-generated images acceptable, leading to similar outcomes as using real images. The authors advocate for a careful approach towards incorporating synthetic content in advertising and marketing strategies [1].

The study from Xu et al. (2024) explores the challenge of source bias in web search engines caused by the increasing sophistication of AI-generated content, which states that it is becoming almost indistinguishable from content created by humans. It specifically investigates how this bias manifests in text-image retrieval systems, where AI-generated images are often ranked higher than real images by retrieval models, despite not being more relevant [5].

Finally, the study from Lu et al. (2023) addresses concerns around the increasing difficulty in distinguishing between real and AI-generated images, an issue that threatens to undermine the credibility of photographs as reliable sources of information. The research introduces a large-scale fake image dataset to evaluate the ability of both humans and advanced AI algorithms to identify AI-generated visual content. In a human perception test, it was found that humans had a significant challenge in distinguishing real from fake images, with a misclassification rate of 38.7%.

III. Data and Methods

In this section, a detailed methodology is outlined, focusing on the integrity and robustness of the findings. Data validation is initially conducted to ensure the accuracy and reliability of the dataset, providing a firm foundation for further analysis. This is followed by a description of the statistical analysis techniques used, which form the backbone for interpreting the data related to the detection of AI-generated visual content. An exploratory data analysis (EDA) is also performed, concentrating on essential procedures such as feature selection, feature engineering, addressing unbalanced data, and utilizing data visualization to reveal underlying patterns and relationships. Lastly, the statistical or machine learning approach employed in this study is described.

A. Data validation

A comprehensive data validation process was done to ensure the reliability of the findings. This process encompassed several stages, from initial cleaning to feature engineering, each designed to enhance the dataset's suitability for sentiment analysis.

The cleaning process involved several steps:

- Duplicate Entries: All duplicate paragraphs were removed, ensuring that each instance within the dataset represented a unique commentary on AI.
- Missing Data: Entries with missing values in key fields, particularly those lacking paragraph text or title information, were excluded to maintain the analysis's focus on complete and informative records.
- Standardization: The 'Article Date' field was normalized across the dataset, converting all dates to a consistent format that facilitates temporal trend analysis.

B. Statistical analysis

After cleaning the data, a time series analysis can be performed to observe how AI mentions have evolved through time. Figure 1 shows how interest in the topic of AI has been increasing constantly throughout the years.

C. Exploratory Data Analysis

After the cleaning process, to prepare the textual data for sentiment analysis, feature engineering was employed to convert raw text into analyzable formats. The main target to

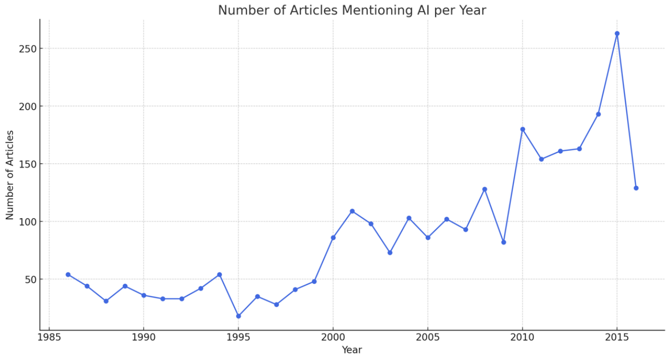


Fig. 1: Number of articles mentioning AI per year

focus on the categorical variable ”AI Mood” which ranges from 1 to 5. All values were set to represent the following values: 1 = ”Very Negative”, 2 = ”Negative”, 3 = ”Neutral”, 4 = ”Positive” and 5 = ”Very Positive”.

The dataset contains 1,574 entries classified as having a low AI mood (mood scores from 1 to 2) and 15,484 entries classified as having a high AI mood (mood scores from 3 to 5). This shows a significant imbalance in the dataset, with a much larger number of entries associated with a high AI mood compared to a low AI mood as seen in Figure 2.

