NewsAlign Technical Specifications and Evaluation

Gabriel Hili
Department of Artificial Intelligence
University of Malta
L-Imsida, Malta
gabriel.hili@um.edu.mt

Dylan Seychell
Department of Artificial Intelligence
University of Malta
L-Imsida, Malta
dylan.seychell@um.edu.mt

Fig. 1: A sample usage of NewsAlign, consisting of a visualisation tool to indicate image-text similarity in articles.

I. Methodology

A. Introduction

While many studies have explored bias-sensitive news aggregation, few have focused on visual content, likely due to limitations in computer vision technology and the complex nature of visual bias. We propose NewsAlign, a bias-sensitive news aggregator that addresses a significant gap in current literature: the analysis of picture-selection and picture-explanation bias in news articles. NewsAlign takes the form of a Chrome extension, where users can browse articles on online news portals as they usually do. However, when viewing an article, NewsAlign aggregates articles on the event level and displays them as points on an image-text similarity spectrum. The 1-D spectrum provides users with multiple sources on the event but is conscious of how the articles are presented with respect to picture-selection and picture-explanation bias. Currently, the newspapers supported are Times of Malta, Malta Today, The Malta Independent, The Shift, and Newsbook.

The position of the points on this line represents the similarity score between that article's thumbnail and headline. With this interface, users are able to compare how different articles' choices of headlines reflect the visual content pertaining to that event. In addition, NewsAlign also supports aggregating articles on the image level; i.e. retrieving articles that utilise visually similar images. A vector database is continuously populated with the latest news articles from the five newspapers currently supported. Event-level articles are retrieved from this database and a VLM (blip-2) is used to compute the image-text similarity score. Lastly, in order to promote better explainability between the end-user and the AI system, GradCAM [1] is used to show a heatmap of the most salient areas of the thumbnail image according to the VLM. Figure 2 showcases the interaction between these components.

B. Implementation Details

The system architecture is made up of two primary components: the data collection phase and the inference stage. During the data collection phase, a web-listener module periodically monitors the online portals of the previously specified newspapers for new articles. These articles are then added to a vector database. The textual content is processed using nomic-embed-text-v1, while image processing is done using blip-2. The metadata of each text entry in the text collection maintains a reference to the corresponding image entry's identifier in the image collection.

In the inference stage, the front-end automatically detects if the user is on an article that belongs to one of the five newspapers, and communicates it to the backend system via an API. The key article is downloaded and added to the vector database if not present. Subsequently, textually similar articles are retrieved from the vector database, as well as visually similar articles, i.e., articles that use visually similar images. For each of these retrieved articles, an image-text similarity score is generated for the thumbnail-headline pair and any other image-caption pairs within the article. The articles and their accompanying similarity scores are sent to the front-end, where they are visualised on the image-text similarity spectrum.

1) Front-end: The front-end of the system takes the form of a Chrome extension and is implemented using a combination of HTML, CSS, JavaScript, and the D3.js

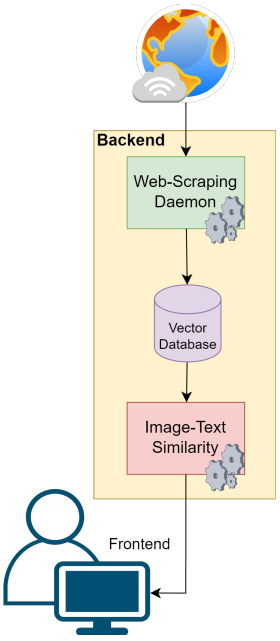


Fig. 2: NewsAlign Component Interactions

library [1]. It displays a 1-D spectrum pop-up over the thumbnail of the currently viewed article (See Fig. 1). Each data point on the spectrum represents an article that reports the same event as the one currently being viewed. The green point represents the thumbnail-headline similarity of the current article, while the blue points represent those of other articles. The position of the point on the line indicates the similarity score of that article's thumbnail to its headline. Users can hover over the points to get a preview (see Fig. 3a, Fig. 3b) or click on them to be redirected to the corresponding article. See Section I-B3 for more details.

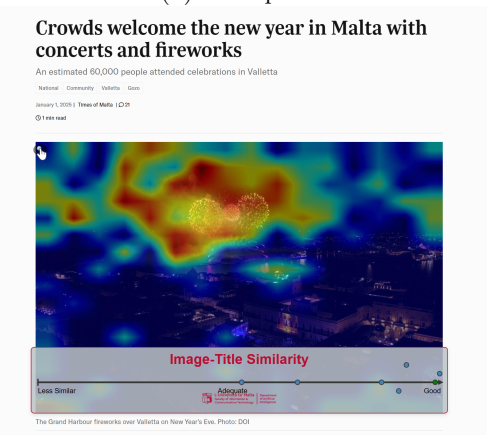
Each third of the image-text similarity spectrum is labelled as "Less Similar", "Adequate", "Good" respectively. Motivated by findings of a crowd-sourced annotation process (see Section II-B4) we split the spectrum into three equal parts to guide users' interpretation of the scores. The specific wording of the labels is intended to describe the results to the user while being careful not to make claims about the credibility of the article, as that is something that the reader must discern for themselves. Moreover, when the cursor is hovered on the icon on the top-left corner of the thumbnail, the GradCAM heatmap (see Section I-B4) is overlayed on the image; highlighting the most visually relevant features that match with the article headline (see Fig. 3c).

D3.js is used to create an interactive SVG-based visualisation in the form of a 1-D spectrum. It conveys to the user the different picture-selection and picture-explanation intricacies of the different articles that explain the key event.

[1]https://d3js.org/

(a) Example 1

(b) Example 2



(c) Example 3

Fig. 3: Examples of NewsAlign on different articles.

Collection	Embedded Content	Metadata
Text Collection (nomic-embed-text-v1)	search_document: {title}. {body}. {captions}	author, body, date, img_ids, newspaper, tags, title, url
Image Collection (blip-2)	Image Data	article_ids, captions, selectors, ascii_encoded_img

TABLE I: Schema for the text and image vector collections.

Model	# Params	COCO (5K test set)	
		Image → Text R@1	Text → Image R@1
$\text{CLIP}_{\text{VIT-L}}$ [2]	428M	57.9	37.1
$\text{SigLIP}_{\text{VIT-L}}$ [3]	-	70.6	52.7
$\text{BLIP}_{\text{VIT-L}}$ [4]	446M	82.4	65.1
$\text{BLIP-2}_{\text{VIT-L}}$ [5]	474M	83.5	66.3

TABLE II: Comparison of various vision-language models image-text retrieval capabilities on COCO

2) Vector Database: The vector database's role is to facilitate similar article searches based on either the textual content or the images within the article. The system uses two vector collections to store an article's textual and visual content. The text collection uses nomic-embed-text-v1 [6] and the image collection uses blip-2 [5]. The vector database is populated by the webscraper daemon (see Section I-B2). The current article viewed by the user is also added to the vector database, in the case that the webscraper daemon has not yet indexed it. Figure 4 showcases the log output of this process.

Nomic AI's nomic-embed-text-v1 [6] was selected for its support of long context lengths, which makes it suitable for representing all the details of a news article without compromise. It demonstrates superior performance compared to other text embedding models across various benchmarks while also being open-source (See Table III). For visual embedding, blip-2 was chosen due to its superior performance with models of its class (See Table II), as well as being the same model used to compute the image-text similarity scores.

3) Webscraper Daemon: A web listener program periodically checks the online portals of the five newspapers (Times of Malta, Malta Today, The Malta Independent, The Shift, Newsbook) for any new articles, which it then adds to the vector database according to the schema as depicted in Table I. Due to the different nature of the

Fig. 4: Log excerpt of NewsAlign retrieving event-level articles.

Name	Context Length	MTEB [7]	LoCo [8]	Jina Long Context [9]	Open Weights	Open Data
nomic-embed-text-v1 [6]	8192	62.39	85.53	54.16	•	•
jina-embeddings-v2-base-en [10]	8192	60.39	85.45	51.90	•	○
text-embedding-3-small [11]	8191	62.26	82.40	58.20	○	○
text-embedding-ada-002 [12]	8191	60.99	52.7	55.25	○	○

TABLE III: Comparison of AI Embedding Models. Data retrieved from [6]

newspapers, a custom scraper was developed for each. The web listener checks the newspapers' sitemap for any new articles by generating a MD5 hash [13] ID using the article's textual and visual content. If both the ID and the URL are unique, then the article is added to the vector database. Suppose the ID is unique, but an entry with the same URL already exists in the database. In that case, it means that the article was modified after being initially indexed and that entry is updated. The webscraper script is implemented using python==3.9.19. The article HTML is downloaded using requests==2.32.3 and parsed using lxml==5.2.1.

