Assignment1-PartA

Poojal Katiyar – 220770 & Nipun Nohria – 220717

26th January 2025

Introduction

Many believe search engines provide unbiased and neutral results due to the assumption that algorithms are impartial. However, this is not entirely true, as search engines often produce biased results influenced by societal prejudices, political agendas, or specific economic interests. For example, former U.S. President Donald Trump accused Google of bias in showing predominantly negative news about him. Google refuted this claim, emphasizing that their results prioritize relevance without political or ideological bias.

Bias becomes particularly significant in controversial topics or those with historical inequalities. Search engine personalization, such as Google's "filter bubble," further narrows the diversity of perspectives available to users, sometimes influencing opinions subtly through "nudging."

Gender biases are another form of bias, reflecting societal stereotypes and reinforcing inequality. Search results for professions or fields, such as cricket, often disproportionately highlight male-dominated narratives while underrepresenting women. Studying these biases sheds light on how digital tools unintentionally perpetuate societal inequalities.

1: Gender Biases in Male dominated Sports

This analysis focuses on identifying gender biases in search engine results related to cricket. According to the United Nations Development Program, significant inequalities persist between genders, particularly in jobs, income, political participation, and the distribution of unpaid domestic work. These societal biases are often mirrored in search results.

The methodology involved the following steps:

1. **Defining Keywords**: Gender-specific keywords commonly associated with men and women in cricket were identified.
2. **Scraping Search Results**: Using Python libraries such as `requests`, `BeautifulSoup`, and `SerpAPI`, search results were scraped from DuckDuckGo, Bing, Yahoo, and Google.
3. **Analyzing Snippets**: The text snippets were processed to count the occurrences of men-related and women-related keywords.
4. **Visualization**: A word cloud was created to highlight frequently used words, and a bar chart was generated to compare gender keyword counts across search engines.

The differences in keyword occurrences across search engines are visualized in the bar graph below:

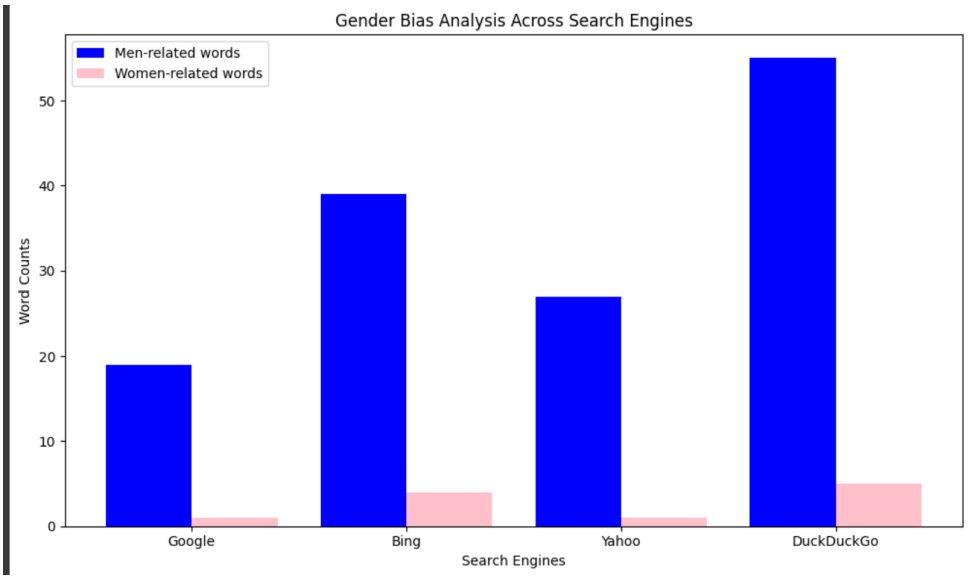

Figure 1: Comparison of gender-specific keyword counts across search engines.

Inference

- The analysis was conducted on the first page of search results for *cricket* in India. Google and Yahoo displayed the least number of women cricket-related terms.
- DuckDuckGo displayed the highest number of women cricket-related words among the analyzed search engines, but the absolute count was still very low.DuckDuckGo (DDG) is said to not gather personal data from its users for its page ranks so this could be the possible reason.
- Repeated analysis consistently demonstrated a disproportionate ratio of men-related to women-related keywords across search engines.
- The observed disparity highlights broader societal biases, especially in male-dominated fields like cricket. These biases are reflected in search engine algorithms and impact public perception.

