"Compassionately": Increasing Plurality Awareness through Community-powered AI

Hala Sheta
Department of Computer Science
University of Waterloo
Waterloo, Canada
hsheta@uwaterloo.ca

Mohamed Ahmed
Department of Computer Science
University of Toronto
Toronto, Canada
mohamed@cs.toronto.edu

Syed Ishtiaque Ahmed
Department of Computer Science
University of Toronto
Toronto, Canada
ishtiaque@cs.toronto.edu

Abstract

The prevalence of Islamophobia has resulted in the continual discrimination of Muslims in North America, where polls have found them to be viewed in a negative light, specifically as a rigid, violent and monolithic community. Furthermore, within the Muslim community, many are ignorant and intolerant of differing perspectives to their own beliefs, viewing Islamic jurisprudence as a one-sided, static truth. However, recent work in this domain rightfully reframe it as a *human-centered field*, and work to amplify minority voices (e.g. women) that are often undermined in textual interpretations. So, our research is centered around utilizing community-powered AI to increase plurality awareness and understanding, both within and outside of the Muslim community. The end goal is to develop a semi-automated hate-speech detection system that educates its users about the multiplicity of perspectives surrounding a topic of interest. This paper will serve as foundational work to this overarching goal by gauging the current climate surrounding the discussion of Islamic perspectives both through a qualitative analysis of online discussions on Reddit and conducting a controlled user study.

1 Introduction

One of the most pervasive issues facing North American Muslims is the prevalence of Islamophobia, with over half of American Muslims having experienced religious discrimination, portraying them as rigid, violent, and monolithic [10]. This is in stark contrast to the plurality of religious and social perspectives found among Muslim communities around the world. Additionally, many Muslims themselves are unaware or intolerant of differing perspectives to what they view as the norm [3]. For example, a majority of North American Muslims believe in the separation of religion and statecraft [11], support LGBTQ+ rights [4], and are more committed to nonviolence than those polled from other religions [5]. This discrepancy between how rigid and monolithic Muslims are viewed as and the diversity in beliefs among North American Muslims is more than a matter of misunderstanding and has led to demonstrable harm to Muslims worldwide [13].

Furthermore, within the realm of Islamic jurisprudence [*fiqh*], Muslims generally view "juristic propositions" as an unchangeable truth comparable to religious scripture (Qur'an), without critical evaluation and contextualization of these propositions that are usually not reflective of issues faced by

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

contemporary Muslims (e.g. as a byproduct of technological advancement) [7]. Another reigning issue in the field of *fiqh* is source reliability and the differing definitions of "proper religious authority". This can range from requiring the scholar to have a high command of Qura'nic Arabic and its etymology [1] to a philosophy-centered understanding that demands utmost due diligence to extract knowledge from principles and their inferences [2].

To combat this and to bring light to the wide range of perspectives found within the Muslim community, we propose "Compassionately", a community-powered AI system meant to elicit and visualize the plurality of religious sentiments among Muslims. In order to reflect the multiplicity of perspectives among Muslims, this tool will be developed with the direct involvement of a diverse range of Muslim communities. To accomplish this, we plan on advancing religious text based Natural Language Processing (NLP) techniques by developing a semi-automated hate-speech detection system that captures and visualizes existing translations of the Qur'an and elicits interpretations of key Qura'nic verses through community-based workshops.

This work builds on the work of Hutchinson [6], who critically analyzed the use of religious text in NLP, pushing for the intentional use of these texts with careful ethical and cultural considerations. He proposed several recommendations for researchers using religious texts in NLP contexts, such as the importance of culturally situating NLP work and attending to the concerns of marginalized cultures. Additionally, this work relies on understanding *fiqh* as a *human-centered* field; where "reinterpretation" is done through a further contextualization of the text (specifically though amplifying minority voices), rather than on the basis of one's own agenda. Prominent works in this domain [9, 12] re-interpret Qura'nic verses using different Islamic sources to contextualize the interpretation such as the Prophet's (PBUH) Sayings [*hadith*] and other Qura'nic verses.

Thus, we propose the following research questions:

RQ1 To what extent does increasing plurality awareness within the Muslim community through **community-powered AI** increase understanding of differing perspectives?

RQ2 How does the integration of community-powered AI tools affect the polarization of beliefs both within the Muslim community and with other communities?