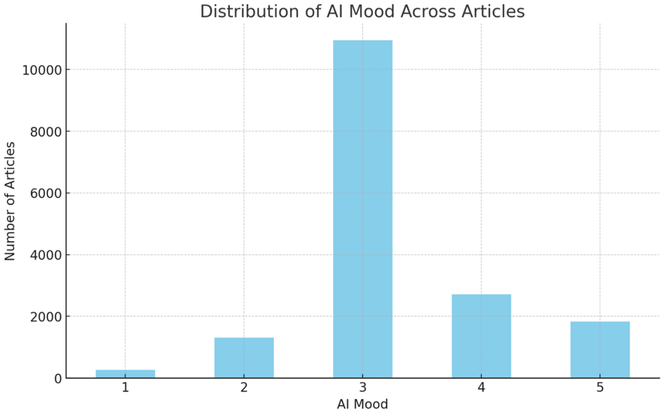


Fig. 2: AI Mood distribution across articles

The next step is eliminating the ”Neutral” categorical variable to make a clear distinction between low and high ai moods. After adjusting the criteria, where low AI mood ranges from 1 to 2 and high AI mood ranges from 4 to 5, the dataset contains 1,574 entries classified as low AI mood and 4,551 entries classified as high AI mood, still presenting imbalance. To address potential bias from imbalanced sentiment classes, the dataset was balanced by under-sampling the overrepresented AI moods, giving as a result.

- Low AI Mood (1-2): 1,574 entries
- High AI Mood (4-5): 1,574 entries

This balanced dataset allows for a more equitable comparison between low and high AI moods without the over representation of any single mood category. It should provide clearer insights into the differences in sentiment towards AI, facilitating a more nuanced analysis of positive versus negative perceptions. The results of this balance can be seen in Figure 3.

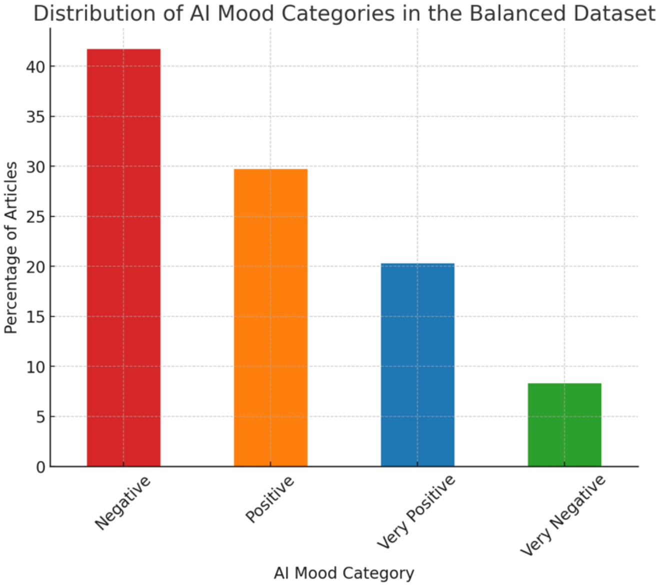


Fig. 3: Balanced data AI Mood distribution across articles

Where the distributions of AI perception shows the following results:

- ”Very Negative”: Approximately 8.29% of articles
- ”Negative”: Approximately 41.71% of articles
- ”Positive”: Approximately 29.73% of articles
- ”Very positive”: Approximately 20.27% of article

With the balanced dataset, a yearly distribution of the AI Mood categories can also be presented to show how different moods have been evolving over time as seen in Figure 4.

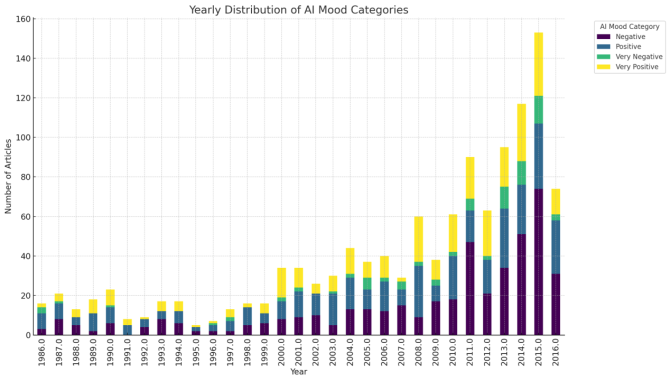


Fig. 4: Distribution of AI Mood categories over time

Additionally, the distribution of these AI mood categories can also be seen in regards to the NYT newspaper section as in Figure 5.

