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.
– Salman Rushdie; *Victory City*; 2023.

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 #مهسا_امینی (*#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 over 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 annota-

*Ashiqur R. KhudaBukhsh is the corresponding author.

[1]Ali Khamenei is the second and current supreme leader of Iran who is in office since 1989.

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: $\mathcal{D}_{protest}$; $\mathcal{D}_{gender}$; and $\mathcal{D}_{baseline}$.

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 $\mathcal{T}_{before}$. We define the time period from September 16, 2022, to January 15, 2023, as $\mathcal{T}_{after}$. 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 $\mathcal{D}_{gender}$

$\mathcal{D}_{gender}$ 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 "غیرت" 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 "جنده" and "کصده" both indicating "a prostitute" or "a promiscuous woman" in a pejorative way (the second insult word mostly accompanies with `sister`).
4. all tweets that have at least one word from the two following subsets: {"دختر" (*girl*), "زن" (*woman*), "خواهر" *sister*}; and
{"زندگی" (*life*), "انقلاب"(*revolution*), "حقوق" (*rights*), "آزادی" (*freedom*)}.

2.2 $\mathcal{D}_{protest}$

$\mathcal{D}_{protest}$ consists of:

1. tweets with #مهسا_امینی (*#MahsaAmini*) in them yielding 21,308,449 Tweets posted by 655,303 unique users.
2. tweets that support government through the hashtag #لبیک_یا_خامنه_ای. 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 #اعدام_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 $\mathcal{D}_{baseline}$

In order to estimate baseline Persian Twitter behavior, we consider five Persian stop words (که, از, در, با, به) and collect 6,000 tweets per day (evenly distributed across the hours) that contain at least one of these stop words. Our dataset, $\mathcal{D}_{baseline}$, consists of 2,190,000 tweets.

We compute the unigram distributions of subsets of $\mathcal{D}_{baseline}$ that was authored during $\mathcal{T}_{before}$ and $\mathcal{T}_{after}$. Table 1 lists the top 20 high-frequency non-stop words present when (1) we subtract the unigram distribution of $\mathcal{T}_{after}$ from the unigram distribution of $\mathcal{T}_{before}$ (left); and (2) we subtract the unigram distribution of $\mathcal{T}_{before}$ from the unigram distribution of $\mathcal{T}_{after}$ (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, $\mathcal{T}_{after}$ showed several hashtags relevant to the protest.

More presence during $\mathcal{T}_{before}$	More presence during $\mathcal{T}_{after}$
"خوش" (*happy*), "انسان" (*human*), "زبان" (*language*), "آقا" (*M.R*), "نمیدونم" (*I do not know*), "توییتر" (*Twitter*), "پسر" (*boy*), "قبول" (*ok*), "گوش" (*ear*), "حد" (*limit*), "خواب" (*sleep*), "جدید" (*new*), weareoneEXO, "دلیل"(*reason*), "عشق"(*love*), "روسیه"(*Russia*), "الله"(*god(Allah)*), "نظرم"(*my opinion*), "حس"(*feeling*), "خوبی"(*goodness*),	"مادر" (*mother*) "کشته" (*killed*) "صدای" (*voice*) "مرگ" (*death*) "آزادی" (*freedom*) "خون" (*blood*) "جمهوری" (*republic*) "خبر" (*news*) "اعدام" (*execution*) "بخاطر" (*for sake of*) "هشتگ" (*hashtag*) "ادامه" (*continue*) "خانواده" (*family*) "نظام" (*can be translated to regime but not exact*) "شهر" (*city*) "لطفا" (*please*) "انقلاب" (*revolution*) "اسم" (*name*) "ج.ا" (*I.R (stands for Islamic republic which represents regime)*) "خیابون" (*street*)

Table 1: Biggest shift in token usage in $\mathcal{D}_{baseline}$ between $\mathcal{T}_{before}$ and $\mathcal{T}_{after}$.

[2] 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.