Causes and Improvement that can be done:

The findings reveal a significant underrepresentation of women in cricket-related search engine results, with societal gender biases mirrored in the digital sphere. To address this issue, search engines must prioritize diversity and balance in their algorithms, ensuring fair representation of all genders. This is the reason that when people don't see so many women cricketers on search, they don't watch women's cricket. Similarly, we can say that Tennis is a female dominated sport therefore when we search for it, it mostly shows female players. Similarly, for professions like engineering, Scientist, Technician, we get male more due to societal biases and number.

2: Use of obscene language especially for girls

This analysis explores the disparities in search result content by generating word clouds for specific search phrases: "Russian girls", "Black girls" and "Russian Boys". The objective is to identify patterns, stereotypes, and differences in the language used to describe these groups, revealing underlying societal and algorithmic biases. The methodology was simple, I collected the data across various search engines. Common stop words (e.g., "the", "and", "to") were removed to focus on meaningful content. The processed text generated word clouds, visually representing frequently occurring words associated with each search phrase. section*Inference

- The analysis showed that the most prominent terms include "dating," "marriage," "brides," "photos," and "free." especially for Russian women. This suggests an overwhelming focus on romantic relationships, objectification, and commodification.

- Terms such as "stunning" and "beautiful" emphasize physical appearance, perpetuating stereotypes that portray women from this group as objects of beauty and marriage prospects. Apart from this link related to viewing the images are also coming like hot images, especially for Russian

Figure 2: Word Cloud for Russian Girls.

- Through the analysis I also saw that for Black girls terms related to marriage are not used rather terms like "inspiring", "power", "great", "knowledge", "beautiful", "unique", and "intelligent" are used which explains the difference of our thoughts for both searches

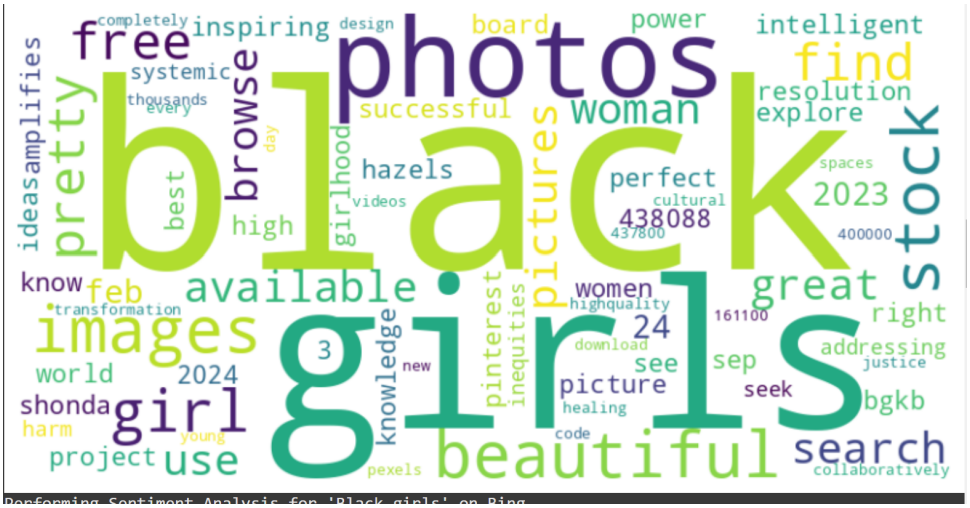

Figure 3: Word Cloud for Black Girls

- Prominent terms include empowering words such as "rich", "modern", "cultural", "unique", and "style" for Russian boys in contrast to Russian girls.

Figure 4: Word Cloud for Russian Boys

Causes and Improvement that can be done:

The main reason I think is that Search engines prioritize content based on engagement metrics, which may amplify existing stereotypes. Biased or incomplete data sources used to train algorithms can reinforce societal biases. Instead, we should use expert data or show only verified sources at our top rank and not be biased on the basis of the stereotypes created on the net or society. These searches will enhance the stereotypes for those newbies and will impact negatively and create a negative environment.

3: Pornographic and Sensational Terms when searched Rape Videos

The analysis indicates a bias in search results where queries related to "rape videos" seem to emphasize pornographic content rather than focusing on informative, educational, or legal aspects. This highlights an algorithmic and societal bias associating serious topics with exploitative and inappropriate content. The cause for this is the access to the Internet which results in people accessing pornography and for them, rape videos are also of that kind which can lead to a poor mindset. To analyse this I plotted a word cloud of the snippets of the data available on the first page and performed sentimental analysis to look at the sentiments associated. section*Inference

- The Average Sentiment Score: -0.30 (Negative). This tells us about the negative context displayed on the search.The word cloud contains words like "porn," "graphic," "unconscious," "brutal," "revenge," "footage," which suggest that search engines may prioritize content with sensational or exploitative angles rather than focusing on educational or advocacy-related materials.
- Expected words like "awareness," "laws," "support," "prevention," "reporting," "resources," and other victim-assistance-related terms are absent or less prominent.

