DS-UA 202: Homework 2

Daphne Ozkan

April 9, 2025

Problem 1

In her lecture on AI ethics and hate speech, Maha Jouini explores how artificial intelligence technologies, particularly those embedded in social media platforms, have contributed to the rise and spread of hate speech in the MENA region. She emphasizes that while platforms like Facebook and Twitter claim to invest in moderation efforts, their algorithms often amplify harmful content due to engagement-driven designs. This amplification is especially concerning in contexts where vulnerable groups are frequently targeted, and where legal frameworks for regulating online speech are either weak or nonexistent.

A key responsible AI concern she raises is the linguistic and cultural bias in AI moderation systems. Most content moderation tools are trained primarily on English-language data and lack the nuance to effectively process hate speech written in Arabic or other regional languages. As a result, harmful rhetoric in these languages often goes undetected. Furthermore, there is no standardized legal definition of hate speech in many MENA countries, which makes it more difficult to implement effective policies or hold users accountable.

The impact of these systems falls most heavily on already marginalized groups. Women, LGBTQ+ individuals, ethnic and religious minorities, and civil society leaders are frequently the targets of hate speech, often with real-world consequences. Jouini shares her own experiences as an activist who has been threatened online, highlighting the personal risks involved. Additionally, communities across the region experience social and political instability, which can be worsened by the unchecked spread of extremist rhetoric online.

Transparency and interpretability are central to understanding the risks posed by these systems. Jouini critiques the fact that most AI moderation tools operate as "black boxes": users do not know how decisions are made, nor are moderation standards communicated in accessible or localized ways. This lack of clarity undermines trust and allows biases to persist without scrutiny or correction.

Jouini also discusses the limited incentives vendors have to improve transparency. Social media companies often prioritize engagement metrics and user

growth over ethical concerns. While platforms have publicized their investments in safety, Jouini points out that enforcement tends to focus on Western contexts, leaving regions like MENA significantly underserved. Without regulatory pressure or strong public accountability, there is little motivation for companies to adapt their systems to regional needs.

She argues for a more inclusive approach to AI governance, one that involves communities in the MENA region in both the development and implementation of ethical frameworks. This includes providing AI ethics resources in Arabic, training local practitioners, and ensuring that the people most affected by hate speech are given a voice in shaping solutions. Addressing hate speech through AI requires more than technical fixes; it also demands cultural awareness, transparency, and meaningful stakeholder engagement.

Overall, the lecture illustrates how hate speech in the MENA region is not only a social issue but also a deeply technical one. Without greater transparency and local involvement, AI systems will continue to reflect and reinforce the inequalities they ought to challenge.

Problem 2

(a)

Applicants from groups with above-average experience, like female and non-binary white (mean = 7.40) and female and non-binary other race (mean = 7.91), may be disadvantaged by this imputation. By replacing missing experience values with a value lower than their group average, their qualifications may be underestimated, leading to lower rankings than they would otherwise receive.

(b)

A better method would be to impute missing experience values using the group-specific mean instead of the overall mean. For example, for a female and non-binary applicant of "other" race, replace missing experience values with 7.91, which better reflects the typical experience of their group. This approach preserves within-group variation and reduces the risk of disadvantaging applicants whose group has systematically higher experience.

(c)

Using the overall mean introduces technical bias by disproportionately lowering the imputed experience for groups with higher average experience. This reflects a form of pre-existing bias, as the imputation reinforces structural inequalities already present in the data, such as certain groups needing to work harder to gain experience. It also creates emergent bias because the model's output, which ranks applicants, becomes skewed against those groups. This

compounds disadvantage during hiring and may result in lower chances of selection for highly qualified individuals simply because their group averages were not properly accounted for.

Problem 3

3: Explaining Text Classification with SHAP

(a)

I trained an `SGDClassifier` on a text classification task to distinguish between "atheism" and "Christianity" using TF-IDF features. The initial model achieved an accuracy of 93.17, with 49 misclassified examples. SHAP was used to analyze and explain the predictions.