Not all of the article text is used to create a vector representation. Table I shows how the article text is formatted before being embedded. search_document: is prepended to the string because the authors of nomic-ai-text-v1 trained the model to embed documents using that prefix [6]. Similarly, the prefix search_query: was used to train the model for document retrieval and is used instead when querying our text collection. In addition, the official Hugging Face page for this model [2] also mandates the use of the prefixes.

4) Thumbnail-Article Similarity: A similarity score for an image-text pair is calculated using blip-2. After the vector database is queried for similar articles, image-text similarity between the thumbnail and title of the article and between images and captions is calculated. The metadata of the articles and the corresponding scores are then passed to the front-end via an API in order to be displayed as points on a 1D spectrum in a pop-up.

To ensure that the generated similarity scores align with human interpretation, we conducted a ground-truth data collection survey involving 142 participants who provided 3765 annotations across 100 image-text pairs. We observed that blip-2 underestimates the similarity scores for weakly related pairs. As a result, we apply a logit-like transformation to the scores to allow the interpretation of low scores to align better with human tendency (See Fig. 13). We demonstrate our findings in Section II-B4.

5) GradCAM Heatmap: After image-text similarity is performed on an image-text pair (see Section I-B3), this process is repeated once more however this time through a custom GradCAM wrapper. This wrapper extracts the value attention scores from the BERT encoder's first layer in BLIP-2's QFormer module. This matrix of scalars is re-interpreted as a grayscale image by mapping the values between 0-255. Additionally, the grayscale image is resized

[2]https://bit.ly/47apwtr

(a) Example 1

(b) Example 2

Fig. 5: Examples of articles displayed image-text similarity spectrum. The green circle indicates the article being viewed by the reader, and the other circles are the news articles related to the same news event.

to the dimensions of the original image (as BLIP-2 resizes the image during preprocessing) and is passed to the front-end where it is overlayed on the original image (see Fig. 3c).

GradCAM allows the end-user to understand the contributions (or lack thereof) of the different salient parts of the image on the image-text similarity score. This feature introduces a level of explainability into the system, allowing users to understand the rationale behind a generated score as well as easily spot misinterpretations by the model.

C. Conclusion

In this chapter, we outline the implementation details of NewsAlign, one of the first bias-sensitive news aggregators that target picture-related media bias, a niche that is unexplored in current literature. The system is interfaced via a Chrome extension. A webscraper daemon periodically scrapes online articles, whose textual and visual content is saved in a vector database in order to facilitate similar article retrieval. A VLM (blip-2) is used to calculate image-text similarity scores on articles similar to the one the user is currently viewing. Additionally, GradCAM is used to extract attention scores from the VLM's layers, which produces a heatmap of the most salient parts of the image that are contributing the most to the generated image-text similarity score. These are then displayed to the front-end via the image-text similarity spectrum, effectively communicating discrepancies in the choice of image or how that image is presented within the article.

The main contribution of NewsAlign is the image-text similarity spectrum, which is the first feature in bias-sensitive news aggregators to communicate picture-related bias. Users are prompted to investigate outliers on the image-text similarity spectrum for any potential signs of visual bias, since they differ significantly from other articles discussing the event. Additionally, the GradCAM heatmap adds an element of explainability to NewsAlign. These elements demonstrate the potential that current AI technologies hold in reshaping the interdisciplinary field of bias-sensitive news aggregators.

II. Evaluation

A. Introduction

This chapter presents the evaluation of NewsAlign, our proposed bias-sensitive news aggregator that addresses picture-related media bias. The evaluation is structured into two main parts: a model evaluation and a system evaluation. In the model evaluation, we assess the effectiveness of Vision Language Models (VLMs) in predicting image-text similarity scores, comparing against our curated ground truth dataset. A total of 142 participants took part in the annotation process, which contributed to a total of 3679 valid annotations across 99 instances of thumbnail-headline pairs taken from the five newspapers considered for this study, namely Times of Malta, Malta Today, The Malta Independent, The Shift, Newsbook. In the system evaluation, we focus on assessing NewsAlign's impact on media bias awareness and news consumption habits. More specifically, we examine NewsAlign's key feature: the image-text similarity spectrum. This part of the evaluation incorporates both quantitative and qualitative methods, including a user study ($n = 22$) and semi-structured interviews with two experienced journalists. Throughout this chapter, we present our findings, discuss their implications, and address the limitations of our approach. We also explore potential areas for future development and research. By the end of this chapter, we aim to provide a clear understanding of NewsAlign's effectiveness in mitigating picture-related media bias and its potential impact on both journalists and news readers.

B. Model Evaluation

1) Overview: To evaluate the degree to which AI-generated scores match that of human judgment, a ground-truth data collection survey was conducted. The survey was accessible via a website and could be done on both desktop and mobile platforms. Participants were presented with randomly selected thumbnail-headline pairs from real news articles and prompted to rate the correlation between the two elements.

Before beginning, users were provided with a brief description of the task, required to consent to data participation, and asked to provide their age and gender for demographic analysis. The survey was limited to people over the age of 15. The image and the respective headline were placed vertically of each other (See Fig. 6). Their placement alternated with each new pair, in order to mitigate any potential biases regarding the placement of the content. The survey consisted of 100 image-text pairs; 20 articles sourced from each of the five newspapers[3] were considered for this study. Every submitted score was

[3]Times of Malta, Malta Today, The Malta Independent, Newsbook, The Shift

(a) Image-text pair with Likert-scale slider

(b) Page displayed every five image-text pairs

Fig. 6: Screenshots of the user interface of the ground-truth data collection process.

immediately recorded to cloud storage, and participants could continue annotating for as long as they wished. This ensured that responses were organic and representative of the participants' opinions. Additionally, a fun fact is displayed every 5 image-text pairs, in order to reduce fatigue from participants.

2) Results:

a) Data Collection and Cleaning: The title of one of the 100 image-text pairs shown to the users was in Maltese. This was an oversight by us, and since neither of the considered vision-language models natively support this language, data collected for that article would not be further considered during the analysis. As a result, only 19 news articles were considered for The Shift newspaper, instead of 20.

While some preliminary responses were received before mass dissemination, the bulk of the survey data was primarily collected between September 9$^{\text{th}}$ and September 14$^{\text{th}}$, 2024, through social media and word-of-mouth dissemination to the general public. A total of 142 participants

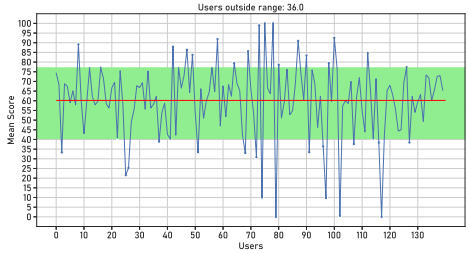


Fig. 7: Distribution of annotators' average score. The red line is the mean and outside the green bands are the 12.5$^{\text{th}}$ percentiles.

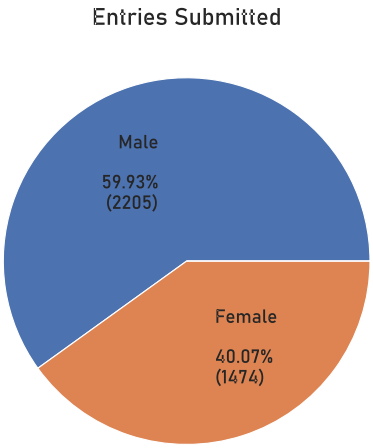


Fig. 8: Ratio of annotations by gender

contributed to the data collection process. To ensure data quality, we calculated the average rating per user and removed those whose ratings belonged to the bottom or top 12.5$^{\text{th}}$ percentile. As shown in Figure II-B2a, 36 users with abnormal average ratings were excluded to prevent outliers or incorrect data from affecting the analysis. This left a total of 106 legitimate users for further analysis.