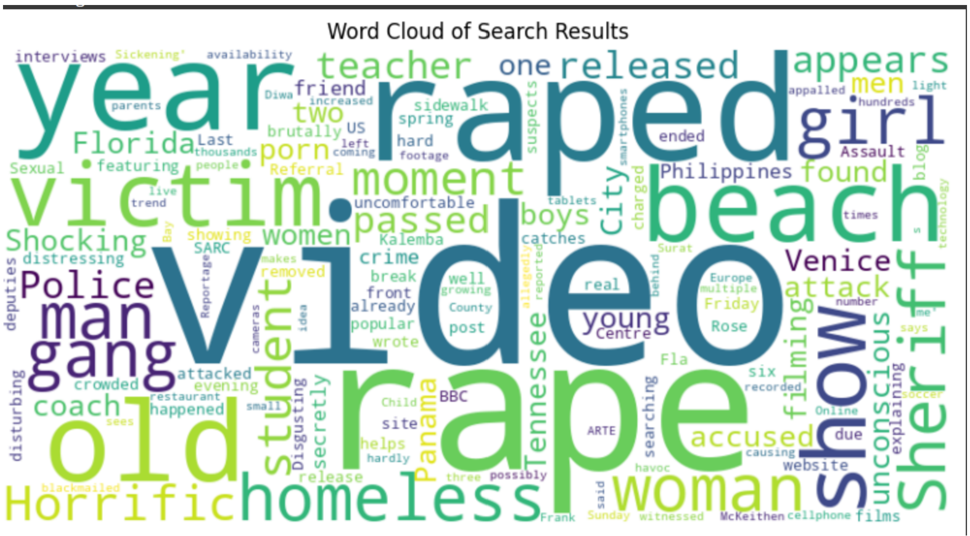

Figure 5: Word Cloud for rapevideos

Causes and Improvement that can be done:

The main cause for this is the rise of pornography. For E.g. Recent rape that happened in Kolkata where people were searching for the video on pornographic sites tells the mentality of the society. The engines also algorithmically put the content first that is popular.Search engines should prioritize informative and awareness-driven content by tweaking ranking algorithms to surface more credible and helpful resources.

I performed some analysis on YouTube Videos search results.

4: Searching celebrity leaked photos on Youtube

The analysis indicates how the titles of YouTube videos can be categorized as misogynistic, obscene, and breaching the privacy of an individual. I scraped the data (basically the titles of YouTube videos) using Selenium and Beautiful Soup. Then I performed a Sentimental Analysis, drew the word cloud, found the most frequent word used, and analyzed the results.

Inference

- A significant number of YouTube video titles related to leaked celebrity photos use sensationalist language, such as “Top 5 Leaked Celebrity Nude Photos” or “Hackers leak nude photos of 100+ celebs.” These titles often exaggerate the content and may mislead viewers, violating ethical norms of truthfulness and respect.
- Many video titles focus on the violation of privacy by mentioning terms like ”private,” ”leaked,” and ”nude,” which indicates a breach of individuals’ privacy for sensational gain. These types of titles tend to objectify and exploit the personal lives of celebrities, contributing to an unhealthy culture around privacy.
- Many titles contain elements of clickbait, using terms like ”shocking,” ”unbelievable,” or ”revealed,” designed to attract views through extreme curiosity.
- When compared across the words used on the basis of frequency, nude and naked were the most common words.
- There are also video titles saying the wardrobe malfunction of an actress which is a breach of privacy.

Figure 6: Word Cloud for ”Celebrity leaked photos”

Causes and Improvement that can be done:

The main cause is to attract more views by Clickbaiting the users. Content creators or platforms use clickbait to increase views, and by extension, ad revenue. More views mean more income from advertisements, so there's a financial incentive to create provocative and controversial content, even at the expense of privacy and dignity. Governments should implement and enforce laws against the sharing of private content without consent. The laws must be updated regularly to tackle emerging methods of privacy infringement. Social media platforms, video-sharing sites, and other online platforms need to be held accountable for the content shared on their networks. Implementing stronger moderation practices and removing harmful or violating content swiftly could reduce exposure. News outlets and websites should commit to ethical reporting standards. Instead of sensationalizing content (such as leaked photos), media should focus on the broader issues of privacy, consent, and ethical boundaries.