Finally, once the balanced dataset was divided into positive and negative categories, keyword analysis can be performed to

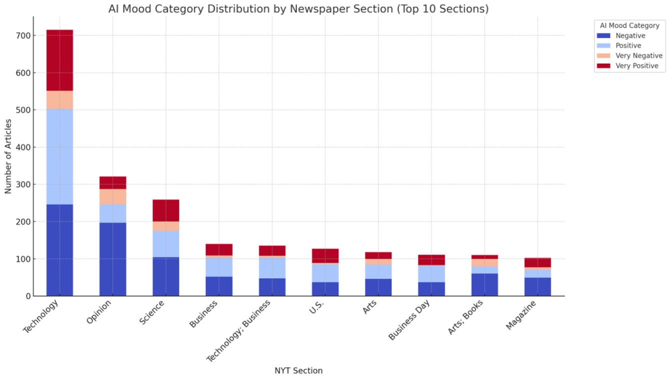


Fig. 5: Distribution of AI Mood categories by newspaper section

generate word clouds to visually represent the most frequent occurring words within each mood category as seen in Figure 6.

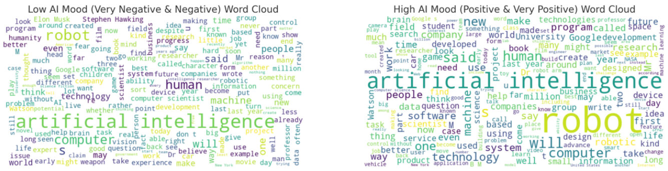

Low AI Mood (Very Negative & Negative) Word Cloud
High AI Mood (Positive & Very Positive) Word Cloud

Fig. 6: AI Mood word clouds

D. Method

Once the different word clouds were generated in the exploratory data analysis, common stop words, which don't provide significant meaning will be eliminated from both groups. Then, the use of TF-IDF, K-means clustering, and transformer-based modeling techniques will be implemented for analysis. TF-IDF allows to a numerical representation of text data for the K-means clustering to be applied to identify sentiment patterns within the data [8], with the optimal number of clusters determined by the elbow method" and "silhouette coefficient" to best reflect the structures present in the text.

In addition to the clustering analysis, the performance of the Bidirectional Encoder Representations from Transformers (BERT) and the Valence Aware Dictionary and Sentiment Reasoner (VADER) models was evaluated on the basis of categorizing texts identified with negative and positive sentiments [9], [10]. These models were trained and tested on the processed dataset, with their performance metrics including precision, recall and F1-score compared to assess their effectiveness in sentiment classification.

IV. Results

A. TF-IDF and K-Means Clustering

With the numerical representation of the text data obtained through TF-IDF, and divided into "positive" and "negative" moods, K-means clustering was performed to group the data into distinct clusters. The procedure first involved determining an optimal number of clusters, which was determined with the elbow method which involves plotting the sum of squared errors (SSE) within clusters against the number of clusters (k). Next, the silhouette coefficient is used to measure how similar a word collection is to its own cluster compared to other clusters. The resulting elbow and silhouette coefficient graphs for both "negative" and "positive" moods can be seen in Figures 7 and 8 respectively.