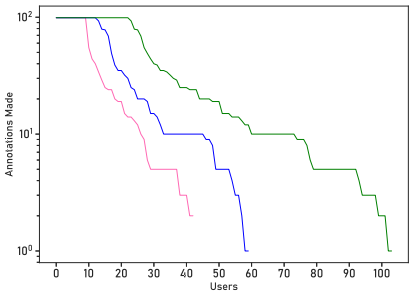


Fig. 9: Number of annotations by male and female users (Log Scale)

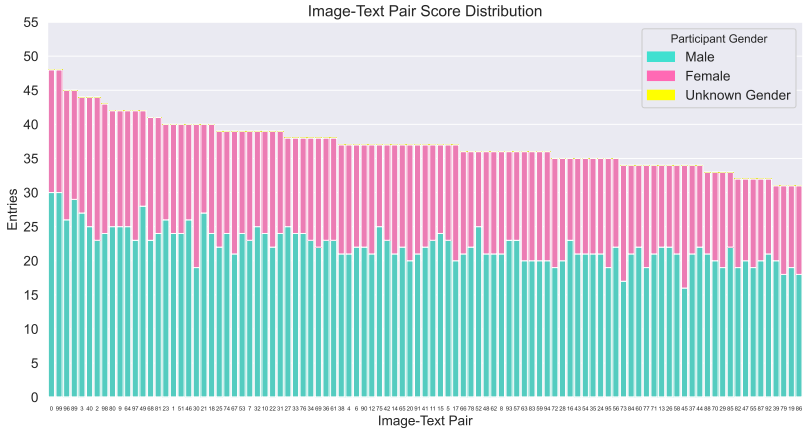


Fig. 10: Annotations per image-text pair by gender

b) Demographics and Response Patterns: A total of 3679 valid annotations were collected across the 99 image pairs. Around 60% of the respondents were male and the remaining 40% were female. A user's rating was recorded the moment it was submitted, despite the fact that the image-text pairs were given to users in batches of five. Moreover, the users could choose to stop at any moment of the survey, in order to ensure that the data collected was organic. Around 25% of respondents annotated all 99 image-text pairs, after which the number of annotations per user dropped exponentially (See Fig. 8). Additionally, Figure 10 shows the number of annotations per image-text pair by gender. No image-text pair received less than 31 annotations, with the highest being 48 annotations. Lastly, with respect to age, the majority of respondents were between 20 and 40 years old (See Fig. 11). The average age was 37, the youngest respondent was 20 years old, and the oldest was 88.

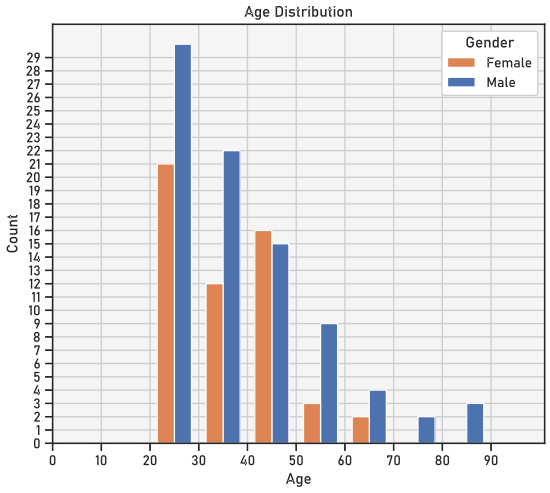


Fig. 11: Age distribution by gender.

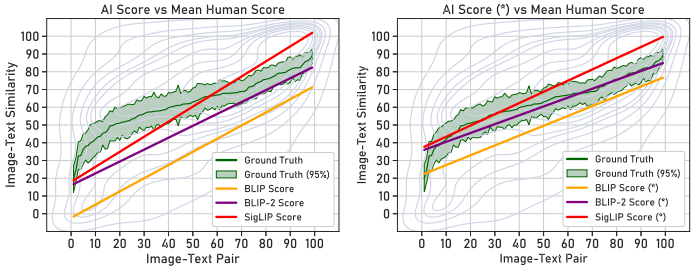


Fig. 12: Ground-truth comparisons for multiple models before (left) and after (right) alignment

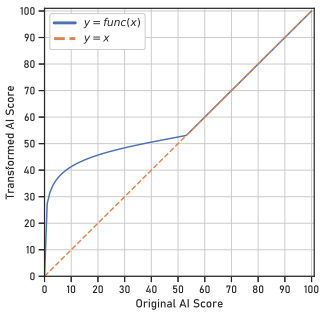
3) Evaluating various Vision-Language Models: We evaluate various vision-language models in order to determine which best predicts image-text similarity as close as possible to the human reviewers. We consider three models namely, BLIP [4], BLIP-2 [5] and SigLIP [3] (see Section ??). Our methodology is as follows: we average the different human scores per image-text pair, compute a similarity score using each of the three models, and plot the results.

As can be seen by Figure 12 and Table IV, all models demonstrated moderate correlation (~0.5) with human ratings. BLIP had the highest RMSE at 41.66%. BLIP-2 achieved the lowest RMSE at 35.64% and was selected for NewsAlign.

Model	Pearson Correlation Coefficient	Root Mean Squared Error
BLIP	0.51	41.66
BLIP-2	0.47	35.64
SigLIP	0.48	37.29
BLIP (*)	0.53	26.33
BLIP-2 (*)	0.47	24.08
SigLIP (*)	0.49	30.25

TABLE IV: Ground-truth analysis of AI scores and aligned AI scores (denoted by *).

Based on our observations, the models often assign scores close to 0, even when the image and text contain an albeit



$$y = \max\left(55 - \frac{x}{5} + 6\ln\left(\frac{x}{100 - x}\right), x\right)$$

Fig. 13: Transformation applied to AI data

weak relation, but enough to merit a higher score. The models tended to assign scores between 1% and 10% to image-text pairs with some degree of correlation. However, this score would be interpreted as being extremely poor by users of NewsAlign because it would be placed on the far end of the image-text similarity spectrum. In instances where the model is certain there is no correlation, the score produced is much closer to 0. To address this discrepancy, we implement a logit-inspired transformation function for scores below 55% (See Fig. 13). This function stretches scores between 0 and 10 to a range of 0 to 40, allowing weakly related pairs to be more accurately interpreted by the end-users of NewsAlign. In cases where the model is certain of no correlation, scores remain close to 0. After applying this transformation, the BLIP-2 scores align more closely with the ground-truth human ratings, as illustrated in Figure 12 and Table IV, which confirms the hypothesis that humans tend to expect the scores of weakly related image-text pairs to be much higher than what is produced by the VLM models. The reasoning for applying this transformation is to align the AI model's values to better fit human interpretability, while still preserving the meaning intended by the model. This alignment resulted in a significantly lower RMSE of 24.08. To ensure the robustness of our approach, we conduct additional tests to confirm that the specific parameters of the transformation are not overfitted to our sample data.

4) Evaluating transformation: Our validation process focuses on two key aspects: confirming that the transformation aligns the data more closely with human interpretability and ensuring that this alignment is not overfitted to the survey data. Figure 13 demonstrates that the linear trend of the aligned model outputs fits better with human scores compared to the raw outputs. Analysis of the residuals also reveals a substantial decrease in the number of scores that exceed a 33% difference from human ratings after applying the transformation (See Fig. 15). Additionally, the Root Mean Square Error (RMSE) for all models significantly decreased post-transformation,

```python
import numpy as np

def func(x: np.array):
    #Avoid log of 0
    x = x.clip(1e-9,100-1e-4)

    #Only apply transformation if larger than x
    x = np.maximum(
        55 - 0.2*x + 6*(np.log(x)-np.log(100-x)),
        x
    )

    x = x.clip(0,100)
    return x
```

Fig. 14: Python code of AI data alignment

without substantially affecting correlation (See Table IV).

Fold	Pearson Correlation Coefficient	Root Mean Squared Error
1	0.2	25.94
2	0.54	27.05
3	0.48	22.43
4	0.69	20.73
5	0.45	23.71

TABLE V: 5-Fold Cross Validation comparing BLIP-2 (*) with human scores

In order to ensure that our choice of transformation was not overfitted to the survey data's noise, we implement a cross-validation approach. Firstly, we sample points into two groups, perform the transformation on each, and observe that the transformation still aligns scores with the ground truth (See Fig. 16). Secondly, we conduct a more comprehensive 5-fold cross-validation to assess correlation and RMSE across different subsets of the data (See Table V). The metrics obtained from the cross validation accurately represented those described in Table IV.