(b)

Below are SHAP explanations for five test documents. Three were correctly classified and two were misclassified.

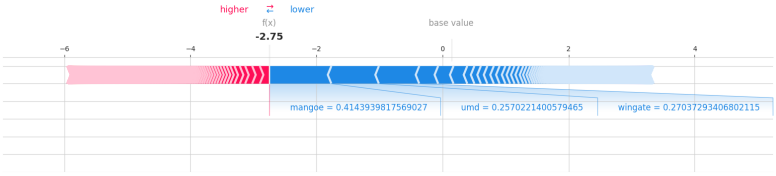


Figure 1: Misclassified example 1

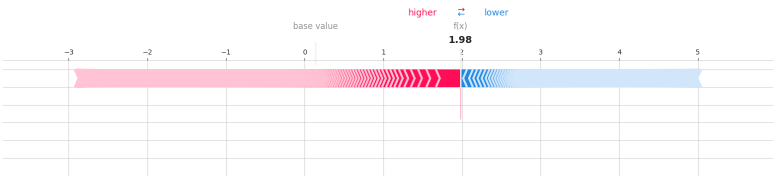


Figure 2: Misclassified example 2

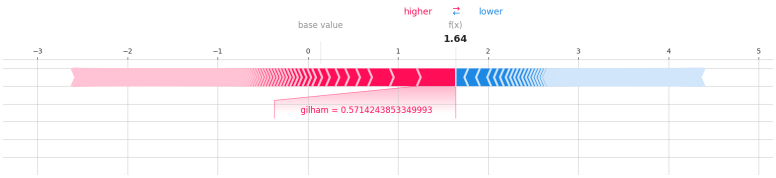


Figure 3: Correctly classified example 1

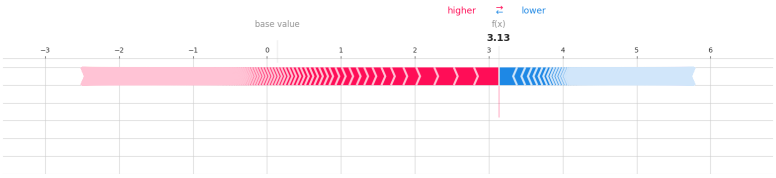


Figure 4: Correctly classified example 2

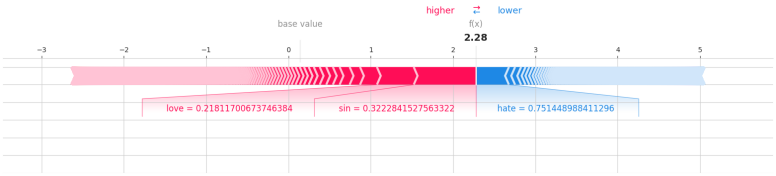
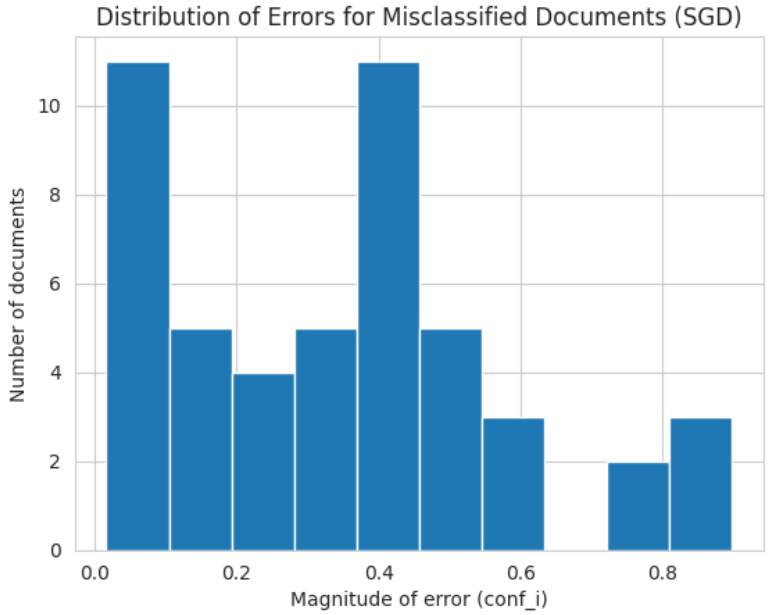
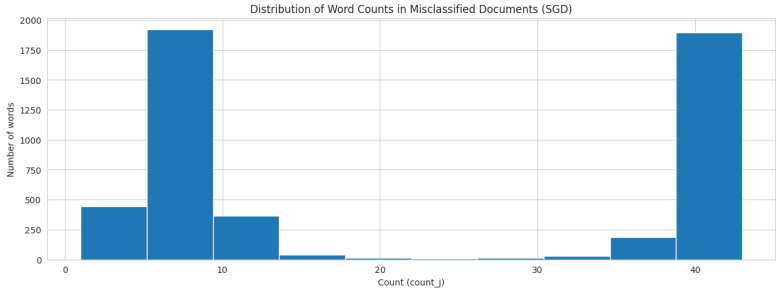