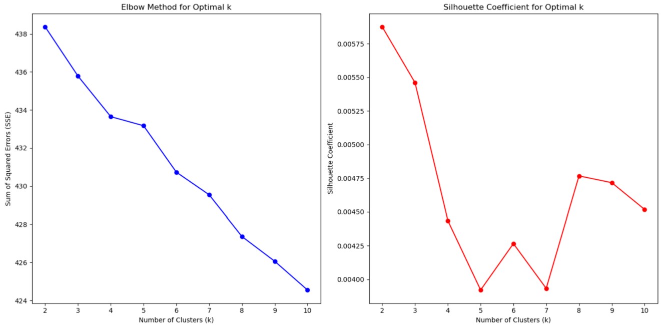


Fig. 7: Elbow graph for "Negative" AI Mood

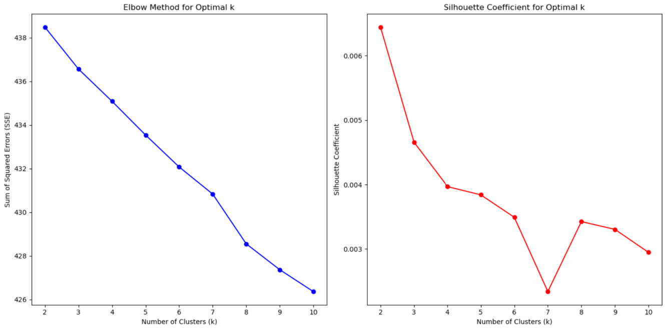


Fig. 8: Elbow graph for "Positive" AI Mood

Analyzing the behavior in the elbow graphs, for the negative AI mood, the k value in which the distortion improvement declines can be observed at k=4, but with a continuous behavior afterwards. For the positive AI mood, the behavior is practically the same with a very slight decline at k=3.

Analyzing the silhouette coefficients for both moods, the highest scores can be observed with k=2 and k=3, with a clear difference afterwards as seen in Figure 9.

After analyzing both of these results, the optimal number of clusters was determined to be k=3. Finally, to get insights from the generated clusters, the most common words for each cluster were analyzed as shown in Figure 10

B. BERT and VADER Models

The BERT model was implemented to classify text into positive and negative sentiments. The implementation process began with tokenizing the text data using BERT's tokenizer, converting the text into a suitable format for the model. The tokenized text was then encoded into tensors, and a custom dataset class was defined to manage the encoded text and corresponding labels.

k	Silhouette Coefficient
2	0.005876
3	**0.005460**
4	0.004437
5	0.003921
6	0.004266
7	0.003933
8	0.004767
9	0.004717
10	0.004518

(a) Negative AI Mood

k	Silhouette Coefficient
2	0.006444
3	**0.004657**
4	0.003969
5	0.003841
6	0.003492
7	0.002339
8	0.003425
9	0.003304
10	0.002948

(b) Positive AI Mood

Fig. 9: Silhouette Coefficients for Various k Values for Negative and Positive AI Moods

Cluster	Top terms
1	mr, musk, hawking, said, like, elon, biggest, could, threat, stephen
2	human, computer, ai, computers, machines, systems, think, us, technology, would
3	robot, said, robots, new, one, people, like, also, two, work

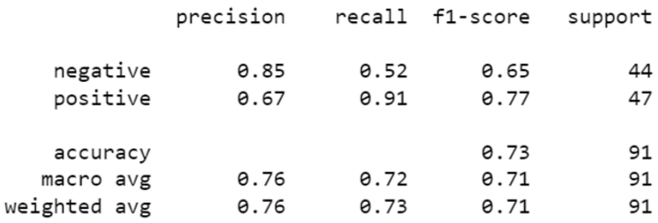
(a) Top terms per cluster for Negative AI Mood

Cluster	Top terms
1	human, software, computer, like, data, technology, ai, new, computers, program
2	said, people, work, robot, mr, years, would, one, could, dr
3	robot, new, arm, robots, one, last, would, also, space, two

(b) Top terms per cluster for Positive AI Mood

Fig. 10: Top terms per cluster for AI Mood

To prepare the dataset for training and evaluation, the data was split into training and validation sets, maintaining a 90-10 ratio. The total set consisted on 910 samples, in which 819 were used for training and 91 for validation and performance evaluation. The BERT model was fine-tuned on the training data for three epochs. During training, a batch size of 8 was used for training, and a batch size of 16 was used for evaluation. Additional training arguments, such as warmup steps, weight decay, and logging configurations, were set to optimize model performance. After training, the model was evaluated on the validation set, and metrics such as precision, recall, and F1-score were computed as seen in Figure 11.

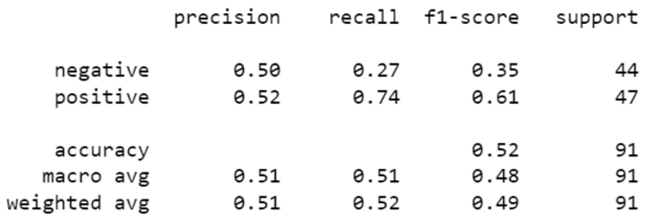


```
              precision    recall  f1-score   support

    negative       0.85      0.52      0.65        44
    positive       0.67      0.91      0.77        47

    accuracy                           0.73        91
   macro avg       0.76      0.72      0.71        91
weighted avg       0.76      0.73      0.71        91
```

Fig. 11: BERT model performance scores

The VADER model was implemented to provide a comparative baseline to the BERT model. The implementation process involved scoring each text paragraph using VADER, which produced compound scores reflecting the overall sentiment. Based on these compound scores, each text was labeled as positive or negative, excluding neutral sentiments from the analysis.

