



*This work is dedicated to the memory of Mahsa Amini, Nika Shahkarami, and Sarina Esmaeilzadeh. Mahsa, a young woman of just 22, faced an untimely death after being detained by the “Morality Police” for improper Hijab. Her tragic passing ignited the Women.Life.Freedom movement. Nika and Sarina, only 16, fearlessly took to the streets to join in the cries of Women.Life.Freedom, but their lives were tragically cut short. Their voices, now silent, still echo in our minds, serving as a stark reminder of the human struggle for basic rights. May their courage inspire us to seek equality, truth, and justice relentlessly.*

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Research Questions . . . . .	2
<b>2</b>	<b>Computational Propaganda: Technologies, Challenges, and Countermeasures</b>	<b>4</b>
2.1	Propaganda and its Techniques . . . . .	4
2.2	Computational Propaganda . . . . .	6
2.2.1	Problems of Computational Propaganda . . . . .	6
2.2.2	Challenges in Addressing Computational Propaganda . . . . .	7
2.2.3	Textual Analysis Perspective . . . . .	8
2.2.4	Network Analysis Perspective . . . . .	9
2.2.5	Future Outlook . . . . .	10
<b>3</b>	<b>Digital Dissidence: The Women Life Freedom Movement in Iran</b>	<b>12</b>
3.1	Grassroots Movements and Social Media Activism . . . . .	12
3.2	Iran's Historical Context . . . . .	13
3.3	Women and Feminist Digital Activism in the Iranian Context . . . . .	14
3.4	Political Field on Iranian Twittersphere . . . . .	15
<b>4</b>	<b>Dataset: Iranian Twittersphere</b>	<b>16</b>
4.1	Data Collection . . . . .	16
4.2	Data Characteristics . . . . .	17
<b>5</b>	<b>Propaganda Analysis: Case Study of Women Life Freedom Movement</b>	<b>20</b>
5.1	Background Information and Related Work . . . . .	20
5.1.1	Topic Analysis . . . . .	20
5.1.2	Social Network Analysis (SNA) . . . . .	22
5.1.3	Related Work in SNA . . . . .	23
5.1.4	Related Work in Iranian Tweetersphere Propaganda Analysis . . . . .	24
5.2	Methodology . . . . .	24
5.2.1	Research Design . . . . .	24
5.2.2	Unsupervised Clustering . . . . .	25
5.2.3	Hashtag Labeling . . . . .	26
5.2.4	Trend Analysis . . . . .	27
5.3	Results and Analysis . . . . .	28
5.3.1	Unsupervised Clustering . . . . .	29
5.3.2	Hashtag Labeling . . . . .	29
5.3.3	Affiliation Assignment . . . . .	31

5.3.4	Trend Analysis . . . . .	45
5.4	Deadends . . . . .	56
5.4.1	Labeling Twitter Data . . . . .	56
<b>6</b>	<b>Explaining Propaganda Techniques With LLMs</b>	<b>57</b>
6.1	Background Information and Related Work . . . . .	57
6.1.1	Natural Language Processing . . . . .	57
6.1.2	Large Language Models . . . . .	58
6.1.3	Explainable AI . . . . .	59
6.1.4	Explaining AI in Natural Language . . . . .	61
6.1.5	Related Work in Explainable Propaganda Detection . . . . .	61
6.1.6	Related Work in AI Generated Explanations . . . . .	62
6.2	Dataset . . . . .	63
6.3	Methodology . . . . .	65
6.3.1	Augmenting PTD with GPT3.5 Explanations . . . . .	65
6.3.2	Curation of Training Data . . . . .	67
6.3.3	Experimental Setting . . . . .	68
6.3.4	Evaluation . . . . .	68
6.4	Results . . . . .	68
6.4.1	Explanation Augmented PTD (exAug-PTD) Dataset . . . . .	68
6.4.2	Evaluation . . . . .	71
6.5	A Methodology for PropagandaTeacherGPT . . . . .	75
<b>7</b>	<b>Discussion, Conclusion and Future Work</b>	<b>78</b>
7.1	Discussion and Conclusion . . . . .	78
7.2	Future Work . . . . .	79
<b>A</b>	<b>Unsupervised Clustered Community Wordclouds</b>	<b>91</b>
<b>B</b>	<b>Labeled Hashtags</b>	<b>101</b>
<b>C</b>	<b>Algorithmically Affiliation Assigned Community Wordclouds</b>	<b>107</b>
<b>D</b>	<b>First 20 M.H.P Tweets</b>	<b>112</b>

## **Abstract**

This thesis investigates Twitter discussions related to the *Women Life Freedom* movement in Iran, which emerged after the tragic death of Mahsa Jina Amini while in police custody. Employing network analytics and topic modeling, the study reveals how the three main political factions engaged with the movement-specific hashtags to implement their propaganda strategies. The findings confirm previous observations that the regime in power manipulated hashtags by intentionally misspelling and utilizing varied encodings, tactics aimed at diluting the movement's online presence. Additionally, the use of emotionally charged hashtags was prevalent, strategically designed to either further the factions' agendas or counteract the movement. These observations contribute to a growing body of knowledge on online manipulation strategies on social media. The rise of Large Language Models (LLMs) introduces concerns about their potential to be exploited, providing propaganda with a cost-effective boost. However, this research demonstrates that LLMs can also be employed in identifying propaganda techniques and offer a promising avenue for public education and awareness of propaganda methods.

# Chapter 1

## Introduction

Social media platforms are key players in global communication, activism, and information sharing [21]. They have the power to bring together diverse communities, highlight marginalized voices, and facilitate large-scale movements for social and political change [117, 21]. However, alongside these benefits, dissemination of misinformation and propaganda on these platforms has emerged as a significant challenge, creating a potential for misrepresentation of facts and influencing public opinion on a broad scale [127].

How does online propaganda manifest, and how can it be detected and analyzed? To answer this, the exploration is bifurcated into two dimensions. Chapter 5 pivots towards a Network Analysis lens, examining the propagation of propaganda within the Women Life Freedom, also known as the Mahsa Amini movement. This section adopts a network-centric analytical methodology to dissect the intricate web of propaganda networks and their pervasive influence on public discourse within the Iranian Twittersphere. Chapter 6 embarks on an understanding from the perspective of Natural Language Processing (NLP) - it delves into the capabilities of Large Language Models (LLMs) to classify and explain propaganda techniques in natural language.

To set the foundation for these explorations, Chapter 2 lays out the background information on computational propaganda, its mechanisms, and potential countermeasures. Chapter 3 provides context about the social movement in Iran, specifically focusing on the Women's movement. Chapter 4 delves into the dataset used in this research, detailing its collection and structure. Finally, Chapter 7 presents a conclusion and directions for future research in this domain.

### 1.1 Motivation

The diffusion of propaganda on social media platforms, such as Twitter, has emerged as a significant concern globally, with potential repercussions for individuals, societies, and governments [127]. The capacity of such digital manipulation to shape public opinion and even steer political trajectories is evidenced by recent significant global events [5, 104]. Despite increasing consciousness of these issues, propaganda strategies, comprehensive understanding, and effective mitigation strategies are still evolving [123]. With the advancement of AI, especially LLMs, propagandists may gain an additional, cost-effective, powerful tool to further their causes. There is a pressing need for research and the development of real-time tools to educate the public and raise awareness about propaganda techniques within the social media landscape.

The dichotomy of social media platforms' role in societal discourse forms the crux of this study's motivation. While these platforms can empower marginalized voices, foster global con-

versations, and act as catalysts for social and political changes, they also offer a prolific environment for propaganda and misinformation spread, potentially undermining the associated benefits [21].

The grassroots movement Women Life Freedom—also known as the Mahsa Amini movement—originating in the context of Iran’s stringent media censorship and socio-political climate, offers an intriguing case to explore these issues. As a manifestation of digital activism and women’s rights advocacy, the movement offers unique insights into the interplay between grassroots activism and computational propaganda [63].

## 1.2 Research Questions

While social media platforms have emerged as dominant channels for global communication, their vulnerability to the dissemination of propaganda remains a pressing concern. The mechanisms and strategies behind this digital propagation are not fully understood [78], and the integration of advanced AI tools, such as LLMs, can compound the problem. The challenge is to discern and decode these evolving propaganda techniques in real time and within the vast datasets inherent to social media platforms. This research seeks to bridge this gap by focusing on the detection and analysis mechanisms of propaganda on social media, particularly within the Women Life Freedom in the Iranian Twittersphere, and investigate the use of LLMs to address online propaganda.

This research aims to answer the following research questions:

**RQ1:** *Who are the political-ideological groupings active on Twitter in the wake of the Women Life Freedom movement?*

**RQ2:** *How do these ideologically conflicting groupings use hashtags, and how do they relate to propaganda techniques?*

**RQ3:** *What are the propaganda strategies and techniques these groupings use dominantly over time in their discourse, and what strategies have exerted the most significant influence on public discourse?*

**RQ4:** *While ChatGPT-like tools can contribute to online manipulation by lowering the cost of creating content, how can we leverage it to counter computational propaganda?*

This thesis makes the following contributions:

### 1. Algorithm for Ideological Affiliation Assignment:

- Proposed an algorithm to determine ideological affiliations utilizing selected Twitter hashtags identified through unsupervised clustering methodology and subsequent labeling.
- Applied this algorithm to the discourse surrounding the “Women Life Freedom” movement in the Iranian Twittersphere.

### 2. Analysis of Political Ideological Communities in Iranian Twittersphere:

- Identified three primary political ideological communities within the dataset.
- Analyzed the usage of specific hashtags by each community, relating several to established propaganda techniques.

- Observed that different ideological groups utilized identical and different hashtags for their respective agendas.
  - (a) *Mahsa Amini Hashtag Utilization*: All groups prominently used Mahsa Amini hashtags in conjunction with their respective leaders to promote their ideologies and agendas.
  - (b) *Manipulative Hashtag Variations*: Confirmed Kermani's [63] observation regarding insertion of intentional misspellings of the Mahsa Amini hashtag to dilute the movement's online presence. A similar approach was identified with the OpIran hashtag, using 'I' instead of 'I'.
  - (c) *Post Mahsa Amini's Death Hashtag Evolution*: Identified a discernible increase in activity and shift towards more radical hashtags after Mahsa Amini's death.
  - (d) *Differential Hashtag Usage of "Women.Life.Freedom"*: This hashtag showed varying usage levels among the communities. It is used by one group in their top 10 hashtags, used less by the other group in their top 50—to blame and denigrate the movement—and is not used by the third group in their top 50.
  - (e) *Emotionally Loaded Hashtags*: Some groups utilized hashtags related to significant events or figures to evoke emotional responses, while others invoked another specific incident in a similar manner.
  - (f) *Common Propaganda Techniques*: Various propaganda techniques were identified in the hashtag usage of different groups, many of which are shared among them.

### 3. N-gram-based Trend Analysis:

- Proposed a methodology for identifying dominant discourse topics and tracking discourse trends over time using N-gram analysis.
- The application of methodology illustrates that these groups seem to use similar and favor different propaganda techniques. However, only one group was identified in successfully shifting the dominant discourse during the observation period.

### 4. AI Generation of Natural Language Explanations of Their Label Prediction:

- Proposed a data augmentation methodology using ChatGPT to generate explanations with varied objectives. This augmented data was then used to fine-tune LaMMA 2 for three distinct tasks: generating explanations using two different prompts and objectives, and standalone label prediction. Empirical findings suggest that not many data annotations with explanations are required to fine-tune large language models in line with the findings of Zhou et al. [134] that fine-tuning LLMs on a carefully curated limited training set can achieve comparable results with state-of-the-art.
- Acknowledging the limitations and inherent biases in AI-generated explanations, a methodology was outlined for the development and assessment of Propaganda Teacher GPT, an educational AI reinforced with human explanation annotations tailored for various audiences.

