Unveiling Gender Disparities in STEM: A Comprehensive Analysis of Encouragement and Interest among Women

Homayra Tabassum
Veronica Mata

Abstract: The underrepresentation of women in STEM fields is a global concern that has sparked debate regarding the need for encouragement versus inherent interest. This project seeks to investigate this issue by analyzing data collected from prominent social media platforms such as Reddit. By identifying meaningful patterns, the study aims to shed light on whether women are discouraged from pursuing STEM careers or if a genuine lack of interest exists. Employing data mining techniques, we analyze social media interactions to understand the dynamics influencing women's participation in STEM. Our findings offer insights into the factors driving this disparity in the STEM community.

Introduction : The persistent underrepresentation of women in STEM is a global issue. Despite efforts to promote gender diversity, a significant gap remains. This project investigates whether this is due to inherent disinterest or societal discouragement. Previous research indicates a complex interplay of factors influencing women's participation in STEM. For instance, Whitcomb highlights that women who leave STEM majors often have higher grades than their male counterparts, suggesting factors beyond academic performance at play [1]. He emphasizes the potential of social media in connecting women with STEM role models [2].

In this project we analyzed data collected from Reddit related to women working or Studying in STEM to identify patterns related to women's lack of participation in the field. By doing so we seek to understand if women are primarily discouraged from pursuing STEM careers or if there is a genuine lack of interest.

First, we collected posts related to our research topic using Reddit API along with comments on the posts with comment scores, followed by data cleaning and preprocessing. After data preprocessing we analyzed frequent patterns to understand sentiments that are commonly discussed among the users in the platform related to our topic.

In the next step we performed clustering on all the comments to see if there were different kinds of sentiments that people were talking about in the posts and the comments. We found four significant clusters each representing separate topics that people talked about in the platform related to women in STEM. In the next step of analysis we used the GPT-3 API as a classifier tool to analyze the sentiment in Reddit comments pertaining to women in the STEM fields into four classification categories : Encouraging women in STEM, Disencouraging women in STEM, Lack of Interest in women in STEM and Interest in women in STEM.

Finally, we calculated the mean of comment score for each category and compared the means for “Discouraging Women in STEM” and “Lack of Interest in women in STEM”, these two categories and used a non parametric test to see if the difference is significant or not. We find that the mean score for “Discouraging Women in STEM” is higher than “Lack of Interest in women in STEM” and the difference is statistically significant. Therefore we can conclude that the majority of the discussion points towards women being discouraged in STEM, rather than lacking interest in STEM.

Dataset: We collected data related to women in stem topics using the reddit API. Initially we collected 450 posts with 22076 comments with their comment scores. There were 11560 unique authors of these posts and comments.

To clean the dataset we removed comments that had been removed or deleted and eliminated duplicate comments. We also deleted comments with characters less than 50 because they are fair to add significance to the dataset. Finally we got rid of stop words and unusual chapters. Then we performed Lemmatization on the comments and tokenized the comments for analysis. After cleaning the data we were left with 450 posts and 14268 comments with 6538 unique authors who posted or commented.

Frequent Pattern Analysis using LDA:

Our application of Latent Dirichlet Allocation (LDA) to the dataset revealed distinct topics reflecting various aspects of the discourse surrounding women in STEM. The LDA model identified five topics, each with a unique set of keywords indicative of the underlying patterns:

- **Topic 0: Socio-Political Identities and Perspectives** — This topic encapsulates conversations around social identities and political affiliations, with keywords such as 'gay', 'Asian', 'feminist', and 'conservative', suggesting discussions on the intersectionality of gender and other identities within STEM fields.
- **Topic 1: Gender Dynamics and Perceptions** — Central to this theme are keywords like 'woman', 'men', and 'gender', pointing towards dialogues on gender perceptions and the dynamics between different genders in STEM environments.
- **Topic 2: Professional Environment and Education** — Keywords such as 'work', 'job', and 'engineer' highlight discussions focused on the professional and educational settings of STEM, addressing the occupational landscape.
- **Topic 3: General Opinions and Experiences** — Frequent terms like 'like', 'say', and 'think' reflect general opinions and anecdotal experiences shared by individuals, possibly relating to personal journeys in STEM.
- **Topic 4: Personal and Recreational Aspects** — This topic diverges towards more personal and leisurely aspects, with words like 'game', 'home', and 'time', indicating a blend of professional and personal life discussions.

