For Women, Life, Freedom: A Participatory AI-Based Social Web Analysis of a Watershed Moment in Iran’s Gender Struggles

Adel Khorramrouz, Sujan Dutta, Ashiqur R. KhudaBukhsh*
Rochester Institute of Technology
{ak8480, sd2516, axkvse}@rit.edu

Abstract
In this paper, we present a computational analysis of the Persian language Twitter discourse with the aim to estimate the shift in stance toward gender equality following the death of Mahsa Amini in police custody. We present an ensemble active learning pipeline to train a stance classifier. Our novelty lies in the involvement of Iranian women in an active role as annotators in building this AI system. Our annotators not only provide labels, but they also suggest valuable keywords for more meaningful corpus creation as well as provide short example documents for a guided sampling step. Our analyses indicate that Mahsa Amini’s death triggered polarized Persian language discourse where both fractions of negative and positive tweets toward gender equality increased. The increase in positive tweets was slightly greater than the increase in negative tweets. We also observe that with respect to account creation time, between the state-aligned Twitter accounts and pro-protest Twitter accounts, pro-protest accounts are more similar to baseline Persian Twitter activity.

1 Introduction
Words are the only victors.

On 16 September 2022, Mahsa Amini, a 22-year-old woman died under police custody in Iran. Reportedly, she was arrested because of not wearing her hijab (headscarf) properly. As media and police presented conflicting accounts of her death [Alkhaldi and Mostaghim, 2022], Mahsa Amini’s death enraged Persian (Farsi) Twitter users in an unprecedented manner [Kermani, 2023]. The hashtag #مهساامینی (#MahsaAmini) became one of the most repeated hashtags on Persian Twitter and initiated a Twitter protest where Iranians expressed their grievances against the government like never before. Support and solidarity for gender equality poured in from prominent world leaders [France-Presse, 2022], artists [Pina, 2022], and sports personalities [Alkhaldi, 2022] across the globe.

#MahsaAmini was undoubtedly the overwhelming top-trending hashtag on Persian Twitter for months during the relentless protest. However, for a brief period of time, hashtags with an opposite stance toward the protest (e.g., #ExecuteThem or #ISupportKhamenei)1 trended. Prior literature conjectured state-aligned trolling in Iran on Instagram [Kargar and Rauchfleisch, 2019]. Also, social bot accounts’ capability to spread extreme ideology is well-documented [Stella et al., 2018; Berger and Morgan, 2015].

Via a substantial corpus of 30.5 million tweets relevant to the protest, this paper makes three key observations:
1. The grievances of protesters against the current government mention a broad range of incidents spanning decades.
2. With respect to account creation time, between the state-aligned Twitter accounts and pro-protest Twitter accounts, pro-protest accounts are more similar to baseline Persian Twitter activity.
3. There was a noticeable shift in positive stance toward gender equality after the protests on Persian Twitter discourse.

To our knowledge, no computational analysis relying on sophisticated natural language processing methods exists that has examined gender equality in Persian social media discourse let alone at this unprecedented scale. That said, we believe our key contribution lies elsewhere. Our paper marks an important effort to include the stakeholders – the Iranian women – in this AI-building process. All examples in our supervised solution’s training set are annotated by Iranian women. Our examples are thus grounded in cultural contexts and first-person experience about the gender struggles in Iran.

Datasets addressing issues faced by vulnerable communities often end up being annotated by annotators with little or no documentation [Guest et al., 2021; Ramesh et al., 2022]. Since annotated examples often form the core of a supervised AI system, it is important to involve stakeholders in the annotation process. For example, Ramesh et al. [2022] present a lexicon of queer-related inappropriate words where one of the annotators identifies as queer. Similarly, Guest et al. [2021] present a misogyny dataset where the majority of the annotators identifies as queer. Ali Khamenei is the second and current supreme leader of Iran who is in office since 1989.

*Ashiqur R. KhudaBukhsh is the corresponding author.
tors identify as women.

Our annotators’ role is not limited to mere annotation. Rather, they take an active role in guiding how to curate more meaningful data by suggesting suitable keywords to curate our dataset and providing a valuable seed set of examples to initiate an active learning pipeline. Our results indicate that the annotators’ contributions yielded a richer seed set than a random baseline.

At a philosophical level, we see this work as a part of the growing conversation of participatory AI [Harrington et al., 2019; Delgado et al., 2022; Bondi et al., 2021; Birhane et al., 2022] where the goal is to develop systems for the people and by the people.