## Chapter 2

# Computational Propaganda: Technologies, Challenges, and Countermeasures

In the digital age, computational propaganda has emerged as a sophisticated tool, harnessing the power of algorithms, AI and automation to influence public opinion on an unprecedented scale [127]. This chapter delves into computational propaganda, which uses algorithms and automation to manipulate public opinion. It examines historical propaganda definition, techniques, and the challenges of detecting and countering this digital manipulation. The text explores two main detection perspectives: analyzing the content and analyzing network behavior. The future outlook suggests an integrated approach combining both perspectives, addressing multimedia propaganda, and fostering public awareness.

### 2.1 Propaganda and its Techniques

Propaganda exhibits various characteristics, prominently its persuasive function, a vast target audience, representation of a certain group's agenda, and the employment of either faulty reasoning or emotional appeals [81]. Originating in the 17th century, the term "propaganda" was initially affiliated with promoting the Catholic faith in the New World [60]. However, it soon garnered a negative connotation, extending its meaning to oppose Protestantism. A definition by the Institute for Propaganda Analysis [42] defines propaganda as:

Expression of opinion or action by individuals or groups deliberately designed to influence opinions or actions of other individuals or groups concerning predetermined ends.

Bolsover and Howard [14] dissect this definition, spotlighting its emphasis on intentional opinion manipulation. While propaganda is an expansive concept, it primarily influences opinions through myriad rhetorical and psychological techniques [30]. Although the definition of propaganda has a consensus in literature, the techniques—and their precise definitions—witness variance across scholars [112]. For instance, while Miller [81] identifies seven techniques, Weston [125] lists 24. Da San Martino et al. [30] compiled a refined list of techniques; the techniques are as follows:

- **Loaded language:** Employing words or phrases with strong emotional implications, either positive or negative, to sway an audience [125].
- **Name calling or labeling:** Designating a target of propaganda with terms that evoke strong emotions, be they negative (e.g., fear, hatred) or positive (e.g., love, admiration) [81].
- **Repetition:** Continuously echoing the same message so that the audience will eventually accept it [112, 81].
- **Exaggeration or minimization:** This technique can take two forms. Exaggeration involves excessively portraying something, magnifying its qualities to make it appear larger, better, or worse than it truly is. Minimization, on the other hand, involves diminishing something's importance or size, making it seem less significant or smaller than its actual magnitude. [60].
- **Doubt:** A technique where the credibility of a person or a piece of information is questioned, casting uncertainty and doubt. [30].
- **Appeal to fear/prejudice:** This involves building support for a particular viewpoint by creating anxiety, fear, or panic in the audience. The fear invoked may be towards an alternative perspective or idea, possibly grounded in pre-existing biases or judgments [30].
- **Flag-waving:** A technique that leverages strong sentiments related to nationalism, race, gender, political preference, or other group affiliations to endorse or rationalize a specific action or ideology [53].
- **Causal oversimplification:** A method that simplistically attributes an issue to just one cause when, in reality, there are several underlying reasons [30].
- **Slogans:** These are concise and memorable phrases that often encompass labeling and stereotyping. Slogans tend to serve as emotional appeals. [33].
- **Appeal to authority:** Asserting a claim's veracity solely based on an expert's endorsement, without additional evidence [46].
- **Black-and-white fallacy:** Presenting a limited set of choices, usually two extremes [112].
- **Dictatorship:** telling the audience specific actions to take, eliminating alternatives or choices [112].
- **Thought-terminating cliché:** Using phrases that deter critical thinking or discussion [55].
- **Whataboutism:** A tactic used to challenge an opponent's stance by accusing them of hypocrisy. Instead of directly countering the opponent's argument, it redirects attention by pointing to an alleged inconsistency or failure in the opponent's position [98].
- **Straw man:** Substituting and refuting an argument similar to, but weaker than, the opposition's actual argument [121].
- **Red herring:** Introducing unrelated content to divert attention [125].
- **Bandwagon:** Persuading the audience to join or follow a cause since "everyone else is" [53].
- **Reductio ad hitlerum:** Also known as Guilt By Association, a strategy that seeks to persuade the audience to reject an action or idea by associating it with groups or concepts that are despised or held in contempt by the target audience [111].

Additionally, the technique of **Glittering Generalities (Virtue)** has also been elucidated in literature [41]. This method leverages words or symbols that are deeply embedded in the value system of the target audience. By associating these virtue words or symbols—like peace, hope, happiness, security, wise leadership, freedom, and "The Truth"—with a person or issue, they can generate a positive image.

Moreover, Propaganda techniques often overlap with persuasion techniques. While there are similarities, it's essential to note that their main goals and applications vary [88]. One technique illustrating this intersection is the **Appeal to Time**. This technique, as described by Piskorski et al. [88], argues that the present moment is right for a specific action or change. Originally identified in the context of persuasion, it highlights the calling for acting now. However, the distinction between persuasion and propaganda isn't straightforward. For instance, while the "Appeal to Time" might typically fit a persuasive context, it has also been found in propaganda settings, as shown in further analysis. Similarly, the **Appeal to Values** technique seeks to ground arguments within a group's recognized values and principles. In persuasion, it aligns a proposal with societal values, exemplified by promoting a park due to communal values of health. In propaganda, it can be manipulative, using values to push a hidden agenda, such as justifying censorship in the name of societal harmony. The line between genuine persuasion and propaganda with this technique can be thin, making discernment crucial.

## 2.2 Computational Propaganda

Computational propaganda refers to the use of algorithms, automation, and human curation to manipulate public opinion, usually through the distribution of misinformation and propaganda on digital platforms. The term was coined to encapsulate the increasingly sophisticated use of technologies for political purposes and the threat it poses to democratic processes [127].

The rise of computational propaganda can be attributed to technological advances, particularly in social automation, machine learning, and artificial intelligence. These technologies make it possible to create and distribute persuasive messages on an unprecedented scale, thereby increasing the potential for manipulation and disruption in the political arena [127].

Social media platforms like Twitter and Facebook are prime environments for computational propaganda due to their broad reach and algorithmically driven content distribution. These platforms' design often leads to echo chambers, where users are exposed mainly to opinions that align with their own, thus making them susceptible to misinformation and propaganda [127].

Computational propaganda is being used in a variety of contexts around the world. Examples include its use in influencing elections, manipulating public opinion during times of crisis, and spreading disinformation. The methods employed in computational propaganda are diverse and continually evolving, making it a challenging phenomenon to track and counteract [127].

The study of computational propaganda is crucial for understanding and mitigating its impact on society. It involves various disciplines, including social sciences, data science, computational social science, computer science, and artificial intelligence. Techniques such as network analysis, machine learning, and natural language processing are often used to detect and analyze patterns of computational propaganda on social media platforms [78, 16].

### 2.2.1 Problems of Computational Propaganda

Computational propaganda introduces several critical problems and concerns:

- **Influence on Democratic Processes:** Computational propaganda's potential to manipulate public opinion and disseminate misinformation may affect the integrity of political

discourse and democratic decision-making [116].

- **Integrity of Information:** Automation in disseminating information can blur the distinction between genuine news and fabricated content, possibly eroding trust in media sources [5].
- **Social Polarization:** By reinforcing existing viewpoints and contributing to the formation of echo chambers, computational propaganda may intensify social divisions [109].

### 2.2.2 Challenges in Addressing Computational Propaganda

Tackling the complexities of computational propaganda presents a formidable array of challenges that demand comprehensive and innovative solutions capable of keeping pace with the rapid evolution of computational propaganda techniques. In this section, we delve into the multifaceted nature of these challenges, ranging from the adaptability of propagandists' techniques to the ethical considerations surrounding data privacy and the intricate balance between freedom of speech and misinformation mitigation.

- **Multimodal Mode:** Propaganda is not limited to textual information. Images, videos, and speech often carry more potent messages. The challenge lies in how to integrate these modalities effectively for analysis. Cross-modal information understanding has seen some research, but applying this understanding to detect propaganda remains underexplored. Developing algorithms that interpret such diverse data sources is a critical and unresolved challenge [78].
- **Explainability:** The explainability of a system, especially in controversial decisions like banning accounts or post removal, is crucial for its societal acceptance. Many current approaches leverage deep learning, known for its lack of transparency. Creating models that provide explainable and justifiable decisions without sacrificing effectiveness is a pressing need [78].
- **Automated Propaganda Generation and Neural Language Models:** Modern language models can manipulate text output and even transfer styles between domains. These capabilities mean that automatically-generated propaganda, indistinguishable from human-created content, may soon become prevalent. The challenge is to develop detection methods that can keep pace with these evolving synthetic techniques [78].
- **Deep Fakes and Multimedia Propaganda:** The surge in deep learning has given rise to "deep fakes," a technique that utilizes neural networks to create realistic-looking video footage, making it appear that a particular individual said or did something they did not. Deep fakes present a monumental challenge in the realm of computational propaganda, as they can potentially spread misinformation and propaganda with a high degree of believability [92]. The implications of this technology in the sphere of misinformation and propaganda are profound, requiring not only technological solutions for detection but also heightened public awareness and education.
- **Understanding Strategy and Goals of Propaganda Campaigns:** While there has been significant progress in detecting and classifying propaganda campaigns, much less has been done to comprehend the underlying strategies and goals of propaganda campaigns. Understanding these aspects is not only essential for a more profound insight into how propaganda functions but also for developing targeted interventions and countermeasures. Without grasping the overarching objectives, merely detecting individual instances may

fall short of addressing the problem comprehensively. This challenge calls for a broader, interdisciplinary approach that synergizes efforts in both detection and strategic understanding of propaganda. The current research landscape in this area is fragmented, and cohesive methodologies for analyzing strategies are largely unsolved [78].

- **Generalization and Dataset Limitations:** Many detectors are evaluated on specific datasets, limiting their broader applicability. Future work must focus on curating large, diverse datasets to test generalizability [78].
- **Ethical Considerations in User Data:** Privacy and ethical concerns can limit data availability, posing challenges in implementing robust privacy-preserving mechanisms [78].
- **Labeling and Bias in Annotation:** The hidden nature of sources and purposes complicates the task of identifying intent in propaganda, demanding intricate analysis and judgment [59]. Subjectivity in labeling can introduce biases, and evolving propaganda techniques impede static definitions and consistent categorization [78].
- **Adaptation by Propagandists:** The propagandists' constant tactic evolution requires continuous research and innovation to maintain detection effectiveness [78].
- **Platform Regulation:** As of 2023, major platforms such as Twitter have halted their Academic API, introducing significant barriers to research on social media. Such changes in data accessibility have implications not only for computational propaganda research but also for the reproducibility of past studies, considering these platforms are central to the largely unregulated spread of propaganda<sup>1</sup>. On the other hand, the European Union's Digital Services Act (DSA) represents a step forward in addressing platform accountability and transparency. It mandates a more structured and open approach to content moderation and data sharing.

Navigating the outlined challenges, the research landscape primarily divides into two core areas: the textual analysis perspective, examining linguistic intricacies of propaganda, and the network analysis perspective, focusing on its propagation patterns. In the subsequent sections, we will delve into these two approaches to understanding and combating computational propaganda.

### 2.2.3 Textual Analysis Perspective

Textual analysis provides a critical lens to unravel the intricacies of digital propaganda. Charting from early computational studies to refined technique-based research, this perspective offers insights into the evolving landscape of text-based propaganda detection.