Top ten hashtags during $\mathcal{T}_{before}$	Top ten hashtags during $\mathcal{T}_{after}$
EXO, "مکتب_امید" (*the_hope_attitude*), "ما_ملت_امام_حسینیم" (*we_are_nation_of_Imam_Hossein*), "ماه_امید" (*month_of_hope*), "ایران_قوی" (*strong_Iran*), "حب_الحسین_یجمعنا" (*love_of_Hossein_gathers_us*), "اوکراین" (*Ukraine*), EXO (*in Korean*), "اللهم_عجل_لولیک_الفرج" (*Oh God, please fasten merging relief (Imam Zaman)for us*), "عید_امید" (*Hope_Eyd*),	"مهسا_امینی" (*mahsa_amini*), "اعتصابات_سراسری" (*Nationwide_strikes*), OpIran, MahsaAmini, StopHazaraGenocide, IRGCterrorists, "نیکا_شاکرمی" (*Nika_Shakarami*), "زن_زندگی_آزادی" (*women_life_freedom*), Mahsa_Amini, "اعتراضات_سراسری" (*Nationwide_protests*), "محسن_شکاری" (*Mohsen_Shekari*),

Table 2: Shift in top hashtags present in $\mathcal{D}_{baseline}$ between $\mathcal{T}_{before}$ and $\mathcal{T}_{after}$.

Line from Baraye	Translation	Percentage of match
برای دختری که آرزو داشت پسر بود	for a girl who wished she were a boy	24.86
برای آزادی	for freedom	15.72
برای خواهرم خواهرت خواهرامون	for my sister, your sister, our sisters	9.76
برای این همه شعار های تو خالی	For all these meaningless slogans	7.91
برای این بهشت اجاری	For this forced "heaven"	7.16
تغییر مغز ها که پوسیدن برای	for changing rusted minds	7.09
برای چهره ای که میخنده	For smiling faces	4.33
برای کودک زباله گرد و آرزو هاش	for child labor and their crushed dreams	3.82

Table 3: Percentage of *because of* tweets that matched with individual lines in Grammy-winning song Baraye by Shervin Hajipour.

3 Baraye – Because Of

A large fraction of tweets of $\mathcal{D}_{protest}$ contains a phrase (*because of*). These tweets were an outlet for Persian Twitter users to vent their frustrations about the situation. Specifically, the tweets aimed at answering a reason for the protests. The tweets expressing a complex collection of emotions on why the current government has failed the nation captured the imagination of Shervin Hajipour, a talented Iranian singer who won the 2023 Grammy award for his song Baraye (*because of*). Each line of this song starts with `because of` and paints a picture of Iranian hope and despair. We collect 1.92 million tweets from $\mathcal{D}_{protest}$ with the phrase *because of*. We train a FastText [Bojanowski *et al.*, 2017] word embedding on $\mathcal{D}_{protest}$ and for each tweet, we compute the line in the Baraye song that is the nearest neighbor in the embedding space. Table 3 lists the top 10 lines from the Baraye song that matched with the tweets.

While the poignant song by Hajipour brilliantly captures Iranian aspirations and struggles, Table 4 indicates there is much more to Iranian angst than what the song could hold. The second most common trigram indicates the access barrier to Twitter. Most prominent social media platforms are blocked in Iran [Kargar and McManamen, 2018; Sohrabi, 2021] and reportedly, Iranians primarily take recourse to VPNs to participate in the social web [Kermani, 2023]. We observe that Navid Afkari[3], an executed Iranian wrestler, is mentioned among the top trigrams. Multiple structural failures got also mentioned in the common trigrams [Bozorgmehr, 2017]. From 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.

Trigram	Translation
حدیث نجفی مهسا	Hadis Najafi Mahsa [Amini]
زن زندگی آزادی	women life freedom
دسترسی توییتر ندارد	have not Twitter access
رمزی میشود آزادی	[your name] will become [the] symbol [of] freedom
نامت رمزی میشود	your name will become symbol
بختیاری نوید افکاری	[Pouya] Bakhtiari Navid Afkari
های ریخته شده	spilled [blood]
پلاسکو سانچی کولبران	Plasco Sanchi kolbar
متروپل آبان پویا	Metropoll Aban (November) Pouya [Bakhtiari]
سینما رکس کوی	Cinema Rex kouye [university]

Table 4: Top ten trigrams from *because of* tweets. We omit two spam trigrams (e.g., please follow me) aimed at gaining followers. Hadis Najafi was killed by a gunshot during the Mahsa Amini protest.