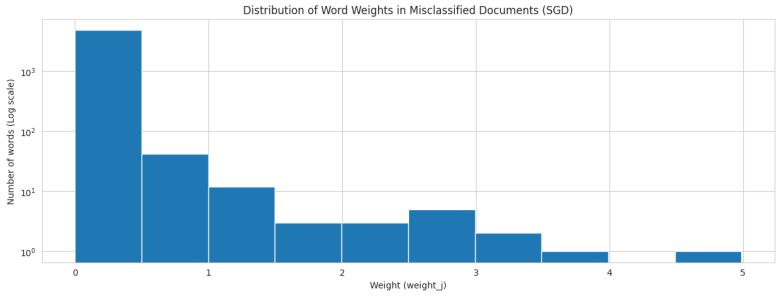
Figure 5: Correctly classified example 3

(c)

I aggregated SHAP values over all misclassified documents and identified the words most frequently contributing to error. These were mostly structural or non-informative words like `posting`, `host`, and `article`.


Distribution of Errors for Misclassified Documents (SGD)
Number of documents
Magnitude of error (conf_i)
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8


Distribution of Word Counts in Misclassified Documents (SGD)
Number of words
Count (count_j)
0
250
500
750
1000
1250
1500
1750
2000
0
10
20
30
40


Distribution of Word Weights in Misclassified Documents (SGD)
Number of words (Log scale)
Weight (weight_j)
10^0
10^1
10^2
10^3
0
1
2
3
4
5

(d)

I implemented a feature selection strategy that filtered out words contributing to 10 or more misclassifications. Retraining the model with these features removed led to an improved accuracy of 94.4%, with only 40 misclassified documents.

To demonstrate the impact of this filtering, I selected a document that was originally misclassified and became correctly classified afterward. The following plots compare the SHAP explanation for this document before and after feature selection.