To support the investigation of these research questions, this foundational work will first gauge the current atmosphere surrounding the discussion of Islamic perspectives both within the Muslim community and with other communities, through a qualitative analysis of public discussions online (Islamic subreddits) and conducting a user study through a custom web interface.

2 Data and Methods

2.1 Scraping Islamic Subreddits

This preliminary study was conducted with data scraped from different Islamic subreddits, which influenced the design and development of the subsequent user study.

2.1.1 Data Collection

The data collection process involved scraping 4 Islamic-related subreddits (`r/islam`, `r/progressive_islam`, `r/Muslim`, `r/CritiqueIslam`) using the Python Reddit API Wrapper[1] (PRAW). These subreddits include a variety of posts; from general, community-based posts between Muslims (e.g. `r/islam`) to analytical and theology-specific discussions between Muslims and others (e.g. `r/CritiqueIslam`). To obtain a more holistic overview, `r/exmuslim` was also considered, but since it is a private subreddit, scraping was not possible. Furthermore, there was a possibility that it served as an echo chamber of hate towards the religion.

Since all of the subreddits are Islam-focused, the queries were able to be short and concise without possibility of overlap with other topics from other subreddits (e.g. `r/christianity`). The queries included specific verse references (e.g. "2:282") and keywords pertaining to those verses (e.g. "inheritance"). A summary of all the query keywords used can be found in Appendix Table 1. Furthermore, since the subreddits chosen are moderated, only "clean" keywords could be used in the

[1] `https://praw.readthedocs.io/en/stable/index.html`

queries. Specifically, in the context of Category 5, which is concerned with "Jihad" (Arabic: striving for one's best), many mistakenly associate the concept with "terrorism", which was used to query the category. However, since only productive discussion is allowed, this limited the number of keywords of the last category to 2. The moderation in these subreddits will naturally introduce biases into the data collected as it may not fully capture the range of extremities within the topics of discussion, however it is still a considerable starting point for this endeavour.

The scraping process was done in November 2023, where the posts were retrieved according to the default post sorting by `"relevance"`. For each subreddit and each query category (See Appendix Table 1), the scraping was done in batches of $n = 200$ for 4 iterations to account for PRAW rate limits. The posts from all subreddits were then grouped by their respective query categories.

2.1.2 Data Categories and Annotation

After scraping the posts, an annotation guideline was developed with detailed descriptions to minimize inter-annotator disagreement. Considering resource and time limitations, 100 posts were randomly sampled from Focus Category 2 ("Gender Equality & Freedom of Expression") and annotated for fine-grained qualitative analysis by the authors. The annotation categories included "Tone", which describes the perceived tone of the post (e.g. `Insinuative`), "Purpose", which describes the perceived intent behind the post (e.g. `Seeking Advice`), "Scriptural References" which describes whether the post included any reference to the Qur'an or Hadith , and "Search Relevance" which describes the relevance of the post to both the query and the focus category. Full details on the annotation guidelines and descriptions can be found in Appendix Figures 1 - 4.

2.2 "Compassionately" User Study

As another preliminary, a custom web interface was developed to allow users to post their perspectives on religious scripture and interact with others' views, visualized in Appendix Figure 5.

2.2.1 Website Development

The website design and interface were loosely based off a typical social media platform, where users are allowed to post interpretations on specific verses and interact with others' interpretations through likes, dislikes and comments (Appendix Figure 6). Currently, the website only displays one English translation of the Qur'an, The Clear Qur'an [8], which was retrieved through the `quran.com` API [2]. Additionally, to give users quick access to other interpretations, a search feature was implemented that allows users to search by `Chapter`, `Verse` and `Post` (Appendix Figure 7). Keyword search for all categories is supported for both Arabic and English queries. To further encourage contextualized interpretations, users are given the ability to cross-reference other verses and include a range of verses on their posts (e.g. `Al-Fatihah 1-2, Al-Baqarah 35`) (Appendix Figure 8). These verses are embedded as hyperlinks on a user's post, allowing other users to quickly access the full verses for reference if needed. Lastly, users are able to add a subtitle under their username (that appears on their posts) and a bio that describes their qualifications, perspectives and interests so that other users can properly situate their perspectives to further encourage understanding and productive discussion.