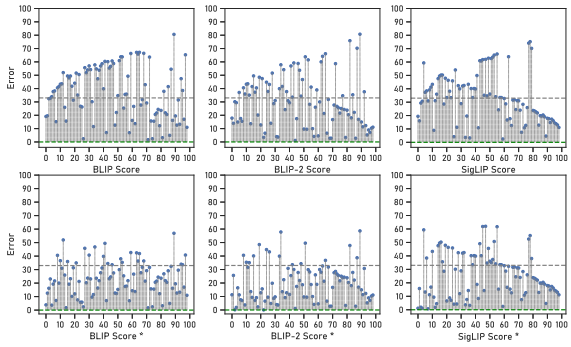


Fig. 15: Residuals of model scores before and after alignment

We theorize that the model manages to convey the similarity of the image-text pair to the user if its prediction is at least within 33% of the ground truth score. Since the spectrum visualization is implemented as a 1-D line,

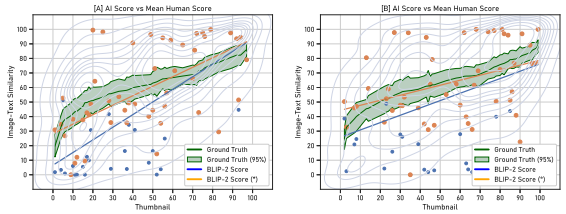


Fig. 16: Ground-truth comparisons for BLIP-2 on different subsets of the data, before and after alignment

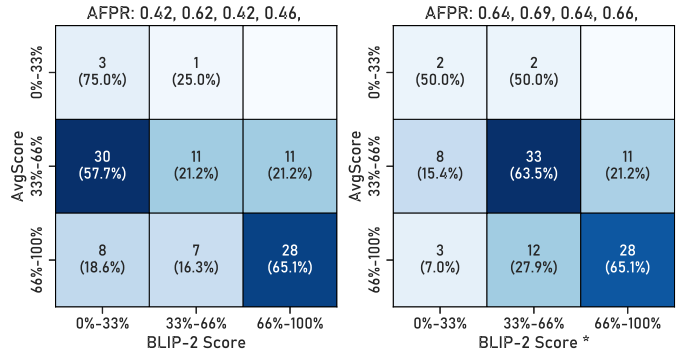


Fig. 17: Confusion Matrix of binned AI scores and human scores. Left: BLIP-2, Right: Aligned BLIP-2. AFPR means Accuracy, F1 Score, Precision, Recall.

users will tend to classify points on the spectrum as having low, medium, or high similarity. Consequently, we believe that the most crucial component for conveying relative similarities between the articles is the AI's ability to accurately predict which of these categories the article belongs to, even if the exact numerical scores differ from human tendency. We bin the human and AI scores into three bins of 33 percentage points each and plot a confusion matrix to analyze the results. The results demonstrate that the aligned model outperforms the raw scores in this binned analysis (See Fig. 17). Specifically, the aligned model achieved an accuracy of 64% and an F1 score of 69%. This result motivates us to believe that the model will succeed in depicting the image-text similarity to the user with decent accuracy. Encouraged by these findings, we label the image-text similarity spectrum into three equally spaced sections as documented in Section I-B1.

a) Conclusion: This chapter evaluated the choice of vision-language model for image-text similarity scoring, as well as focusing on aligning AI-generated scores with human judgments. The evaluation process began with a ground-truth data collection survey, which gathered 3765 annotations across 99 image-text pairs from the five newspapers.

This dataset provided a foundation for us to align our AI generated scores with human interpretability. We evaluate three vision-language models (BLIP, BLIP-2, SigLIP) and find that BLIP-2 gives the best results. The data from the survey showed that participants had a different tendency to rate things as opposed to BLIP-2. As a result, we apply a logit-like function which stretches similarity scores between 0% and 10% to a range of 0% and 40%. This reduces the RMSE of BLIP-2 from 35.64% to 24.08% while also maintaining the same correlation to the crowd-sourced scores. We ensure that the transformation function is not overfitted to our sample data by performing 5-fold cross validation on the data. Additionally, we theorise the idea that humans will tend to subconsciously split the similarity scores into three categories: low, medium, and high. When binning the crowd-sourced and AI-generated scores into these categories, we find that the model and the transformation achieve an accuracy of 64% and an F1 score of 69% in predicting the categories.

C. NewsAlign Evaluation

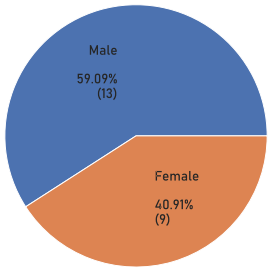
This part of the evaluation aimed to analyse public opinion of the system as well as whether using the system during news browsing leads to increased awareness of media bias, encourages reading more sources about events, and prompts more thought regarding image use and explanations. This evaluation took two approaches: one from the journalists' perspective through in-depth interviews and another from the news readers' perspective after a control group used the software.

Journalist Perspective: We conducted two semi-structured interviews with journalists from Malta: Mr Neil Camilleri and Mr Bertrand Borg. Camilleri is an independent journalist with twenty years of experience, having worked in two newsrooms: Media Link Communications and The Malta Independent. Borg is the online editor for Times of Malta, the country's largest newspaper, and has been working there for about ten years. The interviews focused on gathering their opinions about the state of the Maltese news sphere, their views on visual and textual bias, their experience with AI, and their hopes for and potential of NewsAlign.

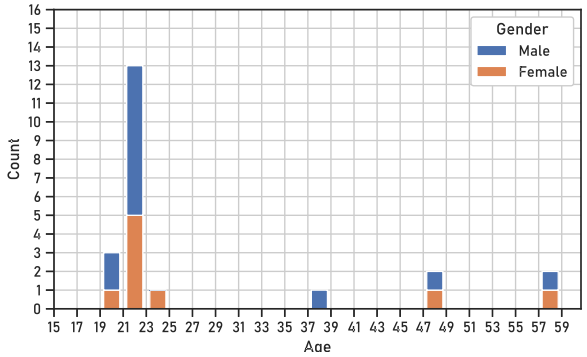
News Reader Perspective: We conducted a user study ($n = 22$) with participants from different backgrounds, age, and gender. Participants were first introduced to the system's functionality and visualisations, then allowed to experiment with it at their own pace. Afterward, they were prompted to complete a questionnaire to gather their opinions about the system. NewsAlign retrieved textually similar articles that discuss the same event as the key article, performed image-text similarity on each of the article's thumbnail and headline, and then displayed those articles on the image-text similarity spectrum. While NewsAlign supports performing image-text similarity on non-thumbnail images, we detail the design choice to restrict NewsAlign on thumbnail-headline pairs in Section 5.4.2.

A total of 5 questions were asked. The user study aimed to determine whether the visualisation effectively com-

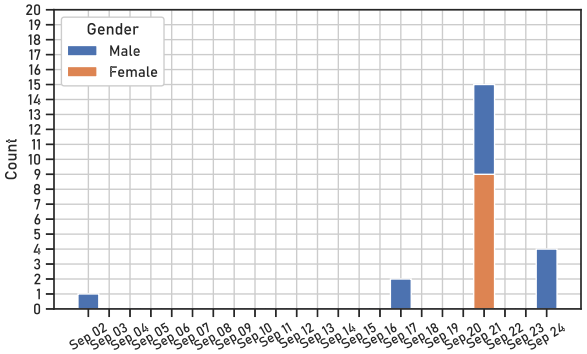
municates picture-related bias and whether it encourages users to read more articles about the event and notice discrepancies between image choices and headlines. Users responded to questions using a 6-point Likert scale, ranging from No to Yes. The complete list of questions and collected responses are shown in VI. Note that the GradCAM heatmap feature was implemented in NewsAlign based on the findings from the evaluation process after the fact. As a result, this section does not evaluate this feature as it was not present during the time. The effectiveness of a system like GradCAM in bias-sensitive news aggregators such as NewsAlign is an avenue for future research.



(a) Ratio of respondents by gender



(b) Age of participants by gender



(c) Date of response received, by gender

Fig. 18: Demographics of NewsAlign User Testing ($n = 22$)

The data collected from the 22 participants was gathered during the month of September 2024. The date range over which data was collected can be seen in Figure 18c. Additional demographic data, including the ratio of respondents by gender and the age distribution of participants by gender, can be seen in Figures 18a and 18b, respectively. From the 22 participants, 13 were male and 9 were female. The bulk of the ages hovered around early 20s. The youngest respondent was 19 while the oldest was 58.

1) Image-Text Similarity Spectrum: Due to the fact that many different news organisations will have their inherent bias and give different renditions of the event, the primary strategy to mitigate media bias is to read multiple sources. This allows readers to form a more holistic opinion based on diverse perspectives. The image-text similarity spectrum in NewsAlign aims to support this approach by aggregating relevant articles and presenting them according to perceived image-text similarity. This feature intends to make users more aware of how different articles about the same event are presenting themselves.

Regarding visual bias in Malta's news landscape, Camilleri noted that while some newspapers lean towards certain political spectrum, most questionable image use stems from the urgency to attract clicks and stand out from competitors, especially in tabloid-like publications. Both Camilleri and Borg highlighted that unintentional bias often results from limited access to stock libraries, content management systems, or dedicated photographers. As a result, newsrooms depend on what is already publicly available, or photos that get sent in. This constraint sometimes leads to the use of repeated images or sub-optimal visual choices. This is especially true when newsrooms are pressed for time to upload a story, causing poor editing choices. Borg mentions that they "very often" get complaints from politicians regarding a poor choice of photo. Borg emphasises his media team does not do this intentionally, but occurs due to limited visual content surrounding that person.