**Early Efforts** Early computational studies, such as Rashkin et al. [94], addressed propaganda detection at the document level. Their approach classified texts into four categories: trusted, satire, hoax, and propaganda using distant supervision. Expanding on this, Barrón-Cedeno et al.[9] investigated writing styles and readability as metrics to discern propaganda from non-propaganda. Their results indicated that enhanced representations combined with distant supervision can determine the source of an article effectively.

---

<sup>1</sup>"The changes to the Twitter API are having catastrophic effects on our research into the spread of disinformation and its harms, the manipulation of social media, and the vulnerability of people and platforms to online abuse," said Filippo Menczer, director of the Observatory on Social Media at Indiana University.

"Twitter is making researchers delete data it gave them unless they pay \$42,000", Independent, May 25, 2023.  
<https://inews.co.uk/news/twitter-researchers-delete-data-unless-pay-2364535>

**Detailed Propaganda Analysis** Da San Martino et al. [32] introduced a detailed approach by developing a dataset annotated with 18 propaganda techniques. Their methodology divided the tasks into detection and classification of technique spans, with the application of a multi-granular gated deep neural network. This dataset facilitated further research, leading to the Prta system’s release [31], methods to address transformer limitations in propaganda detection [24], and initial investigations into interpretable propaganda detection [131].

**Shared Tasks in Propaganda Detection** The EMNLP-IJCNLP 2019 conference organized a shared task on Fine-Grained Propaganda Detection, emphasizing Sentence Level Classification (SLC) and Fragment Level Classification (FLC). The adopted methodologies incorporated ensemble attention neural networks utilizing BERT, tailored BERT models, and hybrid approaches integrating XGboost, BiLSTM, and BERT [29]. Subsequent shared tasks, such as SemEval-2020’s Task 11 and SemEval-2021’s Task 6, furthered this research, with the latter emphasizing both text and image propaganda detection [30]. The WANLP’2022 task concentrated on Arabic tweets [95]. These tasks showcased advancements in strategies like the use of pre-trained models, ensemble techniques, transfer learning, and data augmentation. Notably, SemEval-2023’s Task 3 [89] expanded the scope to multilingual detection across various article genres.

#### 2.2.4 Network Analysis Perspective

Computational propaganda is an evolving field that necessitates a thorough analysis of both content and network behavior to ensure accurate detection and mitigation. Aspects such as textual rhetoric play a role, but understanding the network dynamics—specifically, malicious coordination—is essential [78].

**Early Approaches** Initially detecting malicious coordination focused on classifying individual nodes within a network as either malicious or not. Post-classification, clusters of these flagged nodes were then assessed for potential coordinated behavior. These methodologies did not inherently incorporate the concept of coordination within their detection paradigms but rather appended it as an afterthought. Most techniques during this phase were predominantly grounded in supervised machine learning. The detection process treated each account in isolation. The foundational presumption here was that malicious accounts exhibit discernible features that markedly differentiate them from legitimate counterparts. Classic machine learning models, encompassing decision trees, random forests, SVMs, ensemble methods like boosting and bagging, and with time, deep neural networks were harnessed for this purpose [68, 78].

*Botometer* [102] stands as a testament to this approach. By harnessing over 1,200 features of a social media account, it evaluates diverse dimensions ranging from profile attributes to temporal behaviors. However, it’s essential to note that while some systems, such as Botometer, adopted a holistic approach, others confined their focus to merely network attributes [130], text content [93], or user profiles [70]. Such mono-dimensional detectors, despite their simplicity, are more susceptible to manipulation due to their narrow lens of analysis [78].

**Evolving Challenges** The preliminary success of these methodologies was soon overshadowed by intrinsic challenges. Foremost among these was the dependency on accurate training datasets. More often than not, genuine datasets were unattainable, leading to an over-reliance on human annotators. This was problematic due to the absence of a universally accepted definition of “malicious accounts” [47], compounded by the inherent biases and difficulties human annotators face in identifying sophisticated malicious entities [27].

The evolution of *Botometer* is a prime example of the adaptive nature of computational propaganda detection. Initially released as BotOrNot[35], it has undergone several refinements to become Botometer V4[102], its current iteration. This evolution highlights the dynamic challenges of combating computational propaganda. Botometer V1 (BotOrNot) relied on limited features and models to detect a user’s “bot-like” behavior. However, as malicious actors became more sophisticated, Botometer had evolved as well. V2 [118] introduced a more comprehensive feature and training set and leveraged Random Forest classifiers. The advent of Botometer V3 [129] incorporated deep learning, significantly boosting its accuracy and detection capabilities. By the time V4 was launched, it employed an ensemble of machine learning algorithms and assimilated a vast range of training sets and engineered features reflecting user content, sentiment, network patterns, and temporal behaviors [128].

**Modern Approaches** Acknowledging the shortcomings of early techniques paved the way for methodologies emphasizing group-centric analysis over individual account scrutiny. Contemporary detectors internalized the concept of coordination, recognizing that malicious entities often function in tandem to amplify their reach and impact, as seen in botnets or troll armies [133].

This shift from individual to group-centric analysis also witnessed a transition from traditional machine learning models to bespoke algorithms, especially tailored for coordination detection. The emphasis transitioned from merely engineering features to crafting effective feature representations and algorithms [18]. Unsupervised and semi-supervised approaches gained prominence to counter the limitations posed by the dearth of exhaustive training datasets [40].

One popular strand of modern systems harnesses network-based techniques. Their primary objective is identifying suspicious connectivity patterns among accounts [73, 25, 87, 4]. In this context, coordinated behaviors manifest as closely interconnected communities in graphs, dense clusters in adjacency matrices, or distinctive patterns in spectral subspaces [58].

Other advanced techniques adopted unsupervised methodologies to detect anomalies in account groups’ temporal tweeting and retweeting behaviors. The foundation of such methods rests on the observation that genuine human-driven behaviors are inherently more varied than their automated, inauthentic counterparts [28]. Therefore, if there’s a substantial group of accounts displaying highly similar behavior, it could raise alarms for coordinated inauthentic actions. To measure the distance (or similarity) between activity time series of accounts, techniques like dynamic time warping [23] or the Euclidean distance calculation between feature vectors outputted by LSTM autoencoders have been employed [80].

One crucial observation from the evolution of detection techniques is the reactive nature of the development cycle. Often, advancements materialize in response to identifying new malicious behaviors, offering adversaries a temporal advantage. Additionally, it’s pertinent to understand that machine learning algorithms designed for static and neutral environments may falter when introduced to the dynamic and adversarial landscape of propaganda detection [78].

### 2.2.5 Future Outlook

The computational landscape of propaganda detection is poised for notable advancements, shaped by the nuanced techniques of propaganda dissemination and the technological strides in its detection as computational propaganda evolves. Drawing from the challenges discussed, several pertinent trajectories in the realm of propaganda research and counter-strategies can be discerned.

Da San Martino et al. [78] emphasize an emergent requirement for an amalgamated approach, one that combines both textual and network analysis approaches for efficacious propaganda detection. The advent of refined tools, such as GPT, which can adeptly craft propaganda,

underscores the vulnerability of depending exclusively on linguistic attributes. Therefore, a more integrative strategy, which delves into the network dynamics underlying content distribution, is imperative. This strategy would necessitate a focus on both the content and its dissemination channels, posing a formidable challenge to covert propagandist activities.

Furthermore, the shift towards computational propaganda is especially highlighted by Da San Martino et al. [78]. With a growing inclination among the millennial generation to consume information through visuals rather than text, propagandists are capitalizing on this trend. This transition underscores the urgent need for research to pivot from being primarily text-focused to adopting a multimodal perspective, enabling a thorough analysis across diverse data modalities. One of the most concerning developments in computational propaganda is the rise of deep fakes. Using advanced artificial neural networks, propagandists can now generate hyper-realistic but entirely fictitious audiovisual content. These deep fakes can potentially deceive viewers, distort facts, and manipulate public opinion at an unprecedented scale. Thus, a critical trajectory for future research would be the development of robust deep fake detection mechanisms with the aim of detecting computational propaganda.

Another salient trajectory for future endeavors is the confluence of technology and public education. An informed populace holds intrinsic value in this battle. Hence, future research could strategically leverage technology to apprise and educate the general public about computational propaganda, equipping them with the tools to assess, learn and counter such content critically.

## Chapter 3

# Digital Dissidence: The Women Life Freedom Movement in Iran

The interplay of technology and activism has never been more poignant than in Iran's current political landscape [63]. As digital platforms reshape societal discourses, the country's activism, particularly concerning women's rights, has witnessed a revolutionary transformation [63, 64, 108, 110]. This chapter delves into the nexus of grassroots movements, digital activism, and the broader Iranian historical context to unearth the evolving nature of women's struggles against systemic challenges. Through an in-depth exploration, we underscore how digital spaces, especially Twitter, have become both a bastion for silenced voices and a battleground for political ideologies [66, 62, 7]. Drawing upon seminal events, movements, and political shifts, the chapter sheds light on the complexities and potentials of digital dissidence within the Iranian milieu.

### 3.1 Grassroots Movements and Social Media Activism

Grassroots social movements are forms of collective action that emerge organically within communities, driven by shared concerns or common goals. These movements leverage the power of collective voice to advocate for change, often operating outside traditional political systems or institutional frameworks [39]. The advent of social media has revolutionized the way grassroots movements organize and operate. Through social media platforms, activists can efficiently communicate, mobilize, and coordinate efforts on a global scale [10]. These digital tools allow movements to transcend geographical boundaries and reach a wider audience than traditional media channels, providing new opportunities for engagement and advocacy.

Social media platforms also provide activists with unique means of circumventing traditional power structures, particularly in contexts where freedom of speech is limited or suppressed [117]. In such scenarios, grassroots movements can use social media to make their voices heard, challenge dominant narratives, and push for change. The crucial role of social media in these movements highlights the emergence of a 'networked public sphere'. Conversations that were once confined to private spaces are now taking place in public digital arenas, increasing visibility and attracting international attention [21]. This shift provides a valuable lens for understanding power dynamics in online activism, contributing to broader insights into digital resistance strategies under authoritarian regimes [117].