2 Datasets

As we already mention, #MahsaAmini initiated a protest with global participation. Understandably, tweets in global languages such as English or French are likelier to reflect the global perspective on this issue. Given that Twitter is banned in Iran and users reportedly use VPNs to access Twitter [Kermani, 2023], considering geo-tagged tweets is not a reliable option either to understand and analyze the Iranian perspective. Therefore, we restrict our analyses to only tweets authored in the Persian language. We assume that our choice of language can act as an effective filter to ensure our dataset is less likely to be diluted by the global discourse.

We use Twitter’s official language label as ground truth. We collect three corpora: \(D_{\text{protest}}\), \(D_{\text{gender}}\), and \(D_{\text{baseline}}\).

Our dataset spans the time duration of Jan 15, 2022, to Jan 15, 2023.\(^2\) We define the time period from January 15, 2022, to September 15, 2022, as \(T_{\text{before}}\). We define the time period from September 16, 2022, to January 15, 2023, as \(T_{\text{after}}\). A short description follows next. Throughout the paper, if we use a Persian word or phrase, we present an English translation in parentheses following the word.

2.1 \(D_{\text{gender}}\)

\(D_{\text{gender}}\) consists of 6,036,012 Tweets which has been posted by 700,189 unique users.

1. All tweets that have either “زنان” (women) or “دختر” (girl).
2. All tweets that have either “ناموس” which means the immediate female family members (daughter, mother, sister, wife) whom the male member of the family (father, brother, husband) should protect and sometimes control or “غریب” which means the positive form of jealousy that men have upon their female family members against other men. These two search keywords were suggested by our annotators.
3. All tweets that have gender insult words against women “چندم” and “چندمین” both indicating “a prostitute” or “a promiscuous woman” in a pejorative way (the second insult word mostly accompanies sister).

2.2 \(D_{\text{protest}}\)

\(D_{\text{protest}}\) consists of:

1. tweets with the hashtag #MahsaAmini in them yielding 21,308,449 Tweets posted by 655,303 unique users.
2. tweets that support government through the hashtag #پیامداران_ایران. This hashtag has been used 1,051,792 times by 71,484 unique users across the entire Twitter timeline accessible through the APIs.
3. tweets that have the hashtag عمامه کنید which means execute them. This hashtag has been used 11,292 times by 5,130 unique users across the entire Twitter timeline accessible through the APIs.

2.3 \(D_{\text{baseline}}\)

In order to estimate baseline Persian Twitter behavior, we consider five Persian stop words (که از د. پا. ب. and collect 6,000 tweets per day (evenly distributed across the hours) that contain at least one of these stop words. Our dataset, \(D_{\text{baseline}}\), consists of 2,190,000 tweets.

We compute the unigram distributions of subsets of \(D_{\text{baseline}}\) that was authored during \(T_{\text{before}}\) and \(T_{\text{after}}\). Table 1 lists the top 20 high-frequency non-stop words present when (1) we subtract the unigram distribution of \(T_{\text{after}}\) from the unigram distribution of \(T_{\text{before}}\) (left); and (2) we subtract the unigram distribution of \(T_{\text{before}}\) from the unigram distribution of \(T_{\text{after}}\) (left). In plain English, these are the words that appeared more frequently during one period and much less frequently during the other. From the right column of Table 1, we note that several of these words are not indicative of civic unrest while the left column does not indicate similar unrest. We conduct a similar experiment to track shift in high-frequency hashtag usage between the two time periods. We again observe that even in the baseline Persian Twitter discourse, \(T_{\text{after}}\) showed several hashtags relevant to the protest.

On January 7, 2023, two executions relevant to this protest happened [Radford and Fowler, 2023]. We thus set our end date one week after the executions.

<table>
<thead>
<tr>
<th>(T_{\text{before}})</th>
<th>(T_{\text{after}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>“خوبی” (goodness)</td>
<td>“خون” (blood)</td>
</tr>
<tr>
<td>“پاک” (clean)</td>
<td>“خوشی” (happiness)</td>
</tr>
<tr>
<td>“زمان” (time)</td>
<td>“خیابان” (street)</td>
</tr>
<tr>
<td>“خویش” (life)</td>
<td>“خواب” (sleep)</td>
</tr>
<tr>
<td>“خداوند” (God)</td>
<td>“خواب” (sleep)</td>
</tr>
</tbody>
</table>

Table 1: Biggest shift in token usage in \(D_{\text{baseline}}\) between \(T_{\text{before}}\) and \(T_{\text{after}}\).
Table 2: Shift in top hashtags present in $D_{\text{baseline}}$ between $T_{\text{before}}$ and $T_{\text{after}}$.