To ensure consistency in comparison, the same training and validation sets used for the BERT model were employed for the VADER model. The performance of the VADER model was evaluated by computing metrics such as precision, recall, and F1-score shown in Figure 12. This approach provided a direct comparison between the transformer-based BERT model and the lexicon-based VADER model, highlighting their respective strengths and limitations in sentiment analysis tasks.



```
              precision    recall  f1-score   support

    negative       0.50      0.27      0.35        44
    positive       0.52      0.74      0.61        47

    accuracy                           0.52        91
   macro avg       0.51      0.51      0.48        91
weighted avg       0.51      0.52      0.49        91
```

Fig. 12: VADER model performance scores

Finally, to assess the performance of both the BERT and VADER models, confusion matrices were generated for both. For the BERT model, the confusion matrix was created by comparing the model's predictions on the validation set with the true labels. The true labels and predicted labels were used to construct the confusion matrix, which provided insights into the number of true positives, true negatives, false positives, and false negatives.

Similarly, for the VADER model, the confusion matrix was generated by comparing the predicted sentiments with the actual labels from the validation set. After calculating the compound sentiment scores for each text using VADER, the scores were converted to positive and negative labels. These predicted labels were then compared against the true labels to create the confusion matrix. This matrix highlighted how well the VADER model performed in correctly classifying sentiments and where it faced challenges, offering a clear comparison between the lexicon-based approach of VADER and the deep learning-based approach of BERT. The resulting confusion matrices for both transformer-based models are shown in Figure 13.

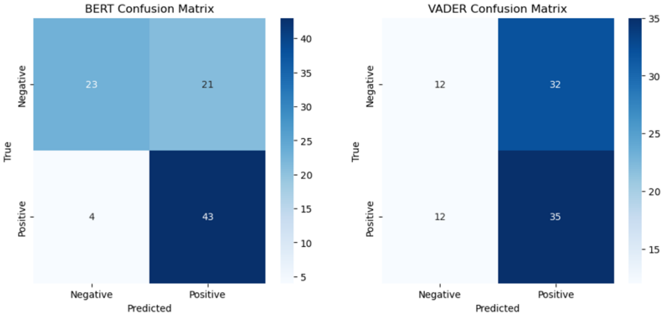


Fig. 13: Transformer-based models confusion matrix comparison

V. Discussion

In this section, an analysis of the various methods employed in this project is presented, including TF-IDF, K-means clustering, and the transformer-based models BERT and VADER. Each method's performance and its implications for sentiment analysis are discussed with the objective of providing a comprehensive understanding of how they contribute to identifying and categorizing sentiments in text data.

Beginning with the TF-IDF method, from Figure 10a, it can be seen that the negative mood clusters has some pattern in each of the clusters. For the first cluster, the most notable aspect are the names and surnames from known scientific and technological personalities such as Elon Musk and Stephen Hawking are present, which could indicate that their quotes on the matter have had a significant impact on negative AI perception. From the second clusters, some aspects about computers and machines related to AI have an impact in people's perception. Finally, cluster 4 is more related to robots and work, potentially indicating an interpretation of how AI could advance into replacing human work, leaving a negative impresion.

From Figure 10b, the first positive mood cluster is related to software and technology, indicating that advances in AI applied to them could be well perceived. The second cluster to people and work, potentially indicating to how AI could be used as a complementary tool, giving positive perception. Finally, the third cluster is related to robots, this aspect also being both well and bad perceived.

For the analysis of the transformer-based models, for the negative sentiment, the BERT model achieved a precision of 0.85, recall of 0.52, and an F1-score of 0.65, with a support of 44 samples. In comparison, the VADER model obtained a precision of 0.50, recall of 0.27, and an F1-score of 0.35 for the same category. These results indicate that the BERT model is significantly better at correctly identifying negative sentiments, balancing both the true positives and false negatives more effectively than the VADER model.