## 3.2 Iran's Historical Context

**Pre-1979: The Pahlavi Dynasty's Governance** From the early 20th century until 1979, Iran was governed by the Pahlavi dynasty. While a constitutional monarchy was formalized in 1906, the monarch's authority expanded notably during the tenure of the Pahlavi family, especially under Mohammad Reza Shah Pahlavi. His regime was characterized by its push for modernization and Western-inspired reforms. One significant policy from this era, introduced by Reza Shah Pahlavi, was the 1930s banning hijab as part of a broader secularization initiative. Reza Pahlavi is the final crown prince of the Pahlavi dynasty. [2]

**1979 Islamic Revolution and the Establishment of the Islamic Republic** The Islamic Revolution in 1979, driven by broad-based discontent. The banning of hijab, rapid Westernization, and the Shah's secularist policies generated religious and cultural tensions. Politically, repression was pervasive, highlighted by the SAVAK's (Iran's secret police) brutal operations. Additionally, foreign interventions, exemplified by the 1953 U.S.-UK orchestrated coup overthrowing the democratic government of Iran, and persisting economic disparities despite vast oil revenues, fostered feelings of economic and national disenfranchisement. This societal unrest culminated in the ouster of Mohammad Reza Shah and the establishment of the Islamic Republic of Iran. Central to the political philosophy of this new system is the "Velayat-e Faqih" (ولایت فقیه) concept, which translates to "Guardianship of the Islamic Jurist." This principle allocates significant religious and political authority to the Supreme Leader, a position first held by Ayatollah Ruhollah Khomeini and subsequently by Ayatollah Ali Khamenei. The foundational ideology of the Islamic Republic is deeply rooted in Shia fundamentalism, emphasizing the creation of a nation prepared for the return of the Mahdi, the prophesized Shia saviour. [79, 2, 37]

**Mujahedin-e Khalq** The Mujahedin-e Khalq (MEK) is an Iranian opposition group with a complex history of ideological evolution and shifting allegiances. Founded in the 1960s, the MEK initially combined Marxist and Islamist ideologies in its opposition to the Pahlavi monarchy. The organization played a role in the 1979 Iranian Revolution but quickly fell out of favor with the emerging Islamic Republic under Ayatollah Khomeini, leading to years of persecution and exile. The group, led by Massoud and Maryam Rajavi, relocated its headquarters to France and later to Iraq, where it found support under Saddam Hussein during the Iran-Iraq War. This alliance notably impacted the MEK's reputation inside Iran. In the early 2000s, the group disarmed and began presenting itself as a democratic alternative to the Islamic Republic. This culminated in a successful lobbying effort to be removed from terrorist lists in the European Union and the U.S. by 2012. However, the MEK remains controversial, with critics questioning its cult-like attributes, human rights records, and genuine commitment to democracy. [100, 1]

**Flight PS752 Incident** On January 3, 2020, American drones targeted Baghdad International Airport, leading to the assassination of Qasem Soleimani, an Iranian major general in the Islamic Revolutionary Guard Corps, and Abu Mahdi al-Muhandis, the secretary-general of Kata'ib Hezbollah, an Iraqi militia. This significant event escalated tensions in the region. In retaliation, on January 8th, Iran initiated a ballistic missile attack named "Martyr Soleimani" against the American Ayn Al-Asad Airbase in Erbil, Iraq. The attack, which commenced at 1:34 a.m. local time and concluded by 4:00 a.m., was reportedly communicated in advance through Iraqi intelligence and other undisclosed channels. Despite the scale of this operation, there were no casualties reported from this strike. However, in a tragic misstep in the heightened security environment, Ukrainian International Airlines Flight PS752, en route from Tehran to Kiev, was caught in the crossfire. Taking off from Imam Khomeini International Airport at 06:12 local

time on January 8th, the plane was struck by a Tor M1 surface-to-air missile (SAM) launched by the IRGC. This devastating incident led to the unfortunate death of all 176 people on board, comprising 167 passengers and 9 crew members. [91]

### 3.3 Women and Feminist Digital Activism in the Iranian Context

The condition of women in Iran represents a complex interaction of legal discrimination, cultural norms, and a persistent spirit of resistance that transcends what might be observed in Western democracies, where women's rights are generally protected by law [83]. The situation in Iran is more profound than merely victim-blaming; it involves fundamental, systemic challenges that intertwine with cultural, ethnic, and political tensions [64].

Legal structures in Iran inherently favor men in sectors such as education, economic participation, and even basic human rights [83]. This legal bias extends to a failure to support victims of sexual assault, leading to further shame and lack of legal recourse [64]. The entrenched political system, coupled with specific policy initiatives like mandatory hijab enforcement, solidifies and perpetuates sexist views [83].

In 2014, Masih Alinejad, an Iranian-born journalist and activist, started the "My Stealthy Freedom" movement, inviting Iranian women to send unveiled photographs of themselves to protest the compulsory hijab [108]. Building on its success, Alinejad launched the #WhiteWednesdays movement on Twitter, Facebook, and later Instagram. This movement brought together women wearing white on Wednesdays, uniting them in a shared demonstration against compulsory hijab laws [110].

In 2018, concurrent with the global prominence of the #MeToo movement, which demanded accountability for instances of sexual harassment and assault, Iran's digital feminist activism underwent a notable shift. Kermani and Hooman [64] analyzed hashtag feminism and the #rape movement on Persian Twitter, uncovering how it served as a space for Iranian users to share abusive experiences. The authors found that Iranian Twitter users not only shared experiences of abuse, aligning with global narratives but also illuminated the distinct structural barriers they faced. These included the country's restrictive political system and prevailing oppressive discourses, which serve to perpetuate rape culture. Their study showcased the transformative power of the movement; once-taboo topics of sexual abuse began to take center stage in digital feminist discourse in Iran. Moreover, these narratives transcended cisgender women, giving voice to other marginalized groups, like the LGBTQIA+ community. Yet, this movement faced multifaceted challenges. State actors, and surprisingly, some anti-regime entities such as Monarchists and MEK, sought to disrupt its momentum. The authors underscore the significance of understanding online feminist activism within restrictive settings and the double-edged role of social media platforms—both empowering and, at times, oppressive.

In 2022, The Women Life Freedom or #MahsaAmini movement, sparked by the tragic death of Mahsa Amini while in the custody of the morality police, evolved to critique not just gender-based violence, but an entire spectrum of sociopolitical norms and structures in Iran [34, 63]. Using social media, especially Twitter, this movement ignited protests across Iran against mandatory hijab and broader authoritarian practices [63].

These feminist movements exemplify a convergence of cultural, political, and technological dynamics [34, 63]. Moreover, they have energized the scope of digital feminism in Iran, not only challenging the pervasive anti-women culture but actively seeking to reshape the national discourse. With the support of diaspora involvement, political allies, and continuous efforts from women and feminist activists, social media, particularly Twitter, has emerged as an essential

space in Iran's political landscape. This transformation has turned Twitter into both a platform for empowerment and a contested battleground, reflecting the complexity of Iran's ongoing sociocultural transformation [66, 64, 63].

### 3.4 Political Field on Iranian Twittersphere

Twitter's role in Iran presents a unique platform for political discourse, distinguishing itself from the political environment of Western democracies. In Iran, where severe constraints are placed on traditional media outlets, Twitter becomes the main channel for political engagement [106, 65]. Unlike Western societies where social media is a choice, for Iranians, it is a necessity. The stringency of media censorship forces political parties and figures to rely on Twitter to express their ideas, making Iranian Twittersphere a more realistic portrayal of Iran's political landscape, with no voices removed. [65]

Iran's political system, primarily divided between reformist and conservative parties, saw both sides turn to Twitter, turning the platform into a political battleground [66, 62, 7]. The 2017 election marked a particular moment where conservative, reformist, and diaspora political figures actively operated accounts to dominate their ideologies [7].

However, recent shifts after the Women.Life.Freedom movement, underscored by the "Iranians' Attitudes toward Political Systems" survey [76] conducted in 2022, have significantly altered the political field on Twitter in Iran. The discourse has evolved, and the focus is now on the different potential governing structures - "secular republic", "constitutional monarchy", and "Islamic republic". This shift reflects the broader societal aspirations, as 88% of the population considers having a democratic political system to be positive, and a large portion of respondents are against a system governed by religious law or having the army rule.

The Iranian diaspora, secular democrats, reformists, conservatives, and newly vocal proponents of constitutional monarchy utilize Twitter to advocate for their preferred regime types, making Persian Twitter an even more complex and contested space [7].

## Chapter 4

# Dataset: Iranian Twittersphere

The rapid digitization and growth of social media platforms have facilitated the emergence of various socio-political movements worldwide [21]. Among these platforms, Twitter stands out due to its microblogging nature, enabling users to express opinions, mobilize support, and document live events [57]. The Women Life Freedom movement, which burgeoned in Iran, is one such example of how socio-political sentiments manifest on digital platforms [63]. This chapter delves into the dataset compiled from Twitter, capturing tweets related to this movement. This chapter presents a comprehensive understanding of the data collection methodology, the reasoning behind choosing specific search queries, and an overview of the results that provide a snapshot of the movement's digital footprint.

### 4.1 Data Collection

This study's primary data source is Twitter, focusing on Women Life Freedom movement. The data was collected using `snsrape`<sup>1</sup>, a Python library that scrapes Twitter data without the need for API access. This approach was chosen because Twitter recently discontinued its Academic API, which previously allowed researchers to access historical and real-time Twitter data.

For data collection, tweets containing specific Persian-language search queries were scraped to identify a relevant subset of Twitter users engaged in or affected by the movement. The search queries are “زن زندگی آزادی” (translated as “Women Life Freedom”), “مرد میهن آبادی” (translated as “Men Homeland Prosperity”), and “مهسا امینی” (“Mahsa Amini”). These queries were employed as they are emblematic of the lexicon recurrently used in discussions and propaganda pertinent to the movement. From March 24, 2023, to April 1, 2023, these specific search queries were used to scrape tweets. During this period, any tweets posted after September 2022 that contained these queries were downloaded. This scraping process resulted in identifying a pool of 464,712 users. The complete tweets of these users were collected between the 1st of April to the 17th of April, 2023. Subsequently, the dataset contains posts from September 25, 2006, to April 17, 2023.

The rationale behind selecting these specific queries for gathering a user pool is based on the understanding that these terms are integral to the discourse surrounding the Women.Life.Freedom movement and are employed by a diverse array of Twitter users. Notably, while activists and supporters of the movement use these terms, they are also employed by other entities, such as state-associated propagandists and various political factions. These groups often use these terms to interact with, influence, or respond to the movement. By using these terms as queries, the

---

<sup>1</sup><https://github.com/JustAnotherArchivist/snsrape>

data collection process can capture a representative pool of Twitter users actively engaging with the movement or its themes, either in support or otherwise. This enables the study to include a wide spectrum of perspectives and dynamics that reflect the complex environment in which the Women.Life.Freedom movement operates.

Unfortunately, the dataset could not be updated beyond April 2023 as snscreape also ceased functioning. Furthermore, due to the limitations of snscreape and the unavailability of Twitter's Academic API, the follower, retweets, and like data could not be collected. This represents a significant limitation in the data collection process, affecting the ability to fully understand the dissemination of information and the structure of interactions among the Twitter users involved in the movement.

## 4.2 Data Characteristics

The dataset for this study includes **464,712** users and **1,006,786,621** tweets. Several visualizations provide insights into the dataset, such as user account creation dates, and tweet volumes from users. A pronounced spike in user account creations in September 2022, as depicted in Figure 4.1, strongly suggests that the Mahsa Amini movement catalyzed significant interest and subsequent user registrations.

When analyzing the distributions of user tweet counts from Figure 4.2, a noteworthy observation emerges. Although the distributions appear similar pre- and post-Mahsa, the median tweet counts tell a more nuanced story. Users created before the Mahsa Amini movement have a median tweet count of 539, whereas the users who joined post-movement have a median of 91 in only seven months. Proportionally, this suggests that post-Mahsa users are exhibiting higher activity levels on the platform when compared to the pre-Mahsa period.