<table>
<thead>
<tr>
<th>Trigram</th>
<th>Translation</th>
</tr>
</thead>
</table>
| #ExecuteThem | at least once in our dataset. Finally, $U_{\text{baseline}}$ is the set of unique users who contributed to $D_{\text{baseline}}$. We compute the account creation time for each user at the granularity of months and obtain normalized histograms for each of these sets. Figure 1 illustrates the account creation temporal distributions of the four user sets. All subsets exhibit a sharp spike after the Cinema Rex fire incident that happened in 1978 [Ali and Ali, 2018], to the attacks on student-dormitory in 1999 [Bozorgmehr, 2017], to the current reality of web censorship, the most striking takeaway from Table 4 is perhaps the time duration between the events people mentioned.

### 4 Account Creation Time

Prior literature has examined state-aligned trolling in Iran on Instagram platforms [Kargar and Rauchfleisch, 2019]. In this section, we present an analysis based on account creation time. We define four sets of users: $U_{\text{pro-protest}}$, $U_{\text{state-aligned}}$, $U_{\text{pro-execution}}$; and $U_{\text{baseline}}$. $U_{\text{pro-protest}}$ represents all users who used the hashtag #MahsaAmini (indicating support for Mahsa Amini) at least once in our dataset. $U_{\text{state-aligned}}$ represents all users who used the hashtag #ISupportKhamenei (indicating support for Ali Khamenei) at least once in our dataset. $U_{\text{pro-execution}}$ represents the set of users who used the hashtag #ExecuteThem at least once in our dataset. Finally, $U_{\text{baseline}}$ is the set of unique users who contributed to $D_{\text{baseline}}$. We compute the account creation time for each user at the granularity of months and obtain normalized histograms for each of these sets. Figure 1 illustrates the account creation temporal distributions of the four user sets. All subsets exhibit a sharp spike around September 2022, however $U_{\text{pro-execution}}$ exhibits two different spikes. In fact, in September 2022, Google Trends indicates one of the most popular search queries from Iran was “دانلود توییتر” (download Twitter).

Table 5 computes the KL divergence of the account creation time distributions with respect to $U_{\text{baseline}}$. We observe that the distribution of $U_{\text{pro-protest}}$ is closest to $U_{\text{baseline}}$ while $U_{\text{pro-execution}}$ is the farthest. The order remains unchanged with other distributional distance measures (e.g., Bhattacharyya distance). Our qualitative findings remain unchanged even if we limit $U_{\text{state-aligned}}$ and $U_{\text{pro-execution}}$ to only those user accounts that used these hashtags during $T_{\text{before}}$ and $T_{\text{after}}$.

---

5 Annotation

All our annotation was conducted by four different annotators. All annotators identify as Iranian women and are fluent speakers of Persian. All of them have undergraduate degrees.

**Annotation task.** Our text prediction task is to predict the stance toward gender equality. For each tweet, we ask the annotator: *does this short document indicate a positive, neutral, or negative stance toward gender equality?*

**Inter-rater Agreement.** Between any two annotators, we have at least 500 overlapping samples. Across all rounds of annotation, the Cohen’s $\kappa$ ranged from 0.41 to 0.52. On a misogyny annotation task, Guest et al. [2021] reported Fleiss’ $\kappa$ of 0.48 and the Krippendorf’s alpha as 0.49. Sanguinetti et al. [2018] report category-wise $\kappa = 0.37$ for offence and $\kappa = 0.54$ for hate. We further note that our observed inter-rater agreement is higher than Gomez et al. [2020] ($\kappa = 0.15$) and Fortuna and Nunes [2018] ($\kappa = 0.17$).

**Disagreement resolution.** Since our task is likely to be subjective, resolving disagreements has to be grounded in the literature. Prior literature has considered diverse approaches to resolving inter-annotator disagreements (e.g., majority voting [Davidson et al., 2017; Wiegand et al., 2019] or third objective instance [Gao and Huang, 2017]). We resolve any disagreement in the following manner. For positives and neutrals, we only consider consensus labels. Following Golbeck et al. [2017], if any annotator marks an example as negative and the other annotator marks it as negative or neutral, we consider the aggregate label as negative. In order to ensure the anonymity of the annotators, we do not conduct any post-annotation adjudication step to resolve disagreements.