> " Even if it's a certain politician, the same photo of that politician would appear very often, not because [we] love that image but because it was very easily accessible... We get complaints very often from politicians about the images we use. 'Why did you pick my- I look bad in this photo', you know, this kind of thing. We don't do that intentionally; it's a bit subjective really. " - Bertrand Borg

The image-text similarity spectrum aims to highlight problematic image-text pairs, flagging them to journalists before publication and drawing readers' attention to them. Camilleri stated that he had never encountered a system like this for highlighting visual bias instances, and is not aware of any similar tools actively being used by journalists. Borg mentioned a similar system at the Financial Times for

№	Question	Likert Scale					
		1	2	3	4	5	6
1	Do you like this feature?	0	0	1	2	7	12
2	This feature made it easier to read multiple sources about an event.	0	0	2	2	4	14
3	Outliers on the image-similarity spectrum made me think about the choice of headline/thumbnail.	0	0	1	2	8	11
4	How the different articles were placed on the image-similarity spectrum made sense.	0	0	1	2	5	14
5	The image similarity spectrum was an effective indicator that made me think about the choice of headline or thumbnail.	0	0	1	0	8	13

TABLE VI: Questions asked and responses gathered from a survey evaluating NewsAlign. ($n = 22$)

balancing gender representation in images, but confirmed that Times of Malta doesn't currently use a bias mitigation system like NewsAlign. Borg states that NewsAlign is very useful when it comes to analysing textual or visual bias, and also suggests future improvements such as finetuning the image-text similarity score on more contextualised content for Malta.

Participants of the user study responded positively to the feature, with the majority answering a score of 4 or higher on the Likert Scale. Additionally, the majority of participants stated that this feature made it easier to research multiple sources about an event due to the fact that event-level aggregation is done automatically. In his interview, Camilleri states that he believes that NewsAlign has the potential to make readers less affected by bias and make journalists more aware if they are producing biased content, but it is ultimately up to the user to adopt the system. From these results, we can conclude that the image similarity feature successfully encourages users to read more sources.

Examples of bias-sensitive news aggregators from literature highlight the importance of the user interface and visualisation, as it is the main communication channel for conveying the differences between the event-level articles. With respect to image bias, the majority of participants stated that outliers on the image-text similarity spectrum made them question the choice of image (Question 3) and also found the spectrum effective in conveying image-text discrepancies (Question 5). Additionally, Camilleri praised the system for its ability to quickly compare headlines and thumbnails from multiple articles together, a task that can be time-consuming for journalists. He mentions that the job's stresses and urgency makes it difficult for this type of work to be done manually, and a system such as NewsAlign is perfect for communicating to the journalist problematic instances of image-text pairs. Borg compares NewsAlign to another bias mitigation system called GroundNews. He states that NewsAlign allows its users to see the current article's thumbnail in the context of all the other article thumbnails, something which according to our knowledge, GroundNews does not yet support. Additionally, he states that only the Times of Malta is listed in one of the biased organisations that GroundNews operates with, meaning applying Ground News for Maltese newspapers is not yet feasible at this moment, unlike NewsAlign. From the responses gathered from both the journalists' and readers' side, we believe that the image-text similarity spectrum is effective at communicating the different nuances in image and headline use across the other articles.

In Section II, we performed a study to determine the effectiveness of blip-2 in predicting image-text similarity scores for articles, as well as how interpretable these are by humans. In Question 4, we ask respondents to rate how cohesive they found the similarity score to be while using the system. The majority of respondents stated that they found the similarity scores to be sensible, with 14 out of the 22 participants giving the highest score possible. The underlying vision-language model responsible for generating the image-text similarity scores is not perfect, as demonstrated by our previous evaluation as well as by the fact that the model did not undergo additional finetuning for this use case. While it is reassuring to know that similarity scores were interpreted as being sensible, additional work is still required to adapt the model to this downstream task. Borg proposes allowing users of the system to submit certain false-positives that they encounter, so the model will be able to predict them correctly in its next iteration. Additionally, an inherent limitation of the current model is its inability to associate public figures with their societal positions. As a result, image-text pair instances that feature a face of a noteworthy person will produce a score no different than if any other face was shown on the image. Borg noted that faces are more likely to feature in images, as his news organisation experiences an increase in readership when faces are used. Lastly, the model is not expected to produce accurate results when dealing with domain-specific knowledge that readers are well aware of but may not have been included in the model's training data.

D. Discussion

The section aims to interpret the key findings from our evaluation of NewsAlign, highlighting notable points and insights that emerged from both the user study and qualitative interviews with journalists. We will examine

the effectiveness of the image-text similarity spectrum, discuss the potential impact of NewsAlign on media bias awareness and news consumption habits, and consider the implications for both journalists and news readers. Additionally, we will address limitations of the current system and propose potential areas for future development and research.

1) Model Evaluation: In order to evaluate and align the VLM model, we compared 3679 annotations made by 106 users over 99 image-text pairs sampled from the five newspapers NewsAlign supports: Times of Malta, Malta Today, The Malta Independent, The Shift, and Newsbook. We believe that the annotations collected are representative of the population. In terms of age distribution, the bulk of the respondents are between 20 and 40 years old; however, the data also includes responses from individuals up to 88 years old. Additionally, there is only a ~19.87% disparity in gender, which, while it could be improved, still represents a good balance.

The National Statistics Office of Malta stated in 2011 that around 30.5% of the population read at least one news article every day in the last 12 months preceding their survey[4]. Taking into consideration the population of Malta in 2011 (416,268)[5], we can calculate that with a sample size of 106 participants, there is a 95% chance that the real value is within 8.76% of the survey value. While these figures are more than 10 years old, they still serve as a good indicator of the effectiveness of the study.

To find the best VLM model capable of predicting image-text similarity, we ran three VLM models on the same image-text pairs used in the annotation process namely, BLIP [4], SigLIP [3], and BLIP-2 [5]. We calculated the RMSE and Pearson's correlation coefficient between the predicted and the actual scores, finding that BLIP-2 produces the least RMSE of 35.64%. From our observations, we found that humans tend to rate weakly related image-text pairs higher than what the model outputs. As a result, we applied a logit-like transformation (See Fig. 13) on the lower half of the scores to align the output of the AI model to better match the scores given by humans, making them more interpretable.

A possible limitation to this approach of aligning the VLM model with the annotation data is that the alignment risks overfitting to a sub-sample of the population. If this is the case, it means that if the rest of the general population has a different tendency to rate the correlation between image-text pairs, then the scores of NewsAlign might not be representative of the rest of the population. Although we believe that the figures achieved are sufficient for the scope of this study, improvements can be made to ensure the data is even more representative. Annotation data can easily be increased due to the fact that the annotation process is hosted entirely online and in the public domain. Additionally, extra attention has been given to making the process easy and fatigue-less for annotators. Future work can involve increasing annotation data through the website to acquire stronger results.

Another counterpoint to aligning model outputs using annotations collected from the general population is that the general population may inherently be biased. As a result, it's possible that the model outputs were aligned to be more biased than they were previously. To address this concern, we consider bias to be defined as the deviation from the ground truth. However, there is no factually objective ground truth since we must depend on human recounts of events. Therefore, we consider the approximate mean value of all different accounts of the story to be the ground truth. Consequently, the more the model aligns with the average human tendency to interpret an event, the more it is considered accurate, as the average human tendency models and represents the ground truth.

Secondly, since bias is a relative concept and not absolute, if the general population, ranging from domain experts to inexperienced individuals, considers an object to be biased (even if, by some idealised Oracle-like entity, the object is unbiased), then the model output should state that it is biased. Ultimately, the model must embody the definitions, beliefs, and ideas of the general population; otherwise, its outputs are not human-interpretable, and are acting according to some other non-human interpretation. All of this is built on the assumption that the annotation data we base our findings on is representative of the general population.

Additionally, to ensure that the choice of transformation was not overfitted on the data, we visualised two random splits of the data and found that the transformation still aligns the model outputs to the human scores. Additionally, we performed a 5-fold cross-validation and found that the RMSE of the different folds matches the RMSE of the entire dataset. We also plotted the residuals of the raw model outputs and the aligned model outputs, finding that the aligned outputs reduce the majority of the errors to under 33% (See Fig. 15). We theorise that humans will tend to interpret the scores on the image-text similarity spectrum as having low, middle, or high accuracy. Motivated by this idea, we binned the average human score per image-text pair and the aligned model outputs into three equal bins. We found that the model's accuracy in predicting the correct category of the similarity score increases from 42% to 64% after applying the alignment process. While this supports our evidence that the model transformation will lead to more accurate scores, there are still some limitations that need addressing. Firstly, in our ground-truth dataset, there is a class imbalance specifically for scores ranging between 0% and 33% (i.e., the bottom third of the scores). While it could be the case that humans tend not to rate articles in this category, it's also possible that the image-text pairs chosen for the annotation did not include any

[4]https://nso.gov.mt/wp-content/uploads/CultureParticipationSurvey.pdf
[5]https://nso.gov.mt/wp-content/uploads/Census2011_FinalReport.pdf

articles that the annotators would have categorised in this range.