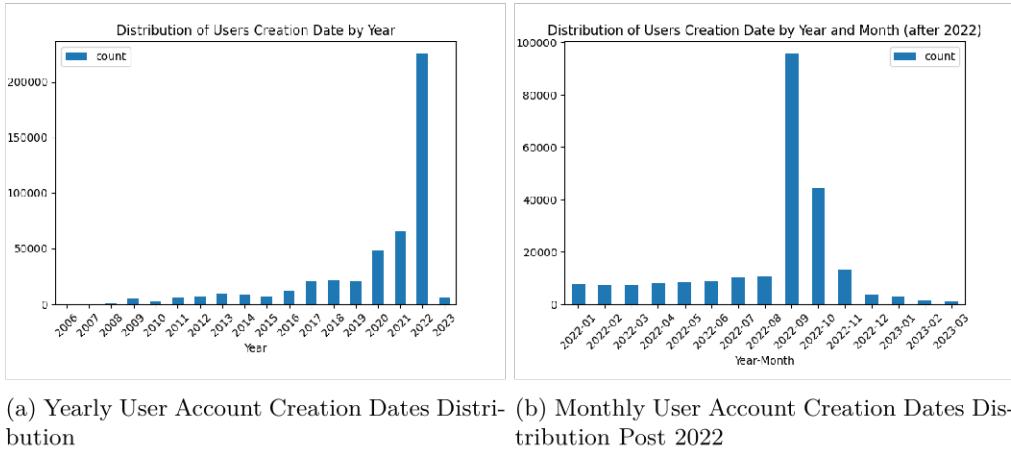


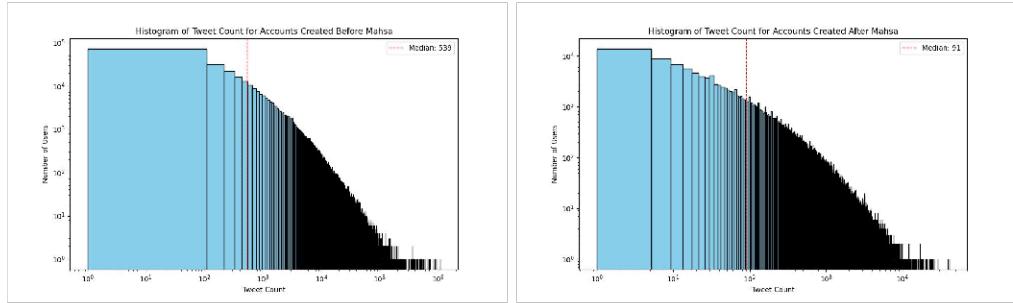
Figure 4.1: User Account Creation Date Distribution

### Top Hashtags

Figure 4.3 provides a graphical representation of the top 25 most frequently used hashtags in the time span from January 1, 2022, to April 1, 2023. A more detailed exploration is offered in Table 4.1, where the top 25 hashtags from this period are listed alongside their respective counts and explanations.

Hashtag	Count	User Count	Translation/Explanation
(Merged) مسا_امینی MahsaAmini Mahsa_Amini مهساامینی mahsaamini	<b>91,576,322</b> 73,373,911 13,588,558 3,750,046 951,482 869,469	<b>419,027</b> 409,504 261,031 170,698 127,641 59,183	Mahsa Jina Amini, a 22-year-old Iranian Kurdish woman who died after arrest by Islamic Republic's morality police. Known as the symbol of Women Life Freedom movement.
(Merged) OpIran opiran Oplran	<b>21,560,913</b> 19,735,970 1,140,939 681,066	<b>237,002</b> 226,159 76,767 74,338	Symbol of Anonymous support for Iran's protests. The second and third hashtags use the letter 'l' instead of 'I', a strategic hashtag inserted to dilute the movement's online presence.
(Merged) اعتصابات سراسری اعتزازات سراسری زن زندگی آزادی	<b>15,632,963</b> 14,441,752 1,089,092	<b>221,025</b> 210,923 76,350	Nationwide Strikes/Protests
نیکا شاکرامی	1,579,093	158,129	Women Life Freedom
IRGCterrorists	3,278,086	147,280	Nika Shakarami: 16-year-old girl who disappeared during the 2022 protests after Mahsa Amini's death and was later found dead.
محسن_شکاری	6,624,805	103,668	IRGC Terrorists, referring to the Islamic Revolutionary Guard Corps.
حسین رونقی	668,122	83,301	Mohsen Shekari was executed by the Islamic Republic after being found guilty of Moharebeh, meaning "waging in war against God." Shekari's execution holds significance as it's believed to be the first state-sanctioned execution directly linked to the 2022 Mahsa Amini demonstrations.
توماج صالحی	830,647	72,785	Hossein Ronaghi: Iranian human rights activist and blogger. He was imprisoned multiple times for his activities before and became a prominent face during the Mahsa Amini protests after his re-arrest in 2022.
اشنویه	1,390,678	72,414	Tomaj Salehi: Iranian hip hop artist known for protest songs. He was sentenced to prison for participating in the 2022 Iran protests and arrested during Mahsa Amini protests.
IranRevolution	1,779,567	64,980	Oshnavieh: City in West Azerbaijan, Iran. Protesters briefly took over the city during the Mahsa Amini protests, from 24 to 25 September 2022.
کریم تو بیت رهبری	813,624	49,679	Iran Revolution
Iran	574,537	45,100	Insult to Leadership (Khamenei)
من_و_کالت_میدهم	553,858	36,508	Refers to the country of Iran.
StopHazaraGenocide	1,225,247	19,730	Translates to I Give Attorneyship. Refers to a petition on change.org for declaring Reza Pahlavi as their representative.
لیک_یا_خامنه_ای	944,117	17,556	Campaign to draw attention to the persecution of the Hazara people, an ethnic minority in Afghanistan.
جاویدشاه	874,987	15,429	At Your Service, Khamenei, showing support or allegiance to Khamenei.
			Everlasting King: Referring to the Pahlavi dynasty and monarchist sentiments

Table 4.1: The top 25 most frequently used hashtags between January 1, 2022, and April 1, 2023, with translations and explanations. Similar hashtags are grouped together.



(a) Tweets Distribution of Accounts Created Before Mahsa (before September 14, 2022)      (b) Tweets Distribution of Accounts Created After Mahsa (after September 14, 2022)

Figure 4.2: Distribution of Users' Tweet Counts Before and After Mahsa. Both x and y axes are set to log scale.

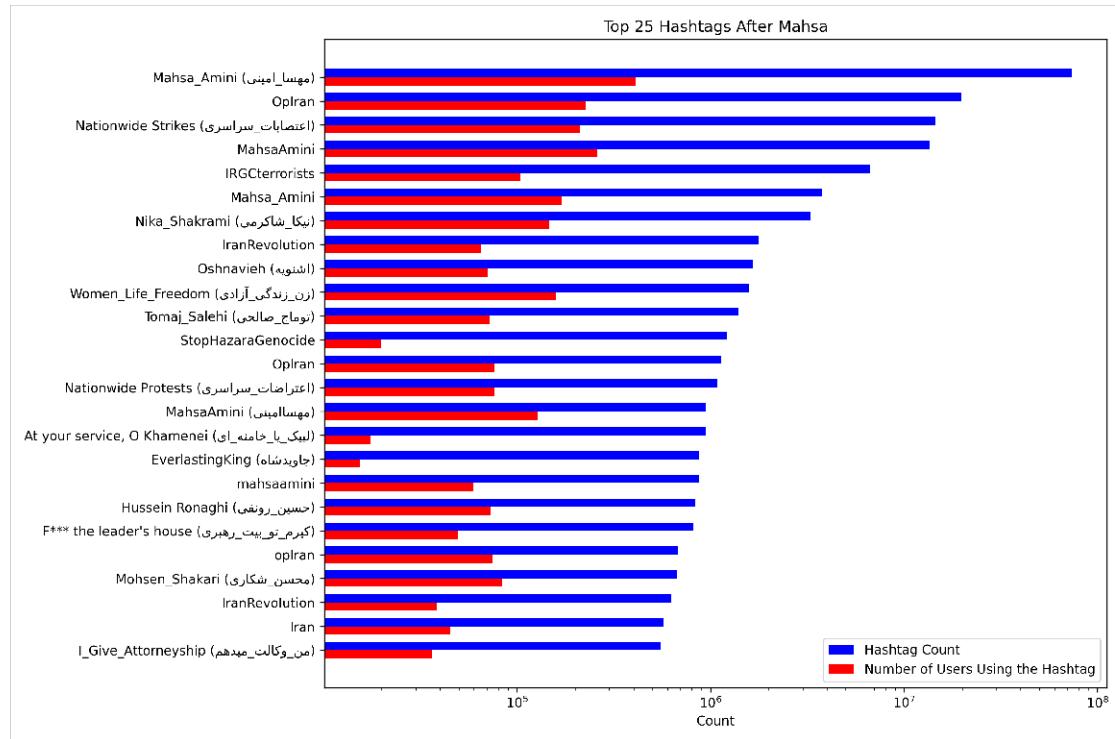


Figure 4.3: Frequency distribution of hashtag total usage count and number of users using the hashtag for the top 25 most frequently used hashtags from September 14, 2022, to April 16, 2023. The x-axis is set to log scale.

# Chapter 5

## Propaganda Analysis: Case Study of Women Life Freedom Movement

This chapter delves into an analysis of the Women Life Freedom Movement within the Iranian Tweetersphere, highlighting the intricate interplay of propaganda. The chapter begins by exploring the underlying technologies and a survey of related academic literature. The subsequent sections detail the research methodology, spotlighting the identification of key actors, the mapping of propaganda networks, and the techniques employed. The chapter culminates in evaluating the impact these entities exert on public discourse, underscoring the significance of digital propaganda in shaping socio-political narratives.

### 5.1 Background Information and Related Work

This section elucidates the core methodologies in topic analysis and social network analysis (SNA). A primary focus is on topic analysis techniques, encompassing both statistical modeling and direct thematic identification from textual datasets. The discussion subsequently shifts to SNA, emphasizing the application of graph theory to discern patterns in social structures. Through an examination of key scholarly contributions, practical applications of these methodologies are highlighted.

#### 5.1.1 Topic Analysis

Topic analysis encompasses a range of techniques in text mining and natural language processing (NLP) aimed at uncovering the primary themes or subjects within a collection of documents. Its overarching goal is to determine patterns or clusters of words that hint at underlying topics, providing insight into the contents of the dataset [77].

##### Statistical Topic Modeling

One of the key strategies in topic analysis involves the use of statistical models. The most notable of these is the Latent Dirichlet Allocation (LDA) [12]. LDA operates on the premise that each document in a corpus encompasses a mixture of topics, and each topic is a distribution

over words. Iteratively, LDA fine-tunes word distributions for each topic and the proportion of topics across documents, thereby illuminating concealed themes. Moreover, Non-Negative Matrix Factorization (NMF) and Latent Semantic Analysis (LSA) also form part of the toolkit for topic extraction [69, 36]. While LSA uses singular value decomposition on term-document matrices, NMF deconstructs high-dimensional vectors into non-negative data representations, offering varied angles on the data.

### Direct Topic Identification

While statistical models offer a probabilistic approach, direct topic identification is more concrete. It zeroes in on specific themes by recognizing discernible patterns such as the frequency of terms or phrases. This method shifts the focus from the probabilistic distributions of LDA to clear thematic and statistical indicators [77].

**N-Grams** N-grams are contiguous sequences of  $n$  words or characters from a text. This ability to encapsulate local context makes N-grams particularly valuable for topic identification, as the combination of words often provides more nuanced insights than individual words [77]. For instance:

- 1-gram (unigram): “apple”
- 2-gram (bigram): “red apple”
- 3-gram (trigram): “rotten red apple”

Quantitatively counting N-gram occurrences in a corpus aids in discerning frequently discussed themes. High-frequency N-grams can often highlight main topics or sub-topics. In relation to interpreting statistical topic modeling, sentiment, or hate speech analysis, analyzing prevalent N-grams in clusters can aid in better understanding the core themes of each topic or label. Furthermore, N-grams have multifarious applications in NLP, including feature engineering, predictive typing, spell-checking, and plagiarism detection [15, 77].

**Term Frequency-Inverse Document Frequency (TF-IDF)** TF-IDF calculates the importance of a term within a document relative to a corpus. It emphasizes terms that are frequent in a document but rare across the corpus. The TF-IDF score,  $\text{TF-IDF}(t, d, D)$ , for a term is:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

Where  $\text{TF}(t, d)$  is the term’s frequency in the document and  $\text{IDF}(t, D)$  is its inverse document frequency, calculated as:

$$\text{IDF}(t, D) = \log \left( \frac{N}{1 + \text{df}(t, D)} \right)$$

Here,  $N$  represents the total number of documents, and  $\text{df}(t, D)$  is the number of documents containing term  $t$ . This method effectively filters out common terms that don’t differentiate the content [77].