### 5.1 Toward Participatory AI

A notable feature in our work is the active involvement of Iranian annotators in both corpus creation and annotation. Our annotators helped us in the following two ways.

**Search keywords.** At a deeper level, which data could contain relevant information may require a clear understanding of the social realities. While constructing $D_{gender}$, choosing woman or girl and gendered insults as search keywords required little cultural context. However, our annotators suggested nuanced keywords such as "ناموس" and "غیرت" to be included in our list of search keywords. Recall that, "ناموس" means the immediate female family members (e.g., daughter, mother, sister, or wife) whom the male members (e.g., father, brother, or husband) should protect and sometimes control; and "غیرت" means a positive form of jealousy that men have upon their female family members against other men.

**Seed set.** A notable feature of our work is the active involvement of Iranian women in the annotation process, where they not only provide labels but also present important representative short documents to construct meaningful seed sets during the guided sampling step described in Section 6.

### 6 Active Learning Pipeline

**Research question:** Is there a noticeable change in support for gender equality in Persian Twitter discourse before and after the demise of Mahsa Amini while in police custody?

To estimate the support for gender equality in Persian Twitter discourse, we build a robust classifier detecting content supportive of gender equality. Since hashtag hijacking [Hadgu et al., 2013] is a common phenomenon where users with opposite views may use the most-popular hashtag to express an opposite stance, our goal is to predict the stance toward gender equality from tweet texts only.

We first estimate to which extent tweets supporting gender equality are present in $D_{baseline}$. We randomly sample 500 tweets weighing both $T_{before}$ and $T_{after}$ equally (i.e., 250 from each time slice). In addition, we randomly sample 1,000 tweets from $D_{gender}$ weighing equally $T_{before}$ and $T_{after}$. Table 6 summarizes the label distribution. We note that a large fraction of $D_{baseline}$ consists of neutral tweets.
In order to construct a dataset that is diverse and representative of the unlabeled pool, we present an active learning pipeline that consists of well-known sampling steps. A short description of active learning follows next.

6.1 Background

Active Learning is a powerful and well-established form of supervised machine learning technique [Settles, 2009]. It is characterized by the interaction between the learner, aka the classifier, and the teacher (oracle or labeler or annotator) during the learning process. At each iteration, the learner employs a sampling strategy to select an unlabeled sample (unlabeled samples) and requests the supervisor to label it (them) in agreement with the target concept. The data set is augmented with the newly acquired label, and the classifier is retrained on the augmented data set. The sequential labeling and re-training process continues until some halting condition is reached (e.g., annotation budget is expended or the classifier has reached some target performance). At this point, the algorithm outputs a classifier, and the objective for this classifier is to closely approximate the (unknown) target concept in the future. The key goal of active learning is to reach a strong performance at the cost of fewer labels. Since retraining the model and running inference on a large, unlabeled pool is computationally costly, prior literature has examined the trade-offs present in a batch active learning setting [Yang and Carbonell, 2013]. In this work, we follow the batch active learning setting.

6.2 Seed Set Construction

Random Sampling. In order to capture a diverse set of examples, we randomly select 1,000 samples from $\mathcal{D}_{gender}$ and 500 samples from $\mathcal{D}_{baseline}$. Table 6 indicates that solely relying on $\mathcal{D}_{baseline}$ to construct the seed set will result in extreme class imbalance with very few positives and negatives and predominantly neutrals. Sampling from $\mathcal{D}_{gender}$ might yield slightly more positives (and negatives), however, a keyword-based starting point runs the risk of biasing the whole active learning pipeline. In what follows, we present a guided sampling approach similar to Palakodety et al. [2020].

Guided Sampling. When faced with the challenge to find high-quality positive examples championing the Rohingya community, Palakodety et al. [2020] proposed a document-embedding-based, guided sampling method where annotators provide example short documents conforming to a given label. We employ a similar technique where we asked three annotators to provide five examples each indicating positive and negative stances toward gender equality. For each example, we select 25 unique nearest neighbors in the document embedding space from the unlabeled pool giving equal weightage to tweets from $T_{before}$ and $T_{after}$. This yields 750 samples. Upon annotation and resolving disagreements, we obtain 166 positives, 145 negatives, and 231 neutrals. We note that our sampling method yielded substantially more positives (and negatives) than the random sampling baseline.

Table 7 presents a few randomly selected positive and negative seed examples provided by our annotators. We observe that the examples are grounded in women’s cultural struggle in Iran [Kazemzadeh, 2002]. Beyond discussions around hijab, inequality in marital and inheritance law [Doherty et al., 2021], restrictions on activities such as visiting stadiums to watch football [Lewis, 2019; Abtahi et al., 2022] echoed in these examples.