Additionally, after applying the transformation on the BLIP-2 outputs, the model's accuracy in predicting scores belonging to the bottom category decreased due to a misclassified instance. Although one misclassified article might not be indicative enough that the model will perform poorly in this range, additional research is needed. Preferably, another set of image-text pairs should be chosen such that there are equal amounts of articles in each category since randomly sampling thumbnail-headline pairs from established newspapers will have a tendency to include more pairs with good similarity due to the fact that these newspapers are already established and popular, and hence have acquired a decent reputation. Moreover, while 64% is a good accuracy for this study (around double the score of the random baseline), there is still room for improvement. For example, additional studies can fine-tune the model on image-text pairs specifically taken from a news context, as image-text pairs outside this context are presented differently than in news articles, which can produce inaccurate similarity scores.

In conclusion, the evaluation of the potential of VLMs as a tool to automate image-text similarity scores produced promising results for further adoption and research using this technique. However, additional improvements in the methodology used to evaluate and align the VLM model warrant future research.

2) NewsAlign Evaluation: Similar bias-sensitive news aggregators in literature evaluate their studies using an approach comparable to ours. Studies like NewsBird [14] and NewsCube 2.0 [15] use a user study with around 20 participants, similar to our methodology. In the user study, we intentionally used an even-numbered Likert scale to prevent users from selecting a neutral middle option, effectively forcing a decisive response. This decision was made since we were expecting a significantly less participants compared to the annotation survey, ensuring that each response contributed meaningful data. While this approach ensures that all participants provide a clear opinion, it may force users who are genuinely undecided to choose between two ends of the scale. On the contrary, participants were allowed to use the system for as long as they wished before providing feedback, allowing them ample time to form a more informed opinion. Although ~77.3% of the responses are in the 20-25 age range, we also include responses from participants up to 58 years old.

The user study focused on the following core aspects:

1) Image-Text Similarity Spectrum
   i. Encouraging readers to consult multiple sources about an event (Question 2).
   ii. Effectively communicating picture-related bias (Questions 3 and 5).
   iii. Ensuring the similarity scores on the spectrum were comprehensible to readers (Question 5).

The positive reception of the above core aspects from both reader and journalist perspectives, as detailed in Section II-C, leads us to conclude that Objective 3 ($O3$) was achieved by our study. While we are satisfied with the results, we also acknowledge areas for potential improvement. One limitation concerning the system's explainability is that it might not always be clear to users why the model assigns a specific score to an image-text pair. There needs to be a more effective indicator to communicate to the user if the model failed to recognise something in the image. A heatmap could be overlayed over the thumbnail to address this, highlighting the most relevant parts according to the model. This could be achieved by implementing a system similar to GradCam [1], which computes a heatmap based on the model's attention weights for different parts of the image.

Regarding the image-text similarity spectrum, Borg suggests incorporating a machine learning component to improve the system's understanding over time. Implementing this feature would require an additional backend module to store flagged image-text pairs and their intended similarity scores. The model could then be finetuned on these instances, learning to associate positive examples and disassociate negative ones. However, safeguards would be necessary to prevent abuse and the introduction of new biases. This warrants a separate study to ensure proper implementation.

While the system that was previewed by the journalists and the readers is still not refined for mass adoption, valuable suggestions were made for improving the usability of the system for a wider implementation. Camilleri notes that NewsAlign has potential to mitigate problematic image-text pairs, but fears many journalists won't be so eager to integrate it into their workflows, citing the hectic nature of journalistic work and time constraints. In light of this claim, future studies aiming to distribute a similar system to a wider audience should consider streamlining how the system is interfaced. Camilleri also mentions skepticism surrounding AI as well as the potential reluctance of some newsrooms to adopt new technologies. This aligns with reviewed literature from countries such as Jordan and Pakistan, that claim a lack of knowledge about AI and poor data infrastructure are some of the major concerns regarding AI adoption. Other additional barriers that might discourage newsrooms from adopting AI are significant financial costs, lack of human resources, and fears over job losses. Hence research investigate the application of a system like NewsAlign from theory to practice should consider these important points.

During the user study, respondents could chose to leave any additional comments they wish to make regarding NewsAlign. A notable comment suggesting a future improvement to NewsAlign was to improve the design of the visuals as users with less technical aptitude may find it

hard to intuitively interpret the results of the image-text similarity spectrum. Responding to a question asking if NewsAlign can be misinterpreted as attacking journalists, Camilleri believes that while some might get offended, citing a "big ego" among journalists, the majority would view it positively. Borg also suggests adding a disclaimer before the system is used to prevent misinterpretation and suggests that the language used could be adjusted to avoid negative reactions.

We believe that NewsAlign successfully manages to integrate VLM to provide valuable insights on picture-related bias, successfully accomplishing Objective 1 ($O1$) and Objective 2 ($O2$). Nevertheless, we also note possible improvements to further exploit the potential of VLMs in addressing picture-related bias. As mentioned in Section I, NewsAlign supports retrieving articles by similar images. However, we observed that many articles do not caption their images with descriptive information regarding the visual content. As a result, the images of these articles cannot be evaluated on the image-similarity spectrum, making the feature difficult to evaluate. Applying this feature only for the thumbnail of the article, i.e searching similar articles by observing the thumbnail instead of the textual content, has the potential to provide insights into how newspapers generally present the same (or visually similar) image across multiple weakly related stories.

Despite the constructive criticism received about NewsAlign, both the reader and journalist perspectives reacted positively to the impact NewsAlign can have on news content production and consumption, indicating to us that we have achieved Objective 3 ($O3$). While we are satisfied with the results, we also acknowledge the potential for further improvements to enhance the system's effectiveness and user experience.

E. Conclusion

The evaluation of NewsAlign has provided valuable insights into the effectiveness of one of the first bias-sensitive news aggregators that addresses picture-related media bias. We have introduced a novel dataset consisting of 3679 valid annotations across 99 thumbnail-headline pairs taken from real news articles. Additionally, we evaluated the feasibility of different VLMs in predicting these ground-truth similarity scores. When binning the scores into three categories (low, medium, high similarity), we found that BLIP-2 is able to achieve an accuracy of 64% in predicting the correct category. Additionally, we evaluated applying a transformation to the raw BLIP-2 scores to improve interpretation of the scores, while still maintaining the underlying meaning of the model.

The system evaluation, incorporating both reader and journalist perspectives, revealed a positive reception of NewsAlign's core features. We conducted semi-structured interviews with two experienced journalists as well as a user study with 22 participants. The interviews revealed valuable insights about media bias in Maltese newspapers, particularly how visual bias is often unintentional, stemming from limited resources and time constraints. NewsAlign's image-text similarity spectrum was found to effectively encourage users to consult multiple sources and raised awareness of potential picture-related bias. We also highlighted areas for improvement, such as enhancing the system's explainability, implementing user feedback mechanisms for continuous learning, and refining the user interface. The positive feedback from both readers and journalists suggests that such tools have the potential to play a valuable role in the future of digital journalism and news consumption. Ongoing research and development will be essential to refine these technologies and tackle the challenges of media bias in the digital age.

References

[1] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," International Journal of Computer Vision, vol. 128, no. 2, p. 336–359, Oct. 2019. [Online]. Available: http://dx.doi.org/10.1007/s11263-019-01228-7

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[3] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 11 975–11 986.

[4] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in International conference on machine learning. PMLR, 2022, pp. 12 888–12 900.

[5] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in ICML, 2023.

[6] Z. Nussbaum, J. X. Morris, B. Duderstadt, and A. Mulyar, "Nomic embed: Training a reproducible long context text embedder," arXiv preprint arXiv:2402.01613, 2024.