Each topic represents a spectrum of discussions that contribute to our understanding of the multifaceted issue of gender disparities in STEM. The prevalence of certain keywords within topics provides an interpretative lens to explore the narrative structures and sentiment in social media discourse related to our research objectives.

Clustering Analysis Using OPTICS: To explore the comments on women in STEM, we employed the OPTICS (Ordering Points To Identify the Clustering Structure) algorithm. This unsupervised machine learning technique is adept at discovering clusters of varying densities, which is suitable for finding differences in text based datasets. Our analysis revealed four distinct clusters.

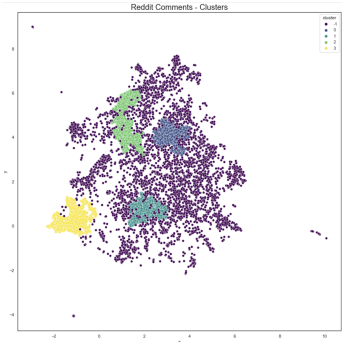

Image : clusters of the comments

- Clusters 0 and 2 exhibited positive sentiments towards women in STEM, suggesting a supportive dialogue within these groups.
- Clusters 1 and 3 reflected negative views, indicating critical or discouraging perspectives.

However, the initial clustering did not distinctly categorize comments regarding encouragement or interest in STEM fields. To gain deeper insights into these specific aspects, we proceeded with a classification analysis of the comments.

Classification : We employed the GPT-3 API as a classifier tool to analyze the sentiment in the Reddit comments we had collected. The classification categories are : “discouraging_women” , “encouraging_women” , “women_lack_interest” and ”women_interest_in _stem”. A manual review of sample 100 comments was conducted to test the accuracy of the GPT-3 classifier. This initial test showed promising results with an accuracy rate of approximately 80%.

In order to understand the users’ opinion on the argument we calculated the mean if score of comments in each class. The score of each comment represents how many users have given the comment an upvote, which indicates the user is agreeing with the comment.

Category	#Comments	Mean score
discouraging_women	5022	**16.05**
encouraging_women	2152	19.79

women_lack_interest	1057	**13.66**
women_interst_in_stem	556	4.99
Unclassified	1057	7.65

Table : Mean score of comment for each category.

As we can see that classes “discouraging_women” have higher mean scores than “women_lack_interest”.

On the next step, we calculated whether the difference of mean score is statistically significant or not. Since both the groups have a distribution that is not normal we opted for a non-parametric test, Mann-Whitney U test, and set the alpha value to 0.05 with a null hypothesis, “There is no significant difference between the mean scores of the two groups: 'discouraging_women' and 'women_lack_interest”. Since the value of p is 0.00136, which is less than alpha, we can reject the null hypothesis and conclude that there is statistically significant evidence to suggest a difference in the mean scores between the two groups 'discouraging_women' and 'women_lack_interest'.

Conclusion : Our analysis revealed a statistically significant difference of mean scores of comments within the 'Discouraging Women' category compared to 'Women Lack Interest'. This finding suggests that the majority of the discussion in the comments points towards the factors of discouragement, rather than a lack of interest, as a contributor to the gender disparity in STEM fields. These results highlight the need to address the problem of discouragement women face in STEM. For future work, we propose refining the application of GPT-3.5 turbo to enhance the precision of sentiment analysis and topic modeling to better understand the nuances within the discourse.

References:

[1] Whitcomb, K. M., & Singh, C. (2020). Gender inequities throughout STEM: Women with higher grades drop STEM majors while men persist. Department of Physics and Astronomy,

University of Pittsburgh, Pittsburgh, PA, 15260. Published on April 2, 2020. arXiv.org. https://arxiv.org/abs/2202.02438

[2] Kelly He, Lee Murphy, and Jiebo Luo, "Using Social Media to Promote STEM Education: Matching College Students with Role Models", *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery (ECML/PKDD),* Riva del Garda, Italy, September 2016. http://arxiv.org/abs/1607.00405