2.2.2 Participant Recruitment

Currently, this study (which was approved by the IRB) is being conducted as a part of an Introductory Islamic Theology course at the University of British Columbia. Students are introduced to certain verses in class and their context, then invited to add their interpretations on the website. Consent forms were distributed throughout the class and if signed, students' responses will be anonymized and unidentifiable when included in the final paper.

2.2.3 Content Moderation

The website is implemented with full support for content moderation, allowing moderators (who are assumed to have adequate religious knowledge and authority) to review and approve or reject posts and comments by users, customize the Chapters they moderate, and interact with other moderators' reviews through comments. The moderation system has specific rules describing when a post is to be

[2] `https://api-docs.quran.com/docs/category/quran.com-api`

labeled as `"Pending"`, `"Approved"`, or `"Rejected"`, based on the number of inputs from multiple moderators and users are able to dispute these reviews if needed. However, the authors are aware that the definitions of "adequate religious authority" differ depending on the context and one's own beliefs (e.g. Al-Azhar Imam vs Theology Professor), where the regulations vary based on (Qura'nic) Arabic language proficiency, scholarly certification, etc. So, for the sake of simplicity, the moderation system was removed from the initial user study, allowing users to freely post and comment within a slightly controlled context (i.e. academia). These considerations will be revisited following the commencement of the user study.

3 Results and Discussion

3.1 Scraping Islamic Subreddits

This preliminary study was conducted for the purpose of analyzing the current atmosphere in terms of inter-religious (Muslim-other) and intra-religious (Muslim-Muslim) dialogue in 100 randomly sampled posts from Focus Category 2 ("Women's Rights"). Inter-annotator agreement was measured using Cohen's kappa coefficient (κ) for each annotation category (Table 1), where the annotators demonstrated substantial agreement overall with high statistical significance. The "Search Relevance" category was omitted as it is not relevant to the current analysis. Finer-grained analysis is conducted on the posts identified as engaging in inter-religious and intra-religious dialogue by each annotator. Firstly, the Cohen's κ for inter-annotator agreement on the `"Intra-religious Dialogue"` category was moderate ($\kappa = 0.587$ with $p < 0.001$), while for the `"Inter-religious Dialogue"` category, it was lower but still within the moderate agreement scale ($\kappa = 0.407$ with $p < 0.001$). As such, the category distributions within these posts are analyzed separately for each annotator.

Table 1: Results for the Cohen's κ coefficient measuring inter-annotator agreement for different focus categories, and their corresponding p-values derived from a two-tailed t-test.

Focus Category	κ	p-value
Tone	0.604	$p < 0.001$
Purpose	0.665	$p < 0.001$
Scriptural References	0.698	$p < 0.001$

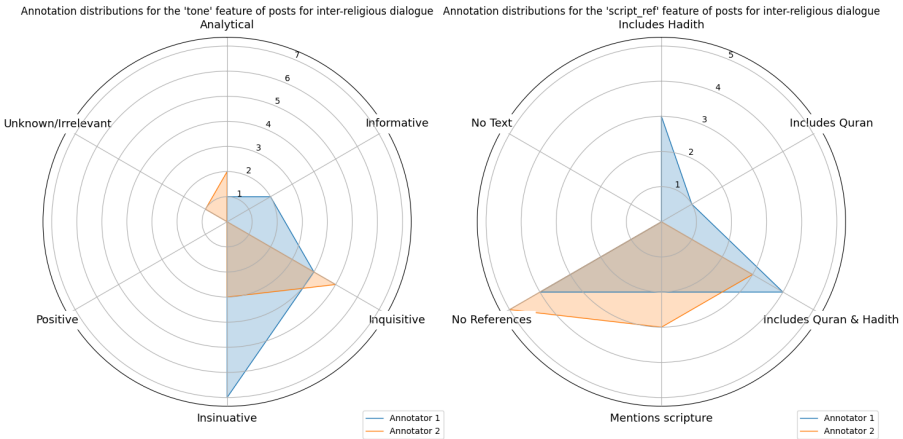

Figure 1: Plots of the `"tone"` (Tone) and `"script_ref"` (Scriptural References) distributions for both the categories on posts that were annotated with `"purpose" = "Inter-religious Dialogue"`.