5: Searches regarding LGBT rights

On the search queries such as "LGBT rights", "LGBT rights in India", there was some variation in the responses given by the search engines. The sentiment analysis for DuckDuckGO and yahoo returned positive results, but for google mixed results were seen, indicating a possibly more critical opinion. On doing a hate speech analysis of the snippets, no signs of hate speech were seen for either of the search engines. Further on investigating the snippets and word clouds for the search results from all the engines, there seems to be a difference in how the search results talk about the issue

- Google- The search results for google were less opinionated and contained more factual content about the history of LGBT rights in India.

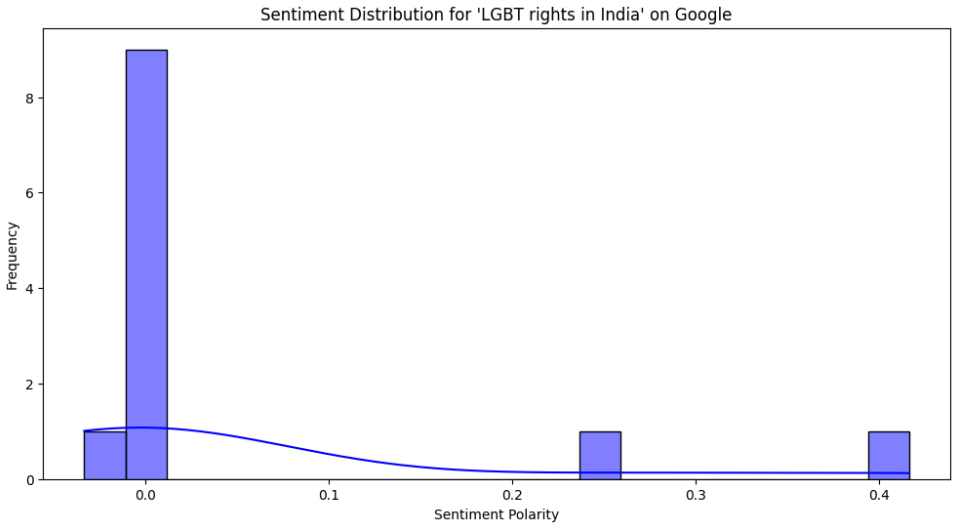

Figure 7:

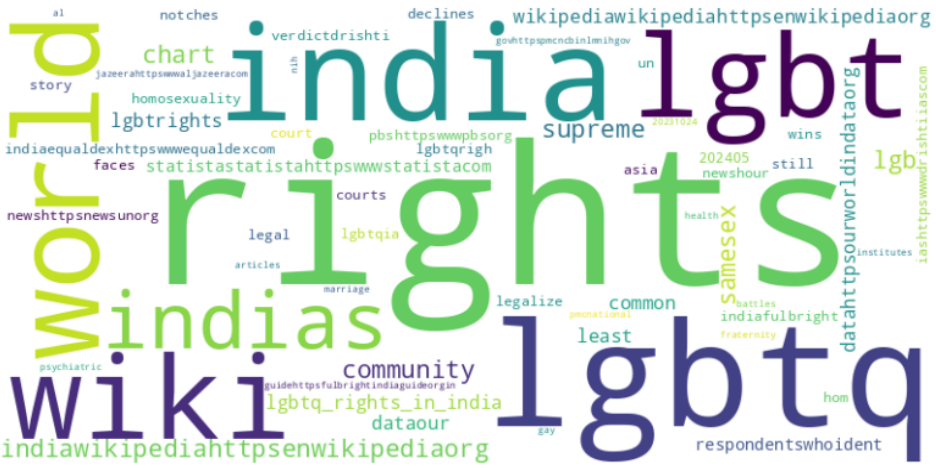

Figure 8:

- yahoo- The search results for yahoo focus on the fact that India has made significant progress in these issues however there is still a greater need for awareness. So yahoo pushed articles more in support of these people.