Table 8 lists a random sample of retrieved tweet texts when we used the guided sampling method. This table shows that not only we found more positives (and negatives) than the random baseline, but the tweet texts also exhibit richness, diversity, and nuance.

Overall, we obtain 343 positives, 440 negatives, and 1,051 neutrals from the random sampling and guided sampling step. In what follows, we describe two well-known sampling strategies that we employ to further expand our dataset.

6.3 Certainty and Uncertainty Sampling

Certainty sampling. Since our goal is to use the trained model for a social inference task, it is important to rectify high-confidence misclassifications. Minority class certainty sampling has found its use in rectifying high-confidence misclassifications involving short documents such as movie reviews and messages [Sindhwani et al., 2009; Attenberg et al., 2010]; search queries [KhudaBuksh et al., 2015]; and comments on YouTube videos [Palakodety et al., 2020; Yoo and KhudaBuksh, 2023]. We conduct certainty sampling for the positive class and select 750 instances that the model predicts as positive with the highest confidence. We also conduct certainty sampling for the negative class and select 750 instances that the model predicts as negative with the highest confidence. In this step, we obtain 338 positives, 345 negative, and 487 neutrals.

Uncertainty sampling. Uncertainty sampling is one of the most well-known sampling strategies used in active learning [Settles, 2009]. Since we have multiple label categories in our prediction task, we use margin sampling, an active learning variant designed for multiple labels [Scheffer et al., 2001]. In this step, we sample 1,500 examples. Upon annotation and resolving the disagreements, we obtain 115 positives, 247 negatives, and 819 neutrals.

To summarize, our active learning pipeline consists of the following steps:

1. Construct an initial seed set by randomly sampling from $\mathcal{D}_{random}$ and $\mathcal{D}_{gender}$, and using guided sampling ($\mathcal{D}_{seed}$ : 343 positives, 440 negatives, and 1,051 neutral instances) using random sampling.
2. Conduct certainty sampling on the positive class and certainty sampling on the negative class ($\mathcal{D}_{certainty}$ : 338 positives, 345 negatives, and 487 neutral instances).
3. Finally, conduct uncertainty sampling (margin sampling) ($\mathcal{D}_{uncertainty}$ : 115 positive, 247 negative, and 819 neutral instances).
Why does a woman, who is an adult, and has the ability to go to another country, need to seek another man's permission to leave regardless of whatever the relation is?

We have to put aside all the differences in our standards and if we consider a behavior appropriate for men, we should allow women as well, and understand that there is no difference. There is no difference between a man and a woman who smoke or dance or laugh loud at a party.

When you see these hijabless girls' attire in the street, you can be sure that they see freedom in being naked and they are not looking for anything else.

If girls like football so much, they can sit at home and watch it. They don't have to go to the stadium and attract attention.

Table 7: Random sample of seed examples presented by annotators. Blue indicates a positive stance toward gender equality and red indicates a negative stance toward gender equality.

<table>
<thead>
<tr>
<th>Seed examples produced by annotators</th>
<th>Translation</th>
</tr>
</thead>
</table>
| جداکن زن که ادم بالایی است و خوشابی این داره که بر هی کشور دیگر، باز از یک نفر مرد دیگر اجازه خروج بگیره؟ حالا با هر نسبت؟
| Why does a woman, who is an adult, and has the ability to go to another country, need to seek another man's permission to leave regardless of whatever the relation is? |
| باید این همه تفاوت در استانداردهای مان را کنار بیانادزیم و اگر رفتاری را برای مرده مناسب می دانیم بر زنها هم رو بداریم و درک کنیم که هیچ فرقی نیست بین مرد و زن که توی یک مهمانی سیگار می کشند یا می رقصند یا با صدای بلند می خندند |
| When it comes to divorce, women should have the same rights as men, that means, they should have the right to custody of the child and the right to divorce, etc. |
| خانمهای باید هنگام طلاق حقوق برای آقایان داشته باشند یعنی حق حضانت بچه و حق طلاق و.. از داشته باشد. |
| When you see these hijabless girls' attire in the street, you can be sure that they see freedom in being naked and they are not looking for anything else. |
| آدم این وضعیت پوشش این دخترازی بی حجاب رو که توی خیابون می بینی بهتر مطمئن می شه که این ازدی رو توی همون این حس نشدن می بینی و غیر از این دنیال هیچی نیستن. دختر آنه خیابون فوتیال دوست دارن حب شبنیم تو خونه بیبنین، نمیخواهند برای جلب توجه برین اسکادیومه |
| It is obvious that a woman's blood money should be half a man's. Men are the breadwinners of the family, and the blood money goes to the women and children. Do not tinker around with the law for no good reason. |