Figures 1 and 2 demonstrate the "Tone" and "Scriptural References" distributions of the annotator-specific posts within the former categories. For the posts perceived as conducting inter-religious dialogue (Figure 1), the predominant annotated tones were `"Insinuative"` and `"Inquisitive"` across annotators with a higher proportion of `"Insinuative"` posts tagged by Annotator 1. These observations are expected and mirror the empirical experiences of the authors when dealing with

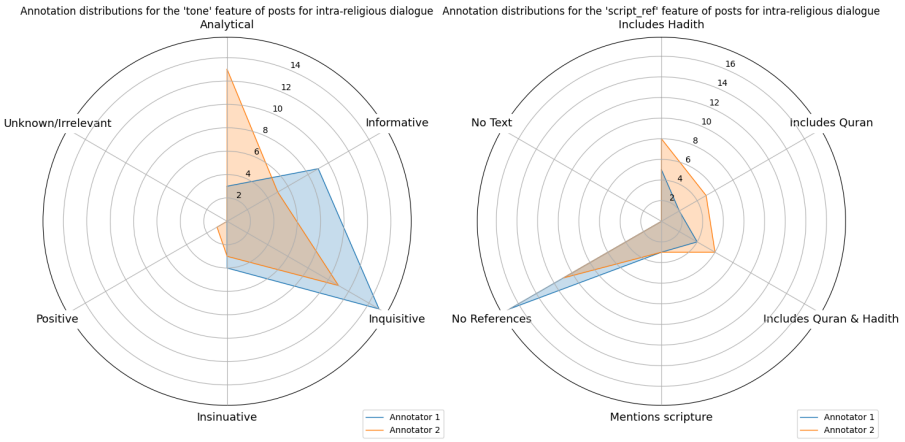

Figure 2: Plots of the annotation distributions for both the `"tone"` (Tone) and `"script_ref"` (Scriptural References) categories on posts that were annotated with `"purpose" = "Intra-religious Dialogue"`.

non-Muslims. In terms of the scriptural references used in these posts, the results were varied, with most posts including no references and some including Qur'an and Hadith excerpts. Interestingly, a fair number of posts were tagged with `"Mentions scripture"`, which indicate that some posts make claims about what is included in the scripture without providing explicit references.

In contrast, the posts identified as conducting intra-religious discussion differ completely in tone (Figure 2), where the predominant "Tone" values were `"Analytical"`, `"Inquisitive"` and `"Informative"`. This is also an expected observation given the authors' experiences, as most Muslims engage in critical analysis to satisfy their curiosity and fill in the gaps in their spiritual knowledge. However, the posts in this category demonstrated a similar distribution of scriptural references to the inter-religious posts, with most containing no references to any scripture, but this could be explained by the large amount of tagged `"Inquisitive"` posts.

Furthermore, subreddit-specific "Tone" distributions were visualized in Appendix Figure 9, where more community-based subreddits like `r/islam` and `r/Muslim` had a larger proportion of `"Inquisitive` and `"Informative"` posts. Although both `r/progressive_islam` and `r/CritiqueIslam` are analytical-based subreddits that are concerned with critically analyzing Islamic views, the latter had a much larger proportion of `"Insinuative"` tagged posts across both annotators, where posts were tagged as `"Insinuative"` majority of the time and no posts were tagged as `"Inquisitive"`. This may be a product of the nature of these subreddits and the people that gravitate towards them; where `r/progressive_islam` seems to be geared more towards Muslims of "all sects and schools of thought" and fosters an "atmosphere of understanding" [3], while `r/CritiqueIslam` is more general and centered around the "respect[ful] discuss[ion] of Islamic theology" [4]. Furthermore, although the subreddits are moderated, we still observe posts that attempt to covertly implicate Islamic beliefs and this is expected to be amplified in unfiltered online discussions on other platforms (e.g. "4chan"). To obtain more representative results, a larger proportion of posts will be annotated from each category, using more annotators with varied backgrounds and more recent posts, to observe tone distributions and the ways in which people interact within those communities.

4 Conclusion

Overall, the studies conducted serve as foundational work to our larger goal of building an automated hate-speech detection tool that can educate its users about the plurality of views within the Muslim community. The Islamic subreddit analysis task was effective in unveiling the current climate of religious discussion within different online communities (subreddits). These observations will be contrasted with the current user study, which is more intentional and held in an academic setting.