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: $\mathcal{U}_{pro\text{-}protest}$; $\mathcal{U}_{state\text{-}aligned}$; $\mathcal{U}_{pro\text{-}execution}$; and $\mathcal{U}_{baseline}$. $\mathcal{U}_{pro\text{-}protest}$ represents all users who used the hashtag #MahsaAmini (indicating support for Mahsa Amini) at least once in our dataset. $\mathcal{U}_{state\text{-}aligned}$ represents all users who used the hashtag #ISupportKhamenei (indicating support for Ali Khamenei) at least once in our dataset. $\mathcal{U}_{pro\text{-}execution}$ represents the set of users who used the hashtag #ExecuteThem at least once in our dataset. Finally, $\mathcal{U}_{baseline}$ is the set of unique users who contributed to $\mathcal{D}_{baseline}$. 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 $\mathcal{U}_{pro\text{-}execution}$ 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 $\mathcal{U}_{baseline}$. We observe that the distribution of $\mathcal{U}_{pro\text{-}protest}$ is closest to $\mathcal{U}_{baseline}$ while $\mathcal{U}_{baseline}$ 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 $\mathcal{U}_{state\text{-}aligned}$ and $\mathcal{U}_{pro\text{-}execution}$ to only those user accounts that used these hashtags during $\mathcal{T}_{before}$ and $\mathcal{T}_{after}$.

[3] https://www.hrw.org/news/2020/09/12/iran-suddenly-executes-wrestler-navid-afkari

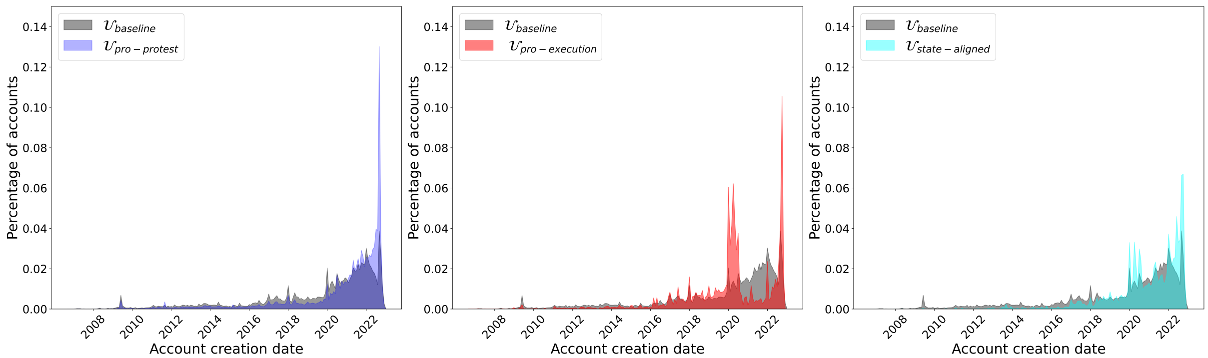

Figure 1: Distributions of account creation dates of different user sets.

User Set	KL-Divergence
$\mathcal{U}_{pro\text{-}protest}$	0.15
$\mathcal{U}_{state\text{-}aligned}$	0.25
$\mathcal{U}_{pro\text{-}execution}$	0.46

Table 5: KL-divergence of the distribution of account creation time for different user subsets with respect to the distribution of account creation time of $\mathcal{U}_{baseline}$.

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 κ ranged from 0.41 to 0.52. On a misogyny annotation task, Guest *et al.* [2021] reported Fleiss' κ 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 $\mathcal{D}_{gender}$, 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 $\mathcal{D}_{baseline}$. We randomly sample 500 tweets weighing both $\mathcal{T}_{before}$ and $\mathcal{T}_{after}$ equally (i.e., 250 from each time slice). In addition, we randomly sample 1,000 tweets from $\mathcal{D}_{gender}$ weighing equally $\mathcal{T}_{before}$ and $\mathcal{T}_{after}$. Table 6 summarizes the label distribution. We note that a large fraction of $\mathcal{D}_{baseline}$ consists of neutral tweets.

Dataset	$\mathcal{T}_{before}$			$\mathcal{T}_{after}$		
	Positive	Neutral	Negative	Positive	Neutral	Negative
$\mathcal{D}_{gender}$	11.9%	51.9%	36.1%	30.4%	35.3%	34.2%
$\mathcal{D}_{baseline}$	0.4%	98.3%	1.2%	2.8%	93.4%	3.6%

Table 6: Label distribution of the first stage of annotation (random sampling) during the seed set construction.

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 label-requesting 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 $\mathcal{T}_{before}$ and $\mathcal{T}_{after}$. 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 [KhudaBukhsh *et al.*, 2015]; and comments on YouTube videos [Palakodety *et al.*, 2020; Yoo and KhudaBukhsh, 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).