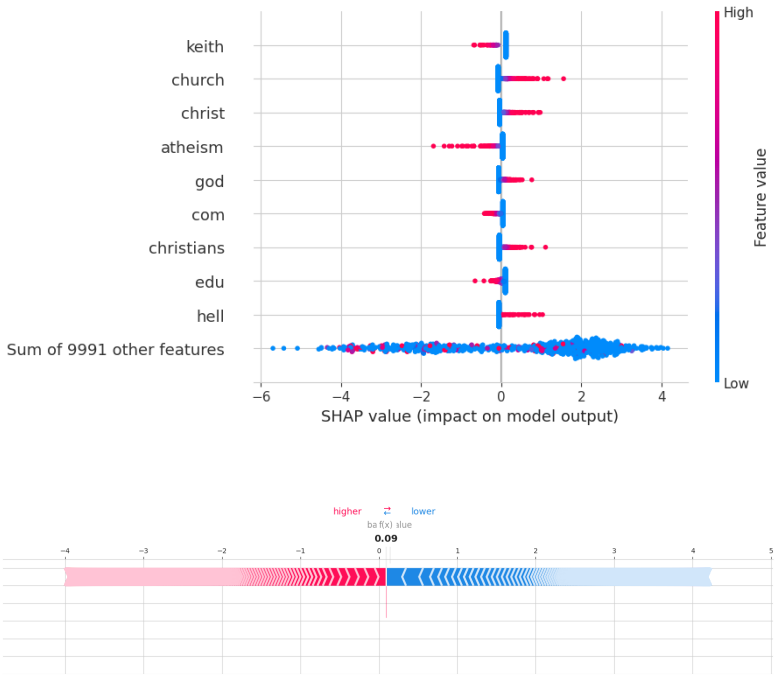


Figure 6: SHAP explanation *before* feature selection

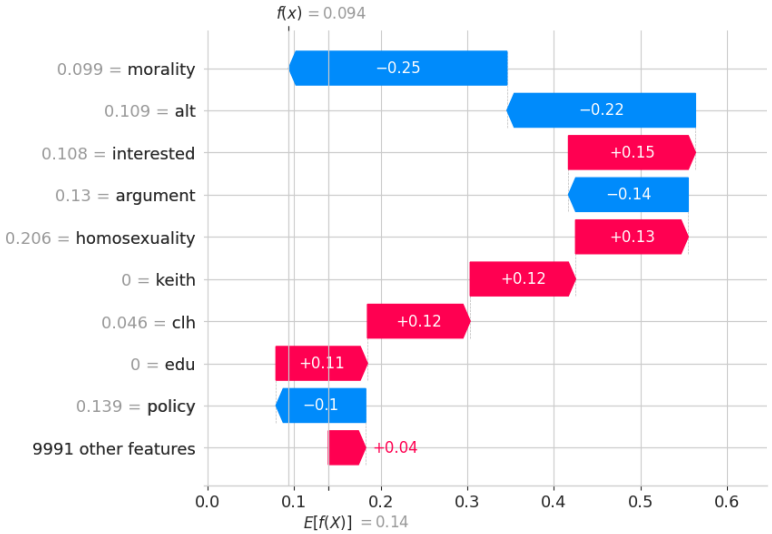


Figure 7: SHAP waterfall plot *before* feature selection

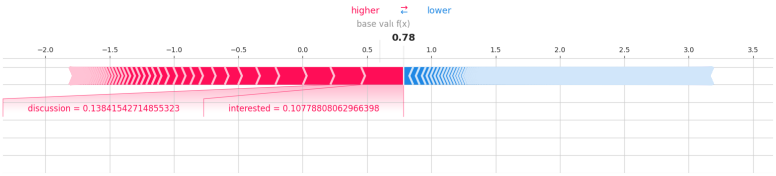


Figure 8: SHAP explanation *after* feature selection

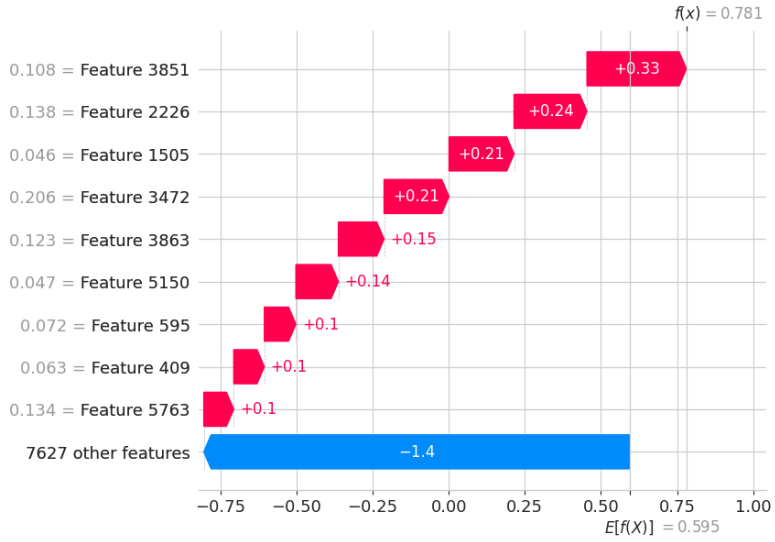


Figure 9: SHAP waterfall plot *after* feature selection

As shown above, the explanation before filtering relied on low-importance or misleading words, while the corrected version emphasized more semantically meaningful features. The SHAP value magnitude also increased, indicating higher classifier confidence.