Figure 2: Temporal trend of tweets expressing positive and negative stance toward gender inequality on $D_{\text{baseline}}$. Figure 2 indicates that the discourse became more polarized during $T_{\text{after}}$ with both percentages of tweets expressing positive and negative stances increasing. However, we also observe that the increase in positive discourse (by a factor of 2.89) is greater than the increase in negative discourse (by a factor of 1.50).

7 Discussions

In this paper, we present the first-ever computational analysis (to the best of our knowledge) of the stance toward gender equality in Persian Twitter discourse following a watershed moment in Iran’s history. Our analyses reveal that the grievances of Persian Twitter users against the government span decades and the protest following Mahsa Amini’s death perhaps presented an outlet for the angst harbored for a long time. Second, we observe that the distribution of account creation time can present important signals. We find that with respect to account creation time, pro-execution and state-aligned user sets are distributionally different from baseline Persian Twitter users.

We follow an ensemble active learning pipeline to construct a robust classifier that detects stance toward gender equality. As a step towards participatory AI, our annotators take an active role in building our machine learning model. There is a growing concern that our ML conversations barely include marginalized community which can further widen the gap of AI-haves and AI-have-nots. All our annotators are Iranian women, with first-person experience of gender struggles. Their role in our system was far more profound than typical annotators. In a guided sampling step, they provided seed examples to expand our dataset lending cultural grounding. They also suggest important keywords to curate our dataset.
Table 8: Random sample of tweet texts retrieved through guided sampling. Blue indicates a positive stance toward gender equality and red indicates a negative stance toward gender equality.

<table>
<thead>
<tr>
<th>Data</th>
<th>Model</th>
<th>Macro F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>D_{seed}</td>
<td>M_{seed}</td>
<td>67.09 ± 1.46</td>
</tr>
<tr>
<td>D_{seed} ∪ D_{certainty}</td>
<td>M_{certainty}</td>
<td>69.28 ± 0.51</td>
</tr>
<tr>
<td>D_{seed} ∪ D_{certainty} ∪ D_{uncertainty}</td>
<td>M_{uncertainty}</td>
<td>73.27 ± 1.87</td>
</tr>
</tbody>
</table>

Table 9: Performance comparison of models trained on various stages of our active learning pipeline. M_{seed} denotes a popular Persian language model [Farahani et al., 2021] trained on D_{seed}. Subsequent models are fine-tuned on top of this. For all models trained in this paper, performance is reported over five different training runs on a fixed evaluation set of randomly sampled 400 instances from our annotated dataset ensuring no overlap between train and tests.

Table 4 suggests that a major Iranian grievance is limited access to the internet. Free and fair access to Twitter was many users’ wish. While we were working on this paper, Twitter as a platform underwent several significant changes. With the depreciation of academic Twitter and developer accounts being monetized, at this point, it is unclear how much of our collected data would be accessible in the future and at what cost. This is a curious juxtaposition of a community longing for access to a platform to voice their concerns while the very same platform is limiting academic researchers’ access to study global politics.

On top of the current uncertainties surrounding Twitter, the inherently transient nature of the social web, censorship, and fear of persecution can contribute to missing content for post-hoc analyses. In that sense, our paper is a humble attempt to preserve a vulnerable chunk of the social web that chronicled a watershed moment in the gender struggles of Iranian history.

**Ethical Statement**

We use publicly available tweets collected using academic Twitter API. Since our data is highly sensitive, we only conduct aggregate analyses without revealing personally identifiable information. We also do not conduct any post-annotation adjudication steps that are typical to many annotation tasks to ensure the privacy of the annotators.

We trained our model on top of a large language model. Several lines of recent research have indicated that large language models have a wide range of biases that reflect the texts on which they were originally trained, and which may percolate to downstream tasks [Bender et al., 2021].

**Acknowledgements**

We thank Lipika Mazumdar and Shawnee Fereydouni for their input.

**References**


[Alkhaldi and Mostaghim, 2022] Celine Alkhaldi and Ramin Mostaghim. Iranian police say death of Mahsa Amini ‘unfortunate’ as protestors take to the streets, 2022. CNN.