[3] `https://www.reddit.com/r/progressive_islam/about/`

[4] `https://www.reddit.com/r/CritiqueIslam/about/`

References

[1] MA al-A Al-Zarqani. Manahil al-'irfan fi'ulumil qur'an. *Beirut: Isa bab Al Halabi. Th*, 1979.

[2] Averroes. *The Book of the Decisive Treatise: Determining the Connection Between the Law and Wisdom*. E. J. Brill, Leiden, 1961. Originally written in the 12th century.

[3] Pew Research Center. The World's Muslims: Religion, Politics and Society, 2013.

[4] Pew Research Center. U.S. Muslims Concerned About Their Place in Society, but Continue to Believe in the American Dream, 2017.

[5] Gallup. Islamophobia: Understanding Anti-Muslim Sentiment in the West, 2023.

[6] Ben Hutchinson. Modeling the Sacred: Considerations when Using Religious Texts in Natural Language Processing. 2024.

[7] Mohammad Hashim Kamali. Methodological issues in islamic jurisprudence. *Arab LQ*, 11:3, 1996.

[8] M. Khattab. *The Clear Quran: A Thematic English Translation of the Message of the Final Revelation*. Book of Signs Foundation, 2016.

[9] Fatima Mernissi. *The veil and the male elite: A feminist interpretation of women's rights in Islam*. Perseus Books Cambridge, MA, 1991.

[10] D. Mogahed and E. Ikramullah. American Muslim Poll 2020, 2020.

[11] D. Mogahed and A. Mahmood. American Muslim Poll 2019: Full Report, 2019.

[12] Amina Wadud. *Qur'an and woman: Rereading the sacred text from a woman's perspective*. Oxford University Press, USA, 1999.

[13] J. Zine. The Canadian Islamophobia Industry: Islamophobia's Ecosystem in the Great White North. *Islamophobia Studies Journal*, 7(2):233–249, 2022.

A Appendix

Table 1: Table describing the query keywords used in PRAW for each Data Focus Category

	Focus Category	Query Keywords
1	Physical Punishments for Theft, Adultery & Slander	`Hudud, hadd, physical punishment, 24:2, 24:3, Nur (Noor) 2, Nur 3, flogging, amputat*, zina (adultery) punishment, slander punishment, theft, 5:38, Ma'idah 38`
2	Womens' Rights: Gender Equality & Freedom of Expression	`feminis*, Baqarah 282, 2:282, women's testimony, inheritance, Nisa 11, 4:11, polygamy, four wives, Nisa 3, 4:3`
3	Womens' Rights: Domestic Violence	`Domestic violence, domestic abuse, guardian*, Nisa 34, 4:34, marital rape, nushuz`
4	LGBTQ+ Rights	`Queer, Lut, Queer muslims, sodomy, LGBT*, gay, lesbian, trans, bisexual, 7:81, Surat Al-A'raf`
5	*Jihad* in the name of Allah	`Jihad, martyr`

Tone

- Positive: This tag is used for posts that are not strictly "informative" and aim to uplift the community through optimistic remarks and helpful practices.
- Informative: This tag is used for posts that aim to pass on information to a specific community within the subreddit, e.g. reposting scriptures or scholarly talks. However, this tag is strictly reserved for information from more validated sources and not just community opinions.
- Inquisitive: This tag is used for posts that are seeking answers to certain questions, whether that be curious non-Muslims looking to revert, or Muslims asking about specific practices and beliefs.
- Analytical: This tag is used for posts that engage in critical interpretation and analysis of scriptures and practices, without covertly implicating the subject(s) of analysis.
- Insinuative: This tag is used for posts that are either inquisitive or analytical but also covertly insinuating certain conclusions, or nitpicking scriptures (Hadiths) out of context and from non-trustworthy sources.
- Unknown/Irrelevant: This tag is used for posts with no discernable tone, such as posts with no body content or posts that *only* contain direct scriptural quotes.

Figure 1: The annotation descriptions for all the values under the category of `Tone`.