Problem 4

a

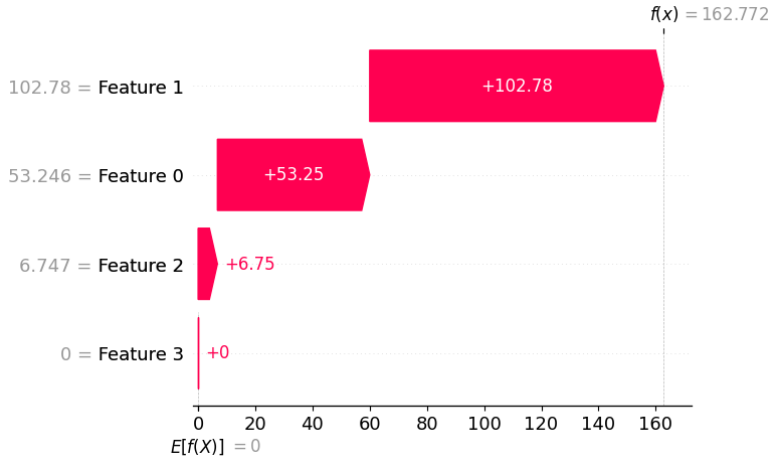


Figure 10: Explanation for the 100th ranked individual for score_function_SCHL

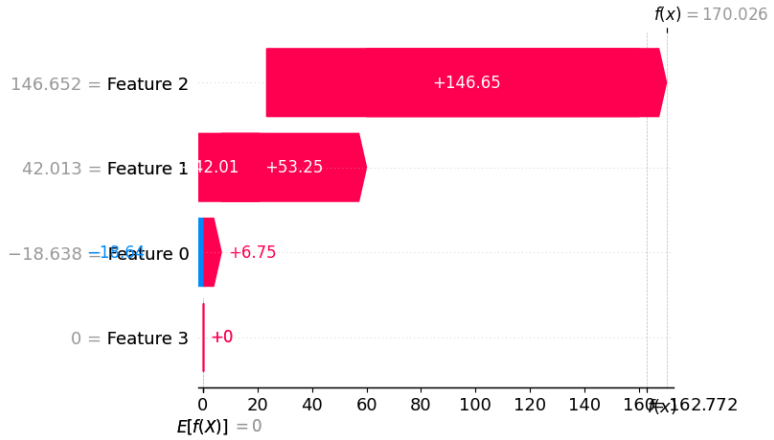


Figure 11: Exaplanation for the same individual for score_function_WKHP

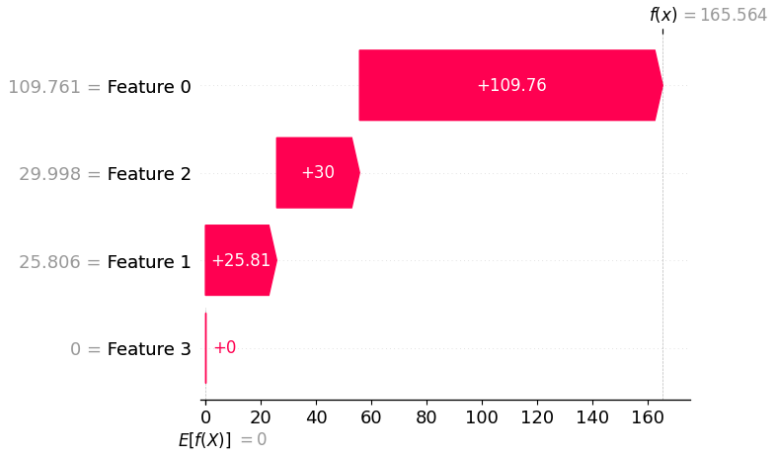


Figure 12: Explanation for the 100th ranked individual for score_function_AGEP