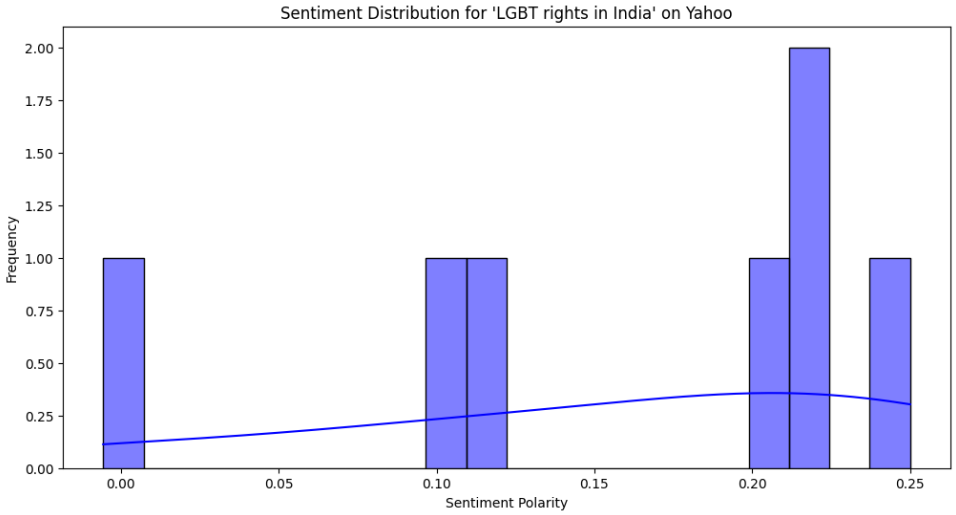

Figure 9:

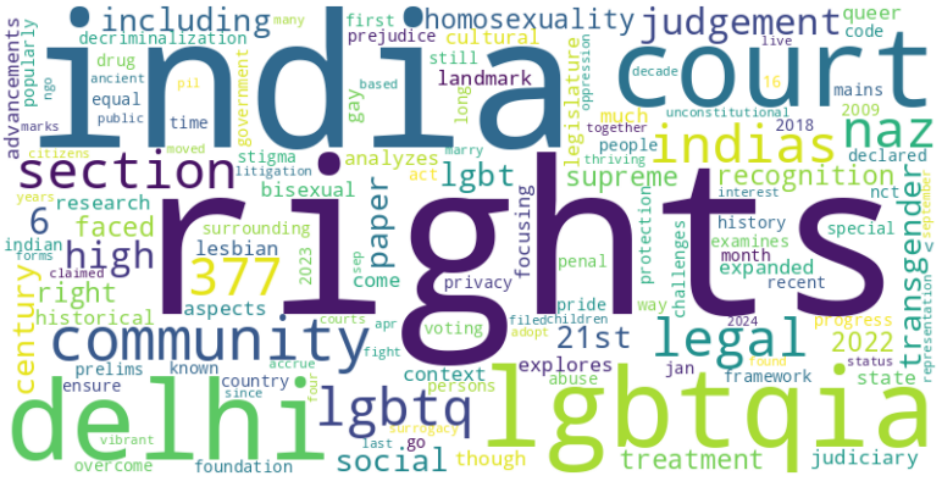

Figure 10:

Overall the search engines might present a different image regarding the situation of LGBT community in India, however none of the search engines can be said to promote hate speech. And in general they promote articles favouring rights for the LGBT community.

6: Some Other Analysis Done:

- When searching for "best smartphones 2025" across different search engines, I observed varying results based on the ranking algorithms used by each engine. For instance, searches on Yahoo and Bing frequently highlighted iPhones, while Galaxy phones appeared more prominently on DuckDuckGo. This variation in results can be attributed to the different ranking strategies and algorithms employed by these search engines, as well as their unique methods of handling user data and personalization.

- There are noticeable biases when changing the location of a search, such as when comparing searches from the UK, US, or India. For example, searching for the term "Women" can yield slightly different results depending on the location. These variations in search results are influenced by factors such as regional preferences, cultural context, and the search engine's localized algorithms.

- Searching for celebrities or public figures, especially women, may often lead to sensationalized or sexualized content, such as "leaked photos" or "wardrobe malfunctions." For example, searching "celebrity women" might show more results about their appearance or personal lives rather than professional accomplishment

Limitations:

The search query I tested may not capture the full spectrum of bias present in search results. For example, searching only for specific terms (e.g., "women leaders") may not account for the more subtle biases present in other types of searches. Titles are often short,

informal, or clickbait-oriented, making it challenging for sentiment analysis algorithms to correctly assess tone. Search results vary greatly depending on the user's location, even with the same search term. This makes it harder to draw generalized conclusions from the data, as search algorithms often prioritize local content, language, and regional preferences.

- **Colab Link for the Code:** Google Colab Link