Purpose

- Inter-religious Dialogue: The purpose of the post is to engage in dialogue *with* the Muslim community, as a supposed outsider
    - Inter-religious conversation was detected using certain mannerisms in the posts such as explicitly stating that they are non-Muslim, targeting the post to "Muslims of Reddit", and other implicit mannerisms such as not respecting the Prophet's name.
- Intra-religious Dialogue: The purpose of the post is to engage in dialogue from *within* the Muslim community, e.g. discussion of different sects, interpretation or analysis of scripture, etc.
- Community Post: The purpose of the post is to address the specific subreddit community, e.g. posting weekly scripture compilations, moderating discussion threads, etc.
- Giving Advice: The purpose of the post is to give advice by either reiterating personal stories (e.g. reverting), or linking scholarly talks and sources.
- Seeking Advice: The purpose of the post is to seek advice about a certain situation or understanding.
- Unknown: The post does not contain any text in the body, and either contains only a title, or a title and visual content (e.g. video, photos…)

Figure 2: The annotation descriptions for all the values under the category of `Purpose`.

Scriptural References

- Includes Quran & Hadith: The post includes direct quotes of both Quranic verses and Hadith excerpts, with references.
- Includes Quran: The post includes direct quotes of Quranic verses.
- Includes Hadith: The post includes direct Hadith excerpts and references.
- Mentions scripture: The post contains mention of scripture, without any direct quotes or references (e.g. "I know that the Quran says <X>...")
- No References: The post contains no references to any scriptures.
- No Text: The post contains no body text.

Figure 3: The annotation descriptions for all the values under the category of `Scriptural References`.

Search Relevance

- Relevant to query: The content of the post is directly relevant to the search query and the overarching category.
- Relevant to category, but not query: The content of the post is relevant to the overarching category, but deviates from the search query.
- Completely unrelated to category: The content of the post is completely unrelated to the overarching category.
- Unknown/Irrelevant: The post contains no discernible (textual) content.

Figure 4: The annotation descriptions for all the values under the category of `Search Relevance`.

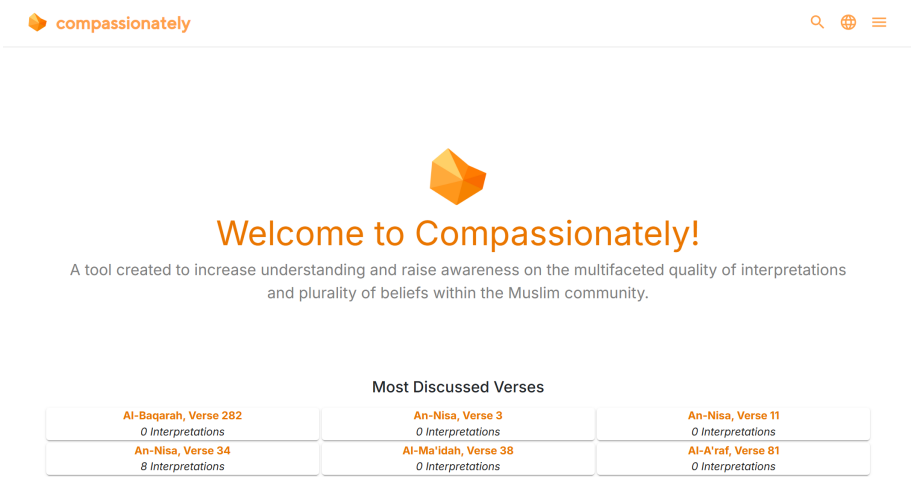

Figure 5: A screenshot of the website developed specifically for the user study.