[7] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "Mteb: Massive text embedding benchmark," 2023. [Online]. Available: https://arxiv.org/abs/2210.07316

[8] J. Saad-Falcon, D. Y. Fu, S. Arora, N. Guha, and C. Ré, "Benchmarking and building long-context retrieval models with loco and m2-bert," 2024. [Online]. Available: https://arxiv.org/abs/2402.07440

[9] M. Günther, L. Milliken, J. Geuter, G. Mastrapas, B. Wang, and H. Xiao, "Jina embeddings: A novel set of high-performance sentence embedding models," 2023. [Online]. Available: https://arxiv.org/abs/2307.11224

[10] M. Günther, J. Ong, I. Mohr, A. Abdessalem, T. Abel, M. K. Akram, S. Guzman, G. Mastrapas, S. Sturua, B. Wang, M. Werk, N. Wang, and H. Xiao, "Jina embeddings 2: 8192-token general-purpose text embeddings for long documents," 2023.

[11] OpenAI, "New embedding models and api updates," https://openai.com/index/new-embedding-models-and-api-updates/, 01 2024, accessed: 2024-09-18.

[12] ——, "New and improved embedding model," https://openai.com/index/new-and-improved-embedding-model/, 12 2022, accessed: 2024-09-18.

[13] R. Rivest, "The md5 message-digest algorithm," MIT Laboratory for Computer Science, Tech. Rep., 1992.

[14] F. Hamborg, N. Meuschke, and B. Gipp, "Matrix-based news aggregation: exploring different news perspectives," in 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE, 2017, pp. 1–10.

[15] S. Park, M. Ko, J. Kim, H.-J. Choi, and J. Song, "Newscube 2.0: an exploratory design of a social news website for media bias mitigation," in Workshop on Social Recommender Systems, 2011, pp. 1–5.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 11 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[17] J. S. Coleman, "Relational analysis: The study of social organizations with survey methods," Human organization, vol. 17, no. 4, pp. 28–36, 1958.

[18] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2641–2649.

[19] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 3558–3568.

[20] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," International journal of computer vision, vol. 123, pp. 32–73, 2017.

[21] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[22] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," arXiv preprint arXiv:2104.08718, 2021.

[23] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.

[24] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2556–2565.

[25] S. W. Rosenberg, L. Bohan, P. McCafferty, and K. Harris, "The image and the vote: The effect of candidate presentation on voter preference," American Journal of Political Science, pp. 108–127, 1986.

[26] B. Van Gorp, "Where is the frame? victims and intruders in the belgian press coverage of the asylum issue," European journal of communication, vol. 20, no. 4, pp. 484–507, 2005.

[27] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," arXiv preprint arXiv:1109.2378, 2011.

[28] S. R. Sommers, E. P. Apfelbaum, K. N. Dukes, N. Toosi, and E. J. Wang, "Race and media coverage of hurricane katrina: Analysis, implications, and future research questions," Analyses of Social Issues and Public Policy, vol. 6, no. 1, pp. 39–55, 2006.

[29] OpenAI, "Introducing gpts," https://openai.com/index/introducing-gpts/, 11 2023, accessed: 2024-09-18.

[30] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, and B. Zoph, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[31] G. Team, "Gemma," 2024. [Online]. Available: https://www.kaggle.com/m/3301

[32] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl, A. Benhaim, M. Bilenko, J. Bjorck, S. Bubeck, M. Cai, Q. Cai, V. Chaudhary, D. Chen, D. Chen, W. Chen, Y.-C. Chen, Y.-L. Chen, H. Cheng, P. Chopra, X. Dai, M. Dixon, R. Eldan, V. Fragoso, J. Gao, M. Gao, M. Gao, A. Garg, A. D. Giorno, A. Goswami, S. Gunasekar, E. Haider, J. Hao, R. J. Hewett, W. Hu, J. Huynh, D. Iter, S. A. Jacobs, M. Javaheripi, X. Jin, N. Karampatziakis, P. Kauffmann, M. Khademi, D. Kim, Y. J. Kim, L. Kurilenko, J. R. Lee, Y. T. Lee, Y. Li, Y. Li, C. Liang, L. Liden, X. Lin, Z. Lin, C. Liu, L. Liu, M. Liu, W. Liu, X. Liu, C. Luo, P. Madan, A. Mahmoudzadeh, D. Majercak, M. Mazzola, C. C. T. Mendes, A. Mitra, H. Modi, A. Nguyen, B. Norick, B. Patra, D. Perez-Becker, T. Portet, R. Pryzant, H. Qin, M. Radmilac, L. Ren, G. de Rosa, C. Rosset, S. Roy, O. Ruwase, O. Saarikivi, A. Saied, A. Salim, M. Santacroce, S. Shah, N. Shang, H. Sharma, Y. Shen, S. Shukla, X. Song, M. Tanaka, A. Tupini, P. Vaddamanu, C. Wang, G. Wang, L. Wang, S. Wang, X. Wang, Y. Wang, R. Ward, W. Wen, P. Witte, H. Wu, X. Wu, M. Wyatt, B. Xiao, C. Xu, J. Xu, W. Xu, J. Xue, S. Yadav, F. Yang, J. Yang, Y. Yang, Z. Yang, D. Yu, L. Yuan, C. Zhang, C. Zhang, J. Zhang, L. L. Zhang, Y. Zhang, Y. Zhang, Y. Zhang, and X. Zhou, "Phi-3 technical report: A highly capable language model locally on your phone," 2024. [Online]. Available: https://arxiv.org/abs/2404.14219

[33] C. Fourrier, N. Habib, A. Lozovskaya, K. Szafer, and T. Wolf,

"Open llm leaderboard v2," https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024.
[34] E. Beeching, C. Fourrier, N. Habib, S. Han, N. Lambert, N. Rajani, O. Sanseviero, L. Tunstall, and T. Wolf, "Open llm leaderboard (2023-2024)," https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard, 2023.
[35] L. Gao, J. Tow, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, K. McDonell, N. Muennighoff, J. Phang, L. Reynolds, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou, "A framework for few-shot language model evaluation," Sep. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5371628
[36] J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou, "Instruction-following evaluation for large language models," 2023. [Online]. Available: https://arxiv.org/abs/2311.07911
[37] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, and J. Wei, "Challenging big-bench tasks and whether chain-of-thought can solve them," 2022. [Online]. Available: https://arxiv.org/abs/2210.09261
[38] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2201.11903
[39] Z. Sprague, X. Ye, K. Bostrom, S. Chaudhuri, and G. Durrett, "Musr: Testing the limits of chain-of-thought with multistep soft reasoning," 2024. [Online]. Available: https://arxiv.org/abs/2310.16049
[40] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen, "Mmlu-pro: A more robust and challenging multi-task language understanding benchmark," 2024. [Online]. Available: https://arxiv.org/abs/2406.01574
[41] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," 2022.
[42] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.
[43] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan, "Constitutional ai: Harmlessness from ai feedback," 2022.
[44] Y. Bar-Haim, D. Lamy, L. Pergamin, M. J. Bakermans-Kranenburg, and M. H. Van Ijzendoorn, "Threat-related attentional bias in anxious and nonanxious individuals: a meta-analytic study." Psychological bulletin, vol. 133, no. 1, p. 1, 2007.
[45] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," Computational linguistics, vol. 16, no. 2, pp. 79–85, 1990.
[46] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.
[47] N. Newman, R. Fletcher, C. T. Robertson, A. R. Arguedas, and R. Kleis Nielsen, Reuters Institute Digital News Report 2024, 2024.
[48] M. M. Butts, D. C. Lunt, T. L. Freling, and A. S. Gabriel, "Helping one or helping many? a theoretical integration and meta-analytic review of the compassion fade literature," Organizational Behavior and Human Decision Processes, vol. 151, pp. 16–33, 2019.
[49] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in European conference on computer vision. Springer, 2020, pp. 104–120.
[50] J. Chen, H. Guo, K. Yi, B. Li, and M. Elhoseiny, "Visualgpt: Data-efficient adaptation of pretrained language models for image captioning," 2022.
[51] F.-L. Chen, D.-Z. Zhang, M.-L. Han, X.-Y. Chen, J. Shi, S. Xu, and B. Xu, "Vlp: A survey on vision-language pre-training," Machine Intelligence Research, vol. 20, no. 1, pp. 38–56, 2023.
[52] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014.
[53] D. D'Alessio and M. Allen, "Media bias in presidential elections: A meta-analysis," Journal of communication, vol. 50, no. 4, pp. 133–156, 2000.
[54] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.
[55] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
[56] R. M. Entman, "Framing: Toward clarification of a fractured paradigm," Journal of communication, vol. 43, no. 4, pp. 51–58, 1993.
[57] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "Gptq: Accurate post-training quantization for generative pre-trained transformers," arXiv preprint arXiv:2210.17323, 2022.
[58] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," Biological cybernetics, vol. 36, no. 4, pp. 193–202, 1980.
[59] M. Khlouf, "Impact of employing artificial intelligence on media institutions in palestine from the viewpoint of those in charge of communication," An-Najah University Journal for Research-B (Humanities), vol. 38, no. 6, pp. 1093–1120, 2023.
[60] S. Jamil, "Artificial intelligence and journalistic practice: The crossroads of obstacles and opportunities for the pakistani journalists," Journalism Practice, vol. 15, no. 10, pp. 1400–1422, 2021.
[61] M. Al Jwaniat, D. Tahat, R. AlMomany, K. Tahat, M. Habes, A. Mansoori, and I. Maysari, "Examining journalistic practices in online newspapers in the era of artificial intelligence," in 2023 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS). IEEE, 2023, pp. 183–189.
[62] K. A. Finnset, "Artificial intelligence in norwegian newsrooms a qualitative study on the uses and assessments of ai technologies in a news context," Master's thesis, University of Oslo, 2020.
[63] Y. Jeon, J. Kim, S. Park, Y. Ko, S. Ryu, S.-W. Kim, and K. Han, "Hearhere: Mitigating echo chambers in news consumption through an ai-based web system," Proceedings of the ACM on Human-Computer Interaction, vol. 8, no. CSCW1, pp. 1–34, 2024.
[64] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, 2016, pp. 1480–1489.
[65] R. Fletcher and R. K. Nielsen, "What does the public in six countries think of generative ai in news?" 2024. [Online]. Available: https://reutersinstitute.politics.ox.ac.uk/what-does-public-six-countries-think-generative-ai-news