**Lemmatization** Lemmatization, in NLP, reduces words to their base or dictionary form. Unlike stemming, it considers word morphology, using a vocabulary and morphological analysis to achieve this conversion [77]. For example, “running”, “runs”, and “ran” are all lemmatized to “run”.

**Preprocessing Pipeline for Topic Identification** A robust preprocessing pipeline is indispensable for effective topic identification. Refining the data and ensuring uniformity, it sets the stage for accurate and meaningful analysis. A typical pipeline for this may include: [22, 77]:

1. **Stop Word Removal:** Eliminating common words that usually don't convey significant meaning such as "be", "is", etc.
2. **Lemmatization:** Words are reduced to their base form.
3. **N-Gram:** N-grams are generated from lemmatized words.
4. **TF-IDF Vectorization with N-grams:** Post lemmatization, TF-IDF can be applied on N-grams to identify the most important n-grams in a document.

### 5.1.2 Social Network Analysis (SNA)

Social network analysis is an interdisciplinary analytical approach utilized for examining social structures via networks and graph theory. [124] Social networks consist of nodes (actors or entities) connected by edges (relationships or interactions). SNA provides both a visual and mathematical analysis of human relationships and has become a vital tool in fields such as sociology, economics, psychology, and computer science. SNA measures and maps the relationships and flows between entities, positions, and affiliations. It aims to understand networks of relations as a whole (sociocentric) or around focal entities (egocentric). Essential metrics in SNA include centrality measures, such as degree centrality, betweenness centrality, and closeness centrality, which provide insights into the most influential nodes within a network. [124, 43]

#### Community Detection

One key aspect of social network analysis is the detection of communities or clusters within a network. These communities represent groups of nodes that are more densely connected with each other than with the rest of the network. Identifying these communities can reveal significant structures and hidden patterns in social networks, leading to a better understanding of complex network dynamics. [85, 43, 99]

The Girvan-Newman method is a hierarchical approach to community detection developed by Michelle Girvan and Mark Newman [45]. It identifies communities by progressively removing edges from the network based on their "betweenness centrality," a measure of their importance in connecting different parts of the network. Specifically, edges that are least central (i.e., least frequently part of the shortest paths between node pairs) and thus most likely to lie between communities, are removed first. This removal process continues iteratively until the network splits into distinct communities. A drawback of the Girvan-Newman method is its computational intensity, especially for large networks. Despite this, it offers valuable hierarchical insights into community structures, revealing both larger clusters and more refined sub-communities.

Furthermore, the Louvain method is a popular heuristic for community detection that relies on the optimization of modularity, a measure quantifying the density of links inside communities compared to links between communities. The algorithm works iteratively, first assigning each node to its community and then moving nodes between communities to optimize modularity. Despite its efficiency and effectiveness on large networks, the Louvain method is sensitive to small changes and has a resolution limit, which can impede the detection of smaller communities [13].

In response to some limitations of the Louvain method, the Leiden algorithm was developed [115]. It improves upon the Louvain method by incorporating a refinement phase, which allows for the movement of individual nodes to enhance the quality of partitions. This inclusion

leads to higher modularity and more stable community detection. Furthermore, the Leiden algorithm has been shown to be robust against the resolution limit problem that hampers the Louvain method. These improvements make the Leiden algorithm an attractive choice for community detection in social network analysis.

### 5.1.3 Related Work in SNA

Gargiulo et al. [44] analyzed Twitter activity related to “Occupy Wall Street” using a network-centric methodology. They formed a bipartite graph of Users-#Hashtags and examined its two major projections: the “semantic network”, with hashtags as nodes, and the “users interest network”, representing users as nodes. They used the Louvain community detection algorithm on the “semantic network” to find semantic communities. Their findings indicate that the discussion topics, represented by semantic networks (#hashtags), showed notable structural heterogeneity.

Conover et al. [26] examined political polarization on Twitter by analyzing the communities detected between mentions and retweets networks. Using a combination of network clustering algorithms and manually-annotated data, their findings revealed a polarization within the retweet network, whereas the mention network displayed more heterogeneity. This divergence was partly attributed to using neutral or mixed-valence hashtags, facilitating interactions across ideological boundaries. However, despite these cross-ideological engagements, the retweet network reflected persistent political segregation, indicating limited dissemination of diverse views within users’ core communities.

Guarino et al. [49] investigated propaganda networks on Twitter during the 2016 Italian referendum. They curated a dataset focused on politically relevant hashtags and constructed a hashtag network, selecting the top 30 by weighted degree based on co-occurrence. After pruning contextually neutral hashtags, they applied Louvain’s algorithm to cluster these hashtags based on their mutual co-occurrence patterns, distinguishing two dominant clusters representing the YES and NO factions. Using these clusters for labeling, they trained a Gradient Boosting Classifier on tweets. They further employed modularity-based network clustering on the retweet network, yielding derived graphs like the complete retweet and disinformation retweet graph.

Alieva et al. [4] conducted a case study to examine a Russian disinformation campaign centered around the topic of U.S. biolabs and alleged biological weapon development in Ukraine by investigating disinformation narrative that was officially propagated by the Russian government and disseminated through state-funded media. The study employed various methods, including network, super influencer, spread, coordination, and qualitative data analysis. They used the Leiden clustering method to identify communities participating in these conversations. To detect coordination, the researchers defined a network based on activities that occurred within a tight five-minute interval, with a focus on actions such as tweets that shared the same user mentions, hashtags, URLs.

Gu et al. [48] proposed a method to detect numerical ideology positions for Twitter users by evaluating their interactions, specifically follow, mention, and retweet links, with a selected set of politicians.

Schuhbauer et al. [103] analyzed German MPs’ Twitter interactions around the 2021 federal election. They utilized cluster analysis, social network analysis, and sentiment analysis to identify shifts in political communication. Interactions, including mentions, retweets, and replies, were examined collectively and separately using the Girvan-Newman Algorithm to determine clusters.

Luber et al. [74] proposed a methodology for community detection in large networks using hashtag-graphs and semi-supervised Non-Negative Matrix Factorization (NMF) topic models. The authors construct hashtag-graphs by considering hashtags as nodes and co-occurrence relationships between hashtags as edges. They then apply semi-supervised NMF topic models to

extract topics from the hashtag-graphs, incorporating labeled data to enhance the accuracy of community detection.

#### 5.1.4 Related Work in Iranian Tweetersphere Propaganda Analysis

Rahmanian [91] investigated the Iranian state’s narrative on Twitter post the Flight PS752 incident, shedding light on how authoritarian regimes exploit social media for manipulating public opinion. The study employs an integrated approach encompassing thematic analysis and narrative analysis. The study identifies four key narratives in the Islamic Republic network: the enemy narrative, intimidation narrative, threat narrative, and fake news narrative. These narratives are strategically harnessed to construct a fabricated reality and stifle opposing voices. The enemy narrative portrays critics as aligned with foreign adversaries, suppressing opposition. The threat narrative seeks to quash dissent by projecting strength and promising retaliation. The fake news narrative entails disseminating misinformation to manipulate perceptions of the incident. This paper underscores the potency of social media as a tool for shaping propaganda narratives, controlling discourse, and influencing public sentiment, even in the context of individual events like the Flight PS752 incident.

Kermani [63] studied computational propaganda around the #MahsaAmini movement on Twitter. Kermani identified eight specific tactics utilized by the Islamic Republic’s propaganda network: downgrading anti-regime discussions to the level of the government, justifying state policies, creating an illusion that everything is normal, redirecting debates, spreading fake news, trending misleading hashtags, mocking dissidents and activists, and encouraging other users who support the regime’s policies. Kermani identified a noticeable shift in the cyber army’s tactics as the movement grew in prominence. Though the cyber army had previously dominated the Iranian Twittersphere, their efforts to manipulate discussions became ineffective during this particular movement. One intriguing case was the cyber army’s attempt to create confusion by replacing the Persian letter *گ* with its Arabic counterpart *گ* in hashtags, a tactic aimed at diluting the movement’s online presence. However, this manipulation was quickly identified and countered by the movement’s supporters. Kermani also argues that the #MahsaAmini movement indicates a transition to a new phase of hybrid movements in Iran, where both online and street protests can effectively counter online manipulation attempts. He underlines the continued relevance of social media, especially Twitter, in fostering democratic movements despite challenges posed by computational propaganda and suppression.

## 5.2 Methodology

In studying the impact of propaganda within the context of the Women Life Freedom movement on Twitter, this research adopts a methodology tailored for unlabeled tweets datasets. The proposed methodology confronts prominent challenges: the dataset’s lack of labels and the absence of a retweet network. This scenario, further complicated by malicious coordination intertwined with simultaneous coordinated hashtag activism, requires a nuanced distinction between genuine activists and deceptive actors. Given these constraints, the approach combines network-centric and analytical techniques, focusing on unsupervised clustering, hashtag labeling, ideological affiliation assignment, and trend analysis.

### 5.2.1 Research Design

This research adopts a mixed-methods approach to rigorously investigate the influence of propaganda on the Women Life Freedom movement within the Iranian Twittersphere. Utilizing

both quantitative and qualitative analytical techniques allows for a detailed examination of user interactions, hashtag dynamics, and ideological affiliations present in the dataset. The design initiates with data extraction from Twitter to capture movement-relevant tweets. Following data collection, unsupervised clustering techniques are applied to segregate users based on hashtag usage. This clustering is succeeded by a mixed approach of qualitative and quantitative labeling of these hashtags according to their ideological leanings. Using these hashtags, affiliations are assigned algorithmically from their hashtag usage. Post the deterministic assignment of users to specific ideological affiliations, the focus transitions to trend analysis, allowing for an in-depth examination of the dominant topics of discourse and its potential manipulation by propaganda actors.

### 5.2.2 Unsupervised Clustering

Here the objective is to cluster community networks to find ideological hashtag network clusters. This step involves clustering the data and identifying hashtags from clusters to enable data analysis. The process involves a cycle of constructing a bipartite graph, applying the Leiden algorithm, and analyzing the resulting clusters for ideological overlaps. The graph is created with user and hashtag nodes. A user node is connected to a hashtag node if the user has used that hashtag during a selected period, with the edge weight equal to the number of times the user has used the hashtag. The top  $n$  hashtags are included in the analysis, with  $n$  being determined through experimentation to optimize the clarity of ideological distinctions among clusters.

This process is conducted over two distinct periods, termed “Before Mahsa” (January 2022 to September 14, 2022) and “After Mahsa” (September 14, 2022 to April 1, 2023). The period “Before Mahsa” is designed to identify pre-existing networks. These results are then evaluated by comparing the consistency between these two periods.

*Procedure:*

1. Choose a study period: “Before Mahsa” or “After Mahsa” and select an initial number of top hashtags ( $n$ ).
2. Construct a bipartite graph with users and the top  $n$  hashtags as nodes for the selected period. An edge connects a user node to a hashtag node if that user has used the corresponding hashtag during the selected period. The weight of an edge between a user and a hashtag node is set to the number of times the user has used the hashtag.
3. Apply the Leiden community detection algorithm to the bipartite graph to cluster the users and hashtags.
4. Analyze the top hashtags within each cluster from both periods for ideological overlaps. If minimal ideological overlap is observed, the algorithm is done. If not, adjust the number of top hashtags ( $n$ ) and repeat steps 1-4.