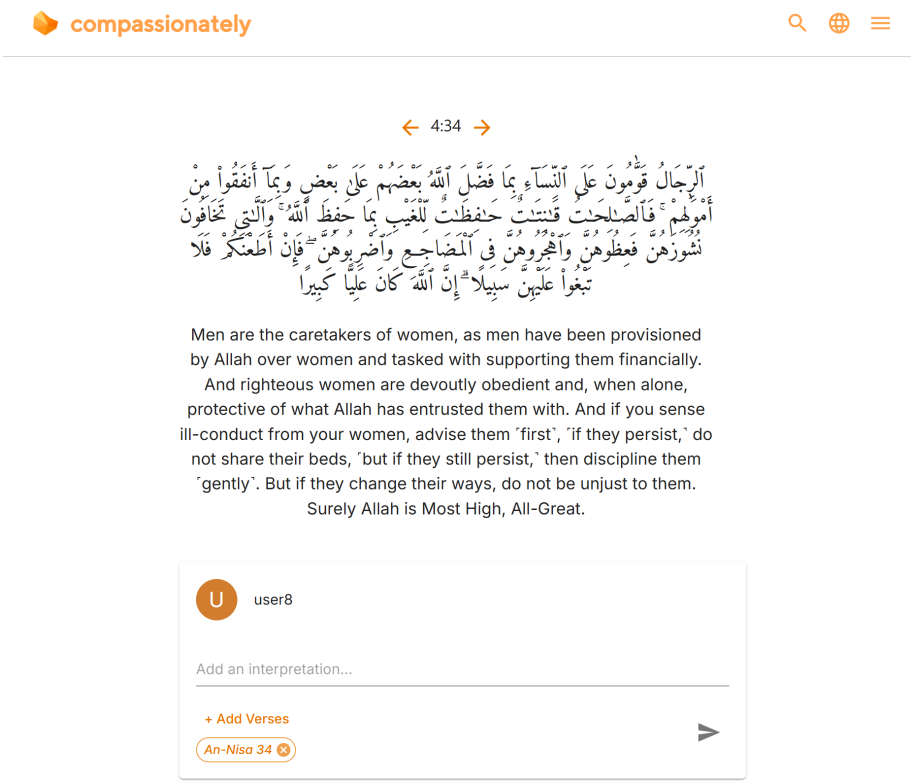

Figure 6: A screenshot of a verse-specific page (An-Nisa, Verse 34) on the "Compassionately" website, where the user can add an interpretation and reference multiple verses in their post.

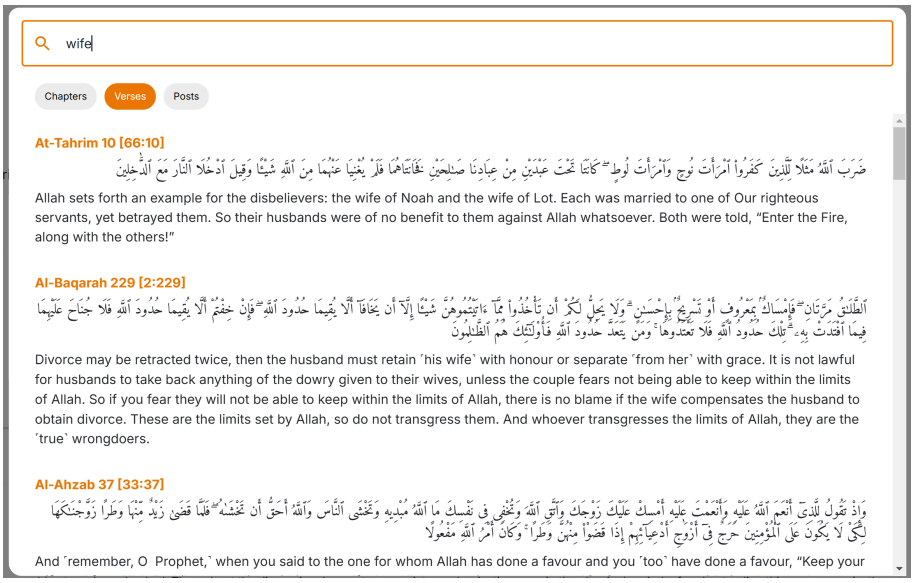

Figure 7: A screenshot of the *Search* feature on the "Compassionately" website, where the input keyword is `"wife"` and the search filter is `"Verses"`.

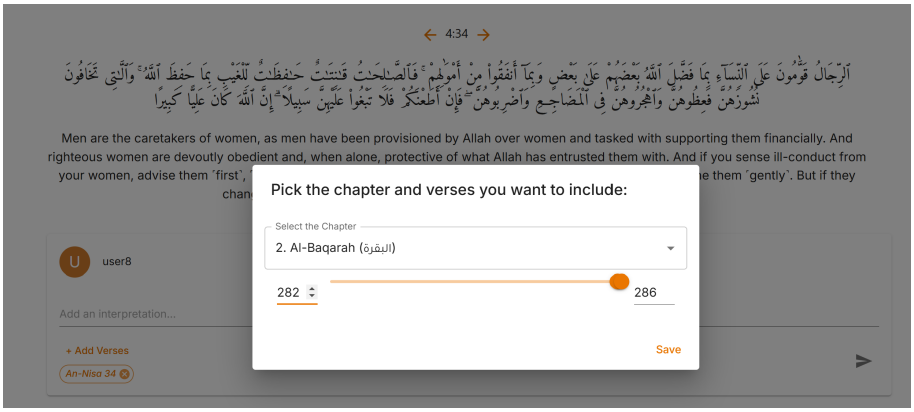

Figure 8: A screenshot of the *Verse Cross-Referencing* feature on the "Compassionately" website, which is triggered through the `+ Add Verses` button, enabling the user to embed quick links to other verses to support their interpretation.