Seed examples produced by annotators	Translation
چرا یک زن که آدم بالغی است و خودش عقل داره و توانایی اینو داره که بره یه کشور دیگه، باید از یک نفر مرد دیگه اجازه خروج بگیره؟ حالا با هر نسبتی	*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 بیاندازیم و اگر رفتاری را برای مردها مناسب می دانیم بر زنها هم روا بداریم و درک کنیم که هیچ فرقی نیست بین مرد و زنی که توی یک مهمانی سیگار می کشند یا می رقصند یا با صدای بلند می خندند	*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.*
خانمها باید هنگام طلاق حقوق برابر با آقایون داشته باشند یعنی حق حضانت بچه و حق طلاق و... را داشته باشند.	*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.*
آدم این وضعیت پوشش این دخترای بی حجاب رو که توی خیابون میبینه بیشتر مطمئن میشه که اینا آزادی رو توی همون لخت شدن میبینن و غیر از این دنبال هیچی نیستن.	*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.*
دخترا اگه خیلی فوتبال دوست دارن خب بشینن تو خونه ببینن، نمیخواد برای جلب توجه برن استادیوم	*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. .*
معلومه که باید دیه زن نصف مرد باشه؛ مردا نون آور خونواده هستن و تازه دیه به زن و بچهه میرسه به نفعشونم هست، گیر الکی به قانون ندید	*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.*

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.

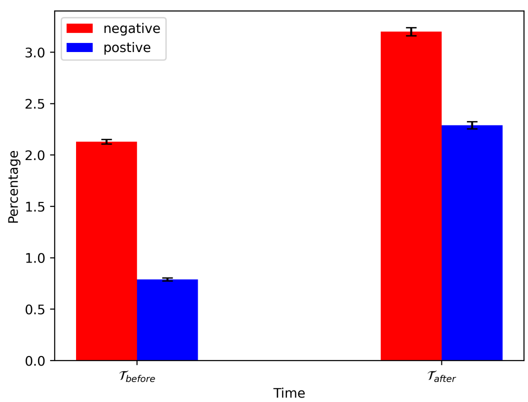

Figure 2: Temporal trend of tweets expressing positive and negative stance toward gender inequality on $\mathcal{D}_{baseline}$.

Overall, we obtain 796 positive, 1,032 negative, and 2,357 neutral examples.

6.4 Model Performance and Analysis

Table 9 summarizes the performance of our trained models. The performance improves at each active learning step and we finally achieve a Macro F_1 performance of 73.27%. We note that if we train a binary classifier with just the positive class and the negatives and neutrals clubbed together as the *notPositive* class, it is possible to achieve slightly better performance (Macro F_1: 77.76 $\pm$ 2.29).

To track shifts in stance toward gender equality, we run inference using $\mathcal{M}_{certainty}$ on $\mathcal{D}_{baseline}$. Figure 2 indicates that the discourse became more polarized during $\mathcal{T}_{after}$ 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.

Examples obtained through guided sampling	Translation
حق حضانت برای تو! درد زایمان برای من نام خانوادگی برای تو! زحمت خانواده برای من سند خانه به نام تو! بیگاری خانه برای من چهار عقد برای تو! حسرت عشق برای من هزار صیغه برای تو! حکم سنگسار برای من هوس برای تو! عفاف برای من این بود برابری حقوق زن و مرد؟	*Custody right for you! Labor pain for me Surname for you! Family trouble for me The document of the house in your name! Hardwork in home for me Four marriages for you! Missing love for me A thousand concubines for you! A sentence of stoning for me Lust for you! chastity for me Are these equal rights for men and women?*
حجاب خانم های خونی ، نسبی، ثبتی من نه تنها به کسی مربوط نیست بلکه به بنده هم مربوط نیست زن یک انسان مستقل هست دست از مالکیت بردارید زنان برده هیچ مردی نیستند آنان که خود را تسلیم مرد میکنند نه تنها به خود بلکه به زنان دیگر هم خیانت میکنند	*Neither me nor anyone can tell my relatives if they should wear a hijab or not. Women are not slaves of men. Those women who submit themselves to men are not only betraying themselves, but also betraying other women.*
میدونی دیگه ایران زن نمیتونه جدا شه مگر دلایلی بیاره که . . . .! درحالی که یک مرد بای دیفالت اون حق و داره . همسرشما داشته از اول. و شما تازه به دستش اوردی و هم سطح شدی.. . .	*You know, Iranian women can't ask for divorce unless they give reasons.! While a man by default has that right. Your husband had it from the beginning. And you just earned it and leveled up. . . .*
مردا بهتر از زنان کار می کنند	*Men work better than women.*
خود زنها عرضه ندارن حقشون رو بگیرن ، مردها مقصرند ؟!! چند بار تا حالا شنیدی که زنها برای گرفتن حق حضور در استادیوم برن جلوی فدراسیون تجمع کنن ؟ ولی از مردها انتظار دارن که به استادیوم نرن تا از زنها حمایت بشه . متاسفانه اکثریت جامعه زنان ایران فقط غر زدن رو بلدن .	*Is this men's fault that women themselves do not have ability to take their rights?!! How many times have you heard that women gather in front of the federation to get the right to attend the stadium? But they expect men not to go to the stadium to support women. Unfortunately, the majority of Iranian women only know how to nag .*
اون آزادی بیشتری که بعضی ها می‌خوان شخصیت زن رو منحصر به زیبایی های ظاهری می‌کنه و این هم برای خودش و هم برای جامعه ضرر داره	*The more freedom people want for women, the more women's character gets overshadowed by physical beauty, which is harmful to both herself and the society.*