**Rationale for Methodology Selection:** In the early stages of this study, various methodologies were explored to optimize the extraction and analysis of meaningful data from hashtags. One such approach involved direct hashtag-to-hashtag network mapping and the subsequent projection of this bipartite user-hashtag graph. However, this method proved to be computationally much more expensive. Moreover, preliminary results from the hashtag-to-hashtag networks were found to be less clear in defining ideological distinctions than the chosen user-hashtag bipartite graph method.

### 5.2.3 Hashtag Labeling

Here, the objective is to label hashtags associated with ideologies in the dataset for a deterministic clustering of the propaganda networks. The common approach to assigning ideological affiliation involves labeling data and training a model on the labeled data. However, when done by a single individual, this process is methodologically flawed due to potential bias and a limited perspective. Hashtags, commonly found in digital communications, particularly on platforms like Twitter, have increasingly been employed as tools for propaganda; specifically, they can serve as slogans that capture and convey political tenets, sentiments, or affiliations [67]. Given these factors, this study opts for an alternative methodology of labeling a minimal set of hashtags. A set of a couple of hundred of these hashtags can be labeled and then verified by other experts in the field. This approach allows for a less biased and more manageable process while maintaining scientific rigor.

Following the unsupervised clustering process, the cluster's usage of hashtags is analyzed in both the Before and After Mahsa periods, with the most frequently and exclusively used hashtag in each community identified and those associated with specific ideologies labeled. Notably, this labeling also includes hashtags that are negatively oriented towards ideologies but do not belong to specific ideologies. These are categorized as "negative towards."

*Procedure:*

1. Find the most frequently and exclusively used hashtags in each community to identify potential ideological affiliation hashtags.
2. Apply the hashtag labeling process, starting by identifying any positive direct ideological affiliations. If no direct positive affiliation can be identified, a multi-step process is initiated:
  - (a) Identify which ideologies are targeted negatively by the hashtag.
  - (b) Determine the cluster(s) where the hashtag is most frequently used.
  - (c) Identify which clusters also predominantly use other hashtags that negatively target the same ideologies.
  - (d) Assign the hashtag to the dominant ideology of these identified clusters. If no specific ideology is dominant, label the hashtag as "negative towards" the targeted ideologies.

After the labeling process, the labeled hashtags are analyzed in terms of their deployment as propaganda techniques. This includes examining frequency and context to discern strategies.

### Assigning Ideological Affiliation

The objective in this phase is to deterministically assign an ideological affiliation to each user in the dataset, following unsupervised clustering and hashtag labeling. The methodology employs a filtering algorithm that considers various factors. These include the frequency of specific hashtag usage, the exclusivity of certain hashtags to specific affiliations, the usage of hashtags negatively against certain affiliations, and a predefined threshold count for considering a user's affiliation. The procedure for assigning ideological affiliations is predicated on the assumption that propagandists use hashtags as a tool for promoting their content and that frequent usage of these hashtags without concurrent negative references suggests support for the corresponding ideology. While not all users employ hashtags to articulate their ideology, this approach helps in potentially identifying frequent members who propagate specific ideological hashtags, indicating a higher likelihood of them being affiliated with certain ideologies or propaganda networks.

*Procedure:*

1. **Count user's usage of affiliated hashtags:** For each user, count the number of times they use affiliated hashtags throughout the dataset. Affiliated hashtags are ones known to be associated with a particular ideological group.
2. **Aggregate counts and filter:** Aggregate these counts for each user and affiliation. If a user has used an affiliated hashtag fewer than a set threshold  $\beta$ , that affiliation is disregarded for the user.
3. **Assign affiliation:** For each user, the algorithm identifies the most frequently used affiliated ideological group. If a user also uses hashtags negatively towards this affiliation, the algorithm checks whether the user employs hashtags from other affiliations. If so, the user is assigned to that other affiliation. Otherwise, the user is considered to use the hashtag negatively to mock or criticize the corresponding ideology and is assigned the 'Unaffiliated People' affiliation.
4. **Unaffiliated users:** Users who do not use any affiliated hashtags after applying the threshold are classified as 'Unaffiliated People'.

In this framework, 'Unaffiliated People' signifies a lack of association with the specifically labeled ideologies within our dataset. It is essential to emphasize that this does not necessarily reflect a general absence of political or ideological leanings on the part of the user.

Following the determination of ideological assignments, the subsequent phase focused on analyzing the frequency of hashtag usage within the identified networks. This involved a comparative study of the deployment patterns both before and after Mahsa. Such an examination offered insights into any shifts or changes in discourse strategies by the respective propaganda networks in response to the event.

Within the scope of this study, a threshold for category inclusion was established based on representation in the dataset. Communities representing less than 1% of the dataset are deemed too small for substantial impact and meaningful analysis and were consequently excluded from trend analysis.

#### 5.2.4 Trend Analysis

The goal of this analysis is to explore the evolution of the topics of discussion over time and the influence of propaganda networks on redirecting discussions. The analysis is performed using n-grams of tweets, computed daily, and then aggregated based on the number of users who used a particular n-gram on a given day per affiliation group.

**Pipeline Overview** The data processing pipeline is as follows:

1. **Cleaning:** The tweet text is cleaned by removing mentions, the five most frequently used hashtags (e.g., #MahsaAmini), Persian punctuation, mentions and URLs.
2. **Normalization and Stop Words Removal:** The tweets are normalized and stop words, in both Persian and English, are removed using the StopWordsCleaner from the sparknlp library.

3. **Tokenization and Lemmatization:** The cleaned tweets are tokenized into individual words (tokens) and these tokens are lemmatized, reducing words to their base or root form.
4. **N-Grams Generation:** N-grams, ranging from unigrams (single words) to 4-grams (four-word sequences), are generated from the lemmatized tokens.
5. **TF-IDF:** A CountVectorizer is used to convert the text documents to vectors of term frequency, followed by computing IDF (Inverse Document Frequency) to reflect the importance of a word to a document in the collection.
6. **Aggregation:** The resulting n-grams are aggregated by the number of users using a particular n-gram per day and per affiliation group.
7. **Output:** The final result of this pipeline is a dataframe that contains, for each date, n-gram, and assigned affiliation usage count of that n-gram, the count of unique users who used that n-gram, and the rank of the n-gram in terms of user count and n-gram size.

**Visualization and Analysis** To methodically assess the dominant themes of discourse, this study focuses on the ten most prevalent n-grams. The n-grams are ranked according to an aggregate measure based on the number of unique users using an n-gram on a weekly basis by affiliation. Bi-grams featured in the top 10 list are accounted for, excluding those already present in tri-grams. The initial phase of analysis concentrates on the "After Mahsa" period, aiming to delineate and visualize salient shifts in the main discourse topics. Subsequently, periods characterized by pronounced activity spikes are examined in more granular temporal brackets. The top discussion topics in the smaller time frames are visualized using the same ranking methodology. However, bi-grams are not excluded if they exist in overall top bi-grams over the extended after Mahsa period. This then facilitates a qualitative exploration into the potential catalysts and subsequent outcomes associated with these shifts in discourse.

**Choosing N-grams Over Other Topic Modeling** Although NLP topic modeling methods like BERT topic modeling can provide a robust approach for topic modeling, their implementation was not feasible given the large magnitude of the dataset. These methods require the entirety of the data to be loaded into memory at once, making them computationally intensive and impractical for this analysis. Moreover, BERT and N-Gram topic modeling often yield better results when combined with sentiment analysis or when applied to labeled datasets to interpret or as a feature for Neural networks or specific categories of text such as hate speech.

Given these challenges and the lack of pre-trained models and labeled datasets in Persian, an n-gram-based approach was chosen as a more viable alternative. N-grams, sequences of n-words used together, provide a practical and efficient way of analyzing large datasets. They can effectively capture the main topics of text by identifying recurring sequences of words that carry significant meaning. Moreover, the generation and use of n-grams are not dependent on the availability of pre-trained models or labeled datasets, thereby circumventing the limitations faced in Persian language processing. This makes n-grams a particularly useful tool for topic modeling and trend analysis in large-scale and resource-limited settings like this analysis.

### 5.3 Results and Analysis

This section presents a comprehensive examination of the study's findings. The analysis centers on the utilization of unsupervised clustering techniques to discern patterns of hashtag usage

across two distinct timeframes: pre and post Mahsa. The subsequent exploration focuses on the development of ideological groups, their associations, and hashtag narratives. Moreover, the investigation extensively examines the propaganda strategies employed by these factions. This includes an inquiry into their propagated narratives, patterns of hashtag usage, and employed propaganda techniques. This section provides insights into the complex online discourse landscape by analyzing trends and linguistic nuances. Additionally, it illuminates the deliberate employment of specific storylines within various ideological groups, underscoring the strategic nature of such utilization.

### 5.3.1 Unsupervised Clustering

In this study, two distinct periods were analyzed using unsupervised clustering, namely “Before Mahsa” spanning from January 2022 to September 14, 2022 and “After Mahsa” from September 14, 2022 to April 1, 2023.

For the “Before Mahsa” duration, the clustering outcome was optimally distinct when the top 1200 hashtags were selected. On the other hand, during the “After Mahsa” period, there was a trend of overlapping clusters. This overlapping could be due to the intertwined usage of hashtags during this latter timeframe. Additionally, incorporating a larger set of top hashtags led to the emergence of more, smaller communities.

Using the Leiden community detection algorithm, distinct communities were discerned. These communities were labeled based on the predominant themes evident in their hashtag usage. The identified communities were: “People”, “Islamic Republic”, “NFT/Crypto Currency”, “Monarchy”, “K-pop BTS Fans”, “International/English Speaking Activists”, “MEK”, “Arabic Media”, and “K-pop EXO Fans”. Notably, the “Islamic Republic”, “Monarchy”, and “MEK” communities emerged as networks with prominent propagandistic tendencies.

A visual representation of the primary hashtags employed by these communities can be found in a series of figures, starting from Figure A.1 for the “People” community, moving to Figure A.9 for the “K-pop EXO Fans” community. Specifically, the “Islamic Republic” community’s hashtag usage is presented in Figure A.2, the “Monarchy” community in Figure A.4, and the “MEK” community’s propagandistic tendencies are evident in their respective figure. Lastly, a pie chart, offering visualization of the community distribution during the “Before Mahsa” period with the top 1200 hashtags, is showcased in Figure 5.1.

#### Addressing RQ1

*Who are the political-ideological groupings active on Twitter in the wake of the Women Life Freedom movement?*

Through applying the unsupervised clustering method of Leiden clustering on user-hashtag graph, three dominant political-ideological clusters were discerned within the Iranian Twitter-sphere, namely: Islamic Republic, Monarchy, and MEK.

### 5.3.2 Hashtag Labeling

The unsupervised clustering identifies three key propaganda networks: “Islamic Republic”, “MEK” and “Monarchy”. The associated hashtags, categorized by direct affiliations or negative orientations towards alternate ideologies, are presented in Appendix B. The Tables 5.1, 5.2, and 5.3 present the top 10 most frequently used labeled hashtags per affiliation. The following is an analysis of the labeled hashtags:

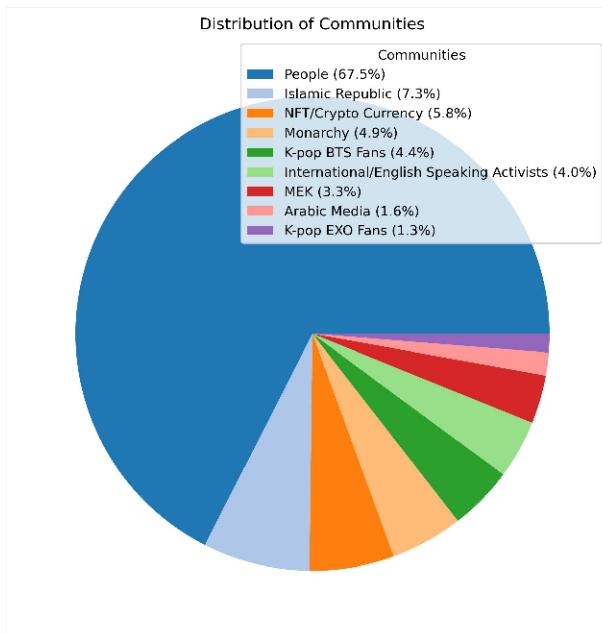


Figure 5.1: Distribution of the Unsupervised Network Clusters: This pie chart visualizes the proportion of various networks clustered with the top 1200 hashtags and 113,087 users before Mahsa (January 2022 to September 14, 2022).