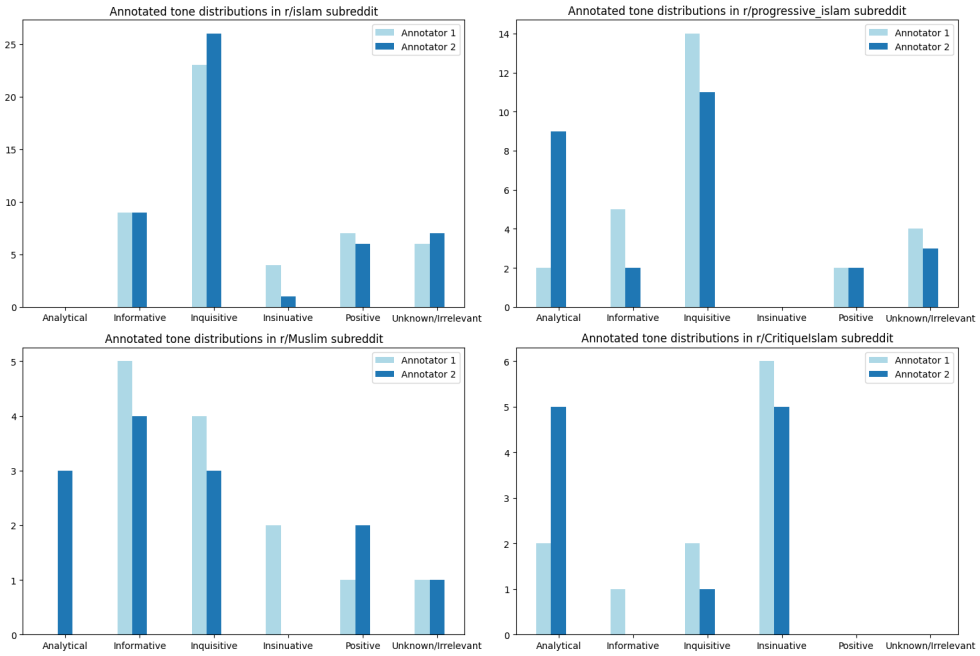

Figure 9: Subreddit-specific plots demonstrating the tone annotation distributions for each subreddit across both annotators.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and the introduction (Section 1) discuss the main contributions of the paper as well as putting the work in context of related work as well as our plans for future work.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: The limitations are not explicitly discussed because of the limited length of the paper, however, since the work was primarily focused on conducting a qualitative analysis of online posts, the design choices were limited anyway. One possible limitation could be that the annotators are of the same religious background (which is mentioned in the analysis in Section 3).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: No theoretical results or proofs were included in the paper.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include the queries in Appendix Figure 1 and the subreddits used for scraping in Section 2, as well as the annotation descriptions used by the annotators in Appendix Figures 1 - 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code used in this paper was all referenced from the official PRAW documentation, which is explicitly mentioned and cited.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We include our experimental setup in the Data and Methods Section 2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: For our experiments, which included a qualitative analysis of annotator results, error bars and statistical significance are not applicable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: No compute resources were used in the study.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors abide by the NeurIPS Code of Ethics through the distribution of consent forms and the preservation of anonymity and privacy throughout the work, as outlined in Section 2.2.2.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We discuss the use case and context within which the system we propose is being developed in and it's potential for combating Islamophobia 1.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: All the data that was scraped from Reddit came from moderated subreddits, which do not allow unsafe or violent posts on their platforms 2.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All assets used are properly cited (i.e. the Quran.com API and the Python Reddit API Wrapper) 2.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No assets are introduced in this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: A description of how the participants are instructed to interface with the website is given under Section 2.2.2.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: We state that this research has been IRB approved under Section 2.2.2.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.