[66] N. Y. Times, "Editor (2015)," 2015. [Online]. Available: https://nytlabs.com/projects/editor.html
[67] C. Beckett, New powers, new responsibilities: A global survey of journalism and artificial intelligence, 2019. [Online]. Available: https://blogs.lse.ac.uk/polis/2019/11/18/new-powers-new-responsibilities/
[68] Brik, "The attitudes of the communicators towards the use of artificial intelligence techniques in the egyptian and saudi journalistic institutions, a field study within the framework of the unified theory for acceptance and use of technology (utaut)," in Media Res. J, vol. 53, no. 2, 2022, p. 447–526.
[69] S. G"achter, H. Orzen, E. Renner, and C. Starmer, "Are experimental economists prone to framing effects? a natural field experiment," Journal of Economic Behavior & Organization, vol. 70, pp. 443–446, 2009.
[70] M. Gentzkow, J. M. Shapiro, and D. F. Stone, "Media bias in the marketplace: Theory," in Handbook of media economics. Elsevier, 2015, vol. 1, pp. 623–645.
[71] S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. Del Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi et al., "Textbooks are all you need," arXiv preprint arXiv:2306.11644, 2023.
[72] F. Hamborg, K. Donnay, and B. Gipp, "Automated identification of media bias in news articles: an interdisciplinary literature review," International Journal on Digital Libraries, vol. 20, no. 4, pp. 391–415, 2019.
[73] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[74] G. Hili and D. Seychell, "Using machine learning to investigate potential image bias in news articles," in 2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON). IEEE, 2024, pp. 174–179.
[75] D. Seychell, G. Hili, J. Attard, and K. Makantatis, "Ai as a tool for fair journalism: Case studies from malta," in 2024 IEEE Conference on Artificial Intelligence (CAI). IEEE, 2024, pp. 127–132.
[76] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities." Proceedings of the national academy of sciences, vol. 79, pp. 2554–2558, 1982.
[77] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, "Pixel-bert: Aligning image pixels with text by deep multi-modal transformers," arXiv preprint arXiv:2004.00849, 2020.
[78] M. AI, "Mistral nemo," 2024. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2019.2926266
[79] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023.
[80] D. Kahneman, "Evaluation by moments: Past and future," Choices, values, and frames, p. 710, 2000.
[81] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in International Conference on Machine Learning. PMLR, 2021, pp. 5583–5594.
[82] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2021.
[83] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," Advances in neural information processing systems, vol. 34, pp. 9694–9705, 2021.
[84] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, "Opt: Open pre-trained transformer language models," 2022.
[85] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan, "Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm," 2022.
[86] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015.
[87] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," Advances in neural information processing systems, vol. 32, 2019.
[88] S. Mullainathan and A. Shleifer, "Media bias," 2002.
[89] Y. Niu, H. Zhang, Z. Lu, and S.-F. Chang, "Variational context: Exploiting visual and textual context for grounding referring expressions," IEEE Transactions on Pattern Analysis and Machine Intelligence, p. 1–1, 2020. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2019.2926266
[90] S. Park, S. Kang, S. Chung, and J. Song, "Newscube: delivering multiple aspects of news to mitigate media bias," in Proceedings of the SIGCHI conference on human factors in computing systems, 2009, pp. 443–452.
[91] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, vol. 3, no. Jan, pp. 993–1022, 2003.
[92] A. Ziashahabi, M. A. Maddah-Ali, and A. Heydarnoori, "Bias-resistant social news aggregator based on blockchain," arXiv preprint arXiv:2010.10083, 2020.
[93] F. Hamborg, K. Heinser, A. Zhukova, K. Donnay, and B. Gipp, "Newsalyze: Effective communication of person-targeting biases in news articles," in 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE, 2021, pp. 130–139.
[94] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, 2014, pp. 55–60.
[95] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman, "Tracking and summarizing news on a daily basis with columbia's newsblaster," in Proceedings of the human language technology conference. Morgan Kaufmann San Francisco, 2002, pp. 280–285.
[96] Y. Ko, S. Ryu, S. Han, Y. Jeon, J. Kim, S. Park, K. Han, H. Tong, and S.-W. Kim, "Khan: knowledge-aware hierarchical attention networks for accurate political stance prediction," in Proceedings of the ACM Web Conference 2023, 2023, pp. 1572–1583.
[97] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[98] E. T. Jaynes, "Information theory and statistical mechanics," Physical review, vol. 106, no. 4, p. 620, 1957.
[99] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "Hellaswag: Can a machine really finish your sentence?" arXiv preprint arXiv:1905.07830, 2019.
[100] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring mathematical problem solving with the math dataset," arXiv preprint arXiv:2103.03874, 2021.
[101] M. Nagel, R. A. Amjad, M. Van Baalen, C. Louizos, and T. Blankevoort, "Up or down? adaptive rounding for post-training quantization," in International Conference on Machine Learning. PMLR, 2020, pp. 7197–7206.
[102] R. Puglisi and J. M. Snyder Jr, "Empirical studies of media bias," in Handbook of media economics. Elsevier, 2015, vol. 1, pp. 647–667.
[103] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, 1989.
[104] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2016.
[105] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, vol. 61, p. 85–117, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.neunet.2014.09.003
[106] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Worts-

man, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "Laion-5b: An open large-scale dataset for training next generation image-text models," 2022.
[107] N. Schwarz, H. Bless, F. Strack, G. Klumpp, H. Rittenauer-Schatka, and A. Simons, "Ease of retrieval as information: Another look at the availability heuristic." Journal of Personality and Social psychology, vol. 61, no. 2, p. 195, 1991.
[108] C. E. Shannon and W. Weaver, The mathematical theory of communication, by CE Shannon (and recent contributions to the mathematical theory of communication), W. Weaver. University of illinois Press, 1949.
[109] K. G. Shaver, "Defensive attribution: Effects of severity and relevance on the responsibility assigned for an accident." Journal of personality and social psychology, vol. 14, no. 2, p. 101, 1970.
[110] F. M. Simon, "Artificial intelligence in the news: How ai retools, rationalizes, and reshapes journalism and the public arena," 2024.
[111] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," arXiv preprint arXiv:1908.07490, 2019.
[112] W. L. Taylor, ""cloze procedure": A new tool for measuring readability," Journalism quarterly, vol. 30, no. 4, pp. 415–433, 1953.
[113] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan et al., "The llama 3 herd of models," arXiv preprint arXiv:2407.21783, 2024.
[114] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023.
[115] M. Tsimpoukelli, J. Menick, S. Cabi, S. M. A. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," 2021.
[116] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, . Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.
[117] E. Walster, "Assignment of responsibility for an accident." Journal of personality and social psychology, vol. 3, no. 1, p. 73, 1966.
[118] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," 2022.
[119] D. M. White, "The "gate keeper": A case study in the selection of news," Journalism quarterly, vol. 27, no. 4, pp. 383–390, 1950.
[120] Q. Xia, H. Huang, N. Duan, D. Zhang, L. Ji, Z. Sui, E. Cui, T. Bharti, X. Liu, and M. Zhou, "Xgpt: Cross-modal generative pre-training for image captioning," 2020.
[121] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," Advances in neural information processing systems, vol. 32, 2019.
[122] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," 2020.