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.

Data	Model	Macro F_1
$\mathcal{D}_{seed}$	$\mathcal{M}_{seed}$	67.09 ± 1.46
$\mathcal{D}_{seed} \cup \mathcal{D}_{certainty}$	$\mathcal{M}_{certainty}$	69.28 ± 0.61
$\mathcal{D}_{seed} \cup \mathcal{D}_{certainty} \cup \mathcal{D}_{uncertainty}$	$\mathcal{M}_{uncertainty}$	73.27 ± 1.87

Table 9: Performance comparison of models trained on various stages of our active learning pipeline. $\mathcal{M}_{seed}$ denotes a popular Persian language model [Farahani *et al.*, 2021] trained on $\mathcal{D}_{seed}$. 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 deprecation 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

[Abtahi *et al.*, 2022] Zahra Abtahi, Leila Zahedi, Zarrin Eizadyar, and Nicole M Fava. # bluegirl: A study of collective trauma on twitter. *Journal of Traumatic Stress*, 35(6):1631–1641, 2022.

[Ali and Ali, 2018] Luman Ali and Luman Ali. Reacting to iran's descent into chaos. *British Diplomacy and the Iranian Revolution, 1978-1981*, pages 67–99, 2018.

[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.

[Alkhaldi, 2022] Celine Alkhaldi. Iranian woman takes part in international chess tournament without mandatory hijab. https://www.cnn.com/2022/12/28/sport/iran-sara-khadem-chess-tournament-hijab-spt-intl/index.html, 2022. CNN.

[Attenberg *et al.*, 2010] Josh Attenberg, Prem Melville, and Foster Provost. A unified approach to active dual supervision for labeling features and examples. In *ECML/PKDD*, pages 40–55, 2010.

[Bender *et al.*, 2021] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *ACM FaccT*, pages 610–623, 2021.

[Berger and Morgan, 2015] Jonathon M. Berger and Jonathon Morgan. *The ISIS Twitter Census: Defining and describing the population of ISIS supporters on Twitter*. Brookings Institution, 2015.

[Birhane *et al.*, 2022] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Díaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. Power to the people? opportunities and challenges for participatory ai. *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8, 2022.

[Bojanowski *et al.*, 2017] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[Bondi *et al.*, 2021] Elizabeth Bondi, Lily Xu, Diana Acosta-Navas, and Jackson A. Killian. Envisioning communities: a participatory approach towards ai for social good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 425–436, 2021.

[Bozorgmehr, 2017] Shirzad Bozorgmehr. Six days that shook Iran. http://news.bbc.co.uk/2/hi/middle_east/828696.stm, 2017. CNN.

[Davidson *et al.*, 2017] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515, 2017.

[Delgado *et al.*, 2022] Fernando Delgado, Solon Barocas, and Karen Levy. An uncommon task: Participatory design in legal ai. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–23, 2022.

[Doherty *et al.*, 2021] William J. Doherty, Seyed Mohammad Kalantar, and Mahdieh Tarsafi. Divorce ambivalence and reasons for divorce in Iran. *Family process*, 60(1):159–168, 2021.

[Farahani *et al.*, 2021] Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. Parsbert: Transformer-based model for Persian language understanding. *Neural Processing Letters*, 53:3831–3847, 2021.