For the positive sentiment, the BERT model also outperformed VADER, with a precision of 0.67, recall of 0.91, and an F1-score of 0.77, against VADER's precision of 0.52, recall of 0.74, and F1-score of 0.61. The support for both models in this category was 47 samples. These metrics suggest that the BERT model is more accurate in identifying positive sentiments and reducing false negatives.

Overall, the BERT model achieved an accuracy of 0.73, whereas the VADER model had an accuracy of 0.52. The macro average and weighted average F1-scores for the BERT model were both 0.71, compared to VADER's 0.48 and 0.49, respectively.

The confusion matrices for both models provided additional information in which the BERT model correctly classified a higher number of both positive and negative samples, whereas the VADER model struggled, particularly with negative sentiments. This discrepancy was evident in the higher false negative rates for VADER.

VI. Conclusion

The TF-IDF and K-Means clustering methods facilitated the segmentation of words used to describe sentiments towards AI adoption, allowing for a detailed analysis of specific topics that may generate these sentiments. Transformer-based methods, namely BERT and VADER, were employed to develop models capable of classifying sentiments. The results demonstrated that the BERT model outperformed VADER in classifying both positive and negative moods towards AI implementation, with higher precision, recall, and F1-scores. Negative sentiment was found to be easier to predict than positive sentiment. This study highlights the potential business applications of sentiment analysis tools, particularly in analyzing historical customer review and feedback data. K-Means clustering can help identify the main sources of both positive and negative feedback, enabling targeted interventions. Additionally, transformer-based models can predict whether customer feedback is likely to result in a positive or negative sentiment, offering valuable insights for businesses seeking to improve customer satisfaction and address specific concerns effectively.

References

[1] Arango, L., Singaraju, S. P., & Niininen, O. (2023). Consumer Responses to AI-Generated Charitable Giving Ads. Journal of Advertising, 52(4), 486–503. https://doi.org/10.1080/00913367.2023.2183285.

[2] Lim, C.V.; Zhu, Y.-P.; Omar, M.; Park, H.-W. Decoding the Relationship of Artificial Intelligence, Advertising, and Generative Models. Digital 2024, 4, 244-270. https://doi.org/10.3390/digital4010013.

[3] Daniel L. Hocutt, Composing with generative AI on digital advertising platforms, Computers and Composition, Volume 71, 2024, 102829,ISSN 8755-4615, https://doi.org/10.1016/j.compcom.2024.102829..

[4] Gołab-Andrzejak, E. (2023). The Impact of Generative AI and ChatGPT on Creating Digital Advertising Campaigns. Cybernetics and Systems, 1–15. https://doi.org/10.1080/01969722.2023.2296253.

[5] Xu, S. and Danyang H. and Liang P. and Jingcheng D. and Jun X. and Huawei S. and Xueqi C. (2023), AI-Generated Images Introduce Invisible Relevance Bias to Text-Image Retrieval, arXIV, 2311.14.084, https://doi.org/10.48550/arXiv.2311.14084.

[6] Baek, T. H. (2023). Digital Advertising in the Age of Generative AI. Journal of Current Issues & Research in Advertising, 44(3), 249–251. https://doi.org/10.1080/10641734.2023.2243496.

[7] Lu, Z., Di H., Lei B., Jingjing Q., Chengyue W., Xihui L. and Wanli O. (2023). arXiv. 2304.13023. https://doi.org/10.48550/arXiv.2304.13023

[8] P. Bafna, D. Pramod and A. Vaidya, ”Document clustering: TF-IDF approach,” 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, India, 2016, pp. 61-66, doi: 10.1109/ICEEOT.2016.7754750. keywords: Clustering algorithms;Entropy;Electronic mail;Classification algorithms;Algorithm design and analysis;Feature extraction;Symbiosis;TF-IDF;entropy;silhouette coefficient;hierarchical;fuzzy k-means;clustering,

[9] Bello, A.; Ng, S.-C.; Leung, M.-F. A BERT Framework to Sentiment Analysis of Tweets. Sensors 2023, 23, 506. https://doi.org/10.3390/s23010506

[10] Bonta, Venkateswarlu & Kumaresh, Nandhini & Naulegari, Janardhan. (2019). A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis. Asian Journal of Computer Science and Technology. 8. 1-6. 10.51983/ajcst-2019.8.S2.2037.