**Islamic Republic’s Hashtag Analysis:** For the “Islamic Republic” network, the prevailing themes in their hashtags lean towards the Shia ideology with references to the savior figure, the 12 Shia Imam, Mahdi. (اللهم اجل لوليك الفرج) (‘O Allah, hasten the return of your relief/Imam Zaman/Mahdi’) and (اللهم صل على محمد وآل محمد وجعل فرجهم) (‘O Allah, bless Muhammad and the family of Muhammad and hasten the reappearance of the saviour’) and ThePromisedSaviour resonate with Shia eschatological beliefs about Imam Mahdi’s reappearance. Their use of the term ‘GreatSatan’ for the USA combines both **Loaded Language** and **Name Calling/Labeling** techniques, casting the USA in a negative light. The terms (ایران قوی) (‘Strong Iran’) and (پاچت بصیرت) (‘Capital of Insight’) exemplify the **Flag-waving** technique, portraying an image of national strength and insight. The term (لیک یا خامنه ای) (‘At your service, O Khamenei’) not only signifies allegiance to Khamenei using the **Appeal to Authority** technique but also acts as a rallying **Slogan** for followers of the Islamic Republic.

**MEK’s Hashtag Analysis** MEK’s hashtags spotlight their organizational aspirations. Hashtags about the “National Council of Resistance” is an organization primarily associated with the MEK [1]. By referencing the “National Council of Resistance” MEK harnesses the **Flag-waving** technique, tapping into nationalistic feelings to garner support. The hashtags “IStandWithMaryamRajavi” and (درود بر رجوی) (‘Salute to Rajavi’) further underscore the **Appeal to Authority** technique, signifying their allegiance and respect to their leader, Maryam Rajavi.

**Monarchy’s Hashtag Analysis:** The Monarchy’s hashtags tap into strong nationalistic feelings. Utilizing the **Flag-waving** technique, hashtags such as (انقلاب ملی ایران) (‘Iranian National Revolution’) and (اتحاد ملی ایران) (‘Iran’s National Coalition’) evoke sentiments of national unity

and revolution. The term جاویدشاه ('EverlastingKing') can be considered both a **Slogan** and an instance of **Flag-waving**, as it emphasizes the idea of a continuous, unbroken monarchy that embodies the nation's spirit. The statement پهلویسم تنها راه نجات ('PahlavismTheOnlyWay-ToSalvation') uses the **Dictatorship** technique, suggesting that there's no alternative but to follow the path of Pahlavism. The hashtag مرگ بر سه فاسد ملا چی مجاهد ('Death To Three Corrupt Mullah Leftist Mujahedin/MEK') employs **Loaded Language**, **Dictatorship**, **Name Calling/Labeling** and **Reductio ad Hitlerum (Guilt By Association)** techniques, the latter by implying that these three distinct entities are all equally corrupt. Further, the **Dictatorship** technique is evident as the hashtag implies a singular solution, the death or demise of the named entities. Moreover, same techniques are used in the hashtag (#DeathToRuhollahMesiAbdollah), which seeks to malign Masih Alinejad, Hamed Esmailion<sup>1</sup>, and Abdullah Mohtadi<sup>2</sup> by associating them with Ruhollah Khomeini, the leader of the Islamic Revolution who is viewed negatively by the followers of the movement.

Table 5.1: Top 10 Most Frequently Used Hashtags from the labeled set for Islamic Republic after Mahsa. The same hashtags with different writing are merged together.

Hashtag and Translation	Propaganda Technique	Count
لبک یا خامنه‌ای - At your service, O Khamenei	Slogan, Appeal To Authority	1025435
اللهم صل على محمد وآل محمد وعجل فرجهم / اللهم عجل لوليك النرج O Al-lah, (bless the Prophet Muhammad and his progeny) and hasten the relief of the savior	Slogan, Loaded language	135979
ThanksIRGC	Slogan	73622
ایران قوی - Strong Iran	Flag Waving, Slogan	58464
ThePromisedSaviour	Slogan, Loaded Language	50971
GreatSatan (USA)	Name Calling / Labelling, Loaded Language, Slogan	38715
حب الحسین یجمعنا - Love of Hussein brings us together	Flag Waving, Slogan, Loaded Language	34903
پایخت بصیرت - Capital of Insight	Flag Waving, Slogan	29204

### 5.3.3 Affiliation Assignment

Figure 5.2 depicts the influence of the threshold on the sizes of the identified ideological groups. Analyzing the consistency of hashtag usage during both “Before Mahsa” and “After Mahsa” periods, a threshold of 5 hashtags was determined to be a robust minimal value, ensuring consistency in ideological affiliations. The overall distribution of affiliations is portrayed in Figure 5.3. Of these communities, the “Islamic Republic” represented 4.1% of the dataset, the “Monarchy” accounted for 3.5%, and notably, the ‘MEK’ community was significantly smaller, constituting only 0.2% of the entire dataset. Figures C.1, C.3, C.4, and C.2 illustrate hashtag usage patterns within ‘Unaffiliated People’, ‘Islamic Republic’, ‘MEK’ and ‘Monarchy’ groups, respectively, across both periods. Figure 5.4 illustrates the distribution shift of affiliations before and after

<sup>1</sup> After the tragic downing of Flight PS752, Hamed Esmailion, having lost his wife and daughter in the incident, became a vocal advocate for justice as the president of The Association of Victims' Families of Flight PS752. He later resigned on 7 March 2023, after announcing he is planning to focus more on the Iranian revolution.

<sup>2</sup> Abdullah Mohtadi is the leader and secretary general of the Komala Party of Iranian Kurdistan.

Table 5.2: Top 10 Most Frequently used Hashtags from the labeled set for MEK after Mahsa. The same hashtags with different writing are merged together.

Hashtag and Translation	Propaganda Techniques	Count
FreeIranTelethon	Slogan, Virtue	5288
کانون شورشی - Rebellion Center	Slogan, Flag-waving	2117
FreeIran10PointPlan	Slogan	1952
WeSupportMEK	Bandwagon, Slogan	1740
(ایران) ملی مقاومت - شورای ملی مقاومت (ایران)	Flag-waving, Appeal to Authority	1705
درود بر رجوی - Salutation to Rajavi	Slogan	1692
IStandWithMaryamRajavi	Bandwagon, Slogan	1661
NCRIAlternative	Slogan, Appeal to Authority	1181
دروド بر کانونهای شورشی - درود بر کانونهای شورشی	Flag-waving, Slogan	1098

Table 5.3: Top 10 Most Frequently Used Hashtags from the labeled set for Monarchy after Mahsa and their propaganda techniques. The same hashtags with different writing are merged together.

Hashtag	Propaganda Techniques	Count
جاویدشاه - EverlastingKing	Flag-waving, Slogan	1112009
مرگ بر سه فاسد ملا چی مجاهد - Death To Three Corrupt Mullah Leftist Mujahedin (MEK)	Name Calling/Labeling, Guilt by Association, Slogan	236759
جاویدشاه رمز پیروزی - Everlasting King symbol of victory	Flag-waving, Slogan	188185
جاوید رضا شاه دوم پهلوی سوم Pahlavi III - Everlasting Reza Shah II	Flag-waving, Slogan	103462
RezaPahlaviIsMyRepresentative	Appeal to Authority, Slogan	82851
KingdomWithPahlavi	Appeal to Authority, Flag-waving, Slogan	62580
انقلاب ملی ایران - Iranian National Revolution	Flag-waving, Slogan, Appeal to Time	34676
بازگشت به مشروطه - Return to Constitutionalism	Slogan, Appeal to Time	31209

the Mahsa: Islamic Republic-affiliated accounts were 81.54% before and 18.46% after; MEK-related accounts constituted 98.79% before and 1.21% after; Monarchy-related accounts were 66.44% before and 33.56% after; and unaffiliated accounts were 65.29% before and 34.71% after. This suggests a significant surge in activity, primarily within the Unaffiliated, Monarchy, and Islamic Republic groups, while the size of the MEK group remained relatively unchanged.

When examining the top 50 hashtags in terms of frequency count before and after the Mahsa period, distinct patterns emerge for each of the three groups, as shown in Figures 5.5, 5.6, and 5.7 a significant increase in hashtag activity is observed across all groups post-Mahsa. Post-Mahsa, these three ideological groups have become notably radicalized. This is evident in their prevalent use of “death to” in hashtags, indicating a shift toward extremism, and in ‘Islamic Republic’ group, its endorsement of executions and ‘decisive response’ to protesters suggests a readiness

for punitive or oppressive actions. Such a slogan, especially in the context of political and social movements, represents a clear and unequivocal denunciation of the target entity, be it a person, group, or concept. It also signifies a call to action beyond mere dissent, pointing to an intention to eliminate or confront the adversary decisively.

In the analysis of hashtags after Mahsa, the prominence of the مهسا امینی (Mahsa Amini) hashtag across ideological groups is evident. It was the top hashtag for the 'Monarchy' group, while it was ranked fifth in both the 'Islamic Republic' and 'MEK' communities. This use of the hashtag post-Mahsa, along with the continuation of their previous hashtags and introduction of other hashtags, can be observed in Figures 5.9, 5.10, and 5.13. This data indicates a strategic approach by these groups, leveraging the visibility of significant events to enhance their messaging reach. All these groups leveraged the Mahsa Amini hashtag to propagate their respective agendas and ideologies. For instance, MEK used the hashtag in combination with قیام تا سرنگونی (Uprising Until Overthrow), مرگ بر سر چه شاه باشه چه رهبر (Death To The Oppressor Whether King Or Leader), مسعود رجوی (Massoud Rajavi). Monarchy used اعتصابات و اعتراضات سراسری (Nationwide protests and strikes), من وکالت میدهم (I Give Attorneyship). The Islamic Republic used hashtags like لبک یا خامنه‌ای (At Your Service, O Khamenei), گشت ارشاد (Morality Police), حجاب (Hijab), مهسا‌امینی (MahsaAmini), this hashtag used Persian half space instead of underline), پرخورد قاطع (Decisive Response), اغتشاشات (Riots) and in particular, by using شاه چراغ (Shah Cherag) hashtag, attributed blame to the Mahsa Amini movement for the terrorist attack on the Shi'ite shrine<sup>3</sup>, deploying the *Reductio ad Hitlerum* propaganda technique.

Furthermore, an intriguing strategy observed in the Islamic Republic community and previously reported by Kermani [63] involves the use of similar-looking hashtags that have slightly different Unicode characters. The top 20 most frequently used hashtags containing مهسا (Mahsa) from this community are presented in Table 5.4. Although they appear the same on the surface, they have different functions: they capitalize on Twitter's auto-suggestion feature to dilute the actual movement's online presence.

Another notable observation is the use and ranking of the زن زندگی آزادی (Women Life Freedom) hashtag. The ranking