[Fortuna and Nunes, 2018] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

[France-Presse, 2022] Agence France-Presse. Iran protests: Joe Biden says US stands with 'brave women' after Mahsa Amini death. https://www.theguardian.com/us-news/2022/oct/15/iran-protests-joe-biden-says-us-stands-with-brave-women-after-mahsa-amini-death, 2022. The Guardian.

[Gao and Huang, 2017] Lei Gao and Ruihong Huang. Detecting online hate speech using context aware models. In Ruslan Mitkov and Galia Angelova, editors, *RANLP 2017*, pages 260–266. INCOMA Ltd., 2017.

[Golbeck *et al.*, 2017] Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, et al. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, pages 229–233, 2017.

[Gomez *et al.*, 2020] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478, 2020.

[Guest *et al.*, 2021] Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online, April 2021. Association for Computational Linguistics.

[Hadgu *et al.*, 2013] Asmelash Teka Hadgu, Kiran Garimella, and Ingmar Weber. Political hashtag hijacking in the us. In *Proceedings of the 22nd international conference on world wide web*, pages 55–56, 2013.

[Harrington *et al.*, 2019] Christina Harrington, Sheena Erete, and Anne Marie Piper. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25, 2019.

[Kargar and McManamen, 2018] Simin Kargar and Keith McManamen. Censorship and collateral damage: Analyzing the Telegram ban in Iran. *Berkman Klein Center Research Publication*, (2018-4), 2018.

[Kargar and Rauchfleisch, 2019] Simin Kargar and Adrian Rauchfleisch. State-aligned trolling in iran and the double-edged affordances of instagram. *New media & society*, 21(7):1506–1527, 2019.

[Kazemzadeh, 2002] Masoud Kazemzadeh. *Islamic fundamentalism, feminism, and gender inequality in Iran under Khomeini*. University Press of America, 2002.

[Kermani, 2023] Hossein Kermani. # mahsaamini: Iranian twitter activism in times of computational propaganda. *Social Movement Studies*, pages 1–11, 2023.

[KhudaBukhsh *et al.*, 2015] Ashiqur R. KhudaBukhsh, Paul N. Bennett, and Ryen W. White. Building effective query classifiers: a case study in self-harm intent detection. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1735–1738, 2015.

[Lewis, 2019] Samantha Lewis. Death of Blue Girl shines light on women's rights in Iran. https://www.theguardian.com/football/2019/sep/21/death-of-blue-girl-shines-light-on-womens-rights-in-iran, 2019. The Guardian.

[Palakodety *et al.*, 2020] Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. Voice for the voiceless: Active sampling to detect comments supporting the rohingyas. In *AAAI 2020*, volume 34, pages 454–462, 2020.

[Pina, 2022] Christy Pina. Asghar Farhadi Invites Artists to Declare Solidarity With the People of Iran. https://www.hollywoodreporter.com/news/general-news/asghar-farhadi-solidarity-people-of-iran-mahsa-amini-1235227129/, 2022. The Hollywood Reporter.

[Radford and Fowler, 2023] Antoinette Radford and Sarah Fowler. Iran protests: Two men hanged over killing of militiaman. https://www.bbc.com/news/world-middle-east-64196635, 2023. BBC.

[Ramesh *et al.*, 2022] Krithika Ramesh, Sumeet Kumar, and Ashiqur R. Khudabukhsh. Revisiting queer minorities in lexicons. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 245–251, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics.

[Sanguinetti *et al.*, 2018] Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

[Scheffer *et al.*, 2001] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden Markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318. Springer, 2001.

[Settles, 2009] Burr Settles. *Active learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences, 2009.

[Sindhwani *et al.*, 2009] Vikas Sindhwani, Prem Melville, and Richard D. Lawrence. Uncertainty sampling and transductive experimental design for active dual supervision. In *ICML*, pages 953–960, 2009.

[Sohrabi, 2021] Hadi Sohrabi. New media, contentious politics, and political public sphere in Iran. *Critical Arts*, 35(1):35–48, 2021.

[Stella *et al.*, 2018] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, 115(49):12435–12440, 2018.

[Wiegand *et al.*, 2019] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers)*, pages 602–608, 2019.

[Yang and Carbonell, 2013] Liu Yang and Jaime Carbonell. Buy-in-bulk active learning. *Advances in neural information processing systems*, 26, 2013.

[Yoo and KhudaBukhsh, 2023] Clay H. Yoo and Ashiqur R. KhudaBukhsh. Auditing and Robustifying COVID-19 Misinformation Datasets via Anticontent Sampling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):15260–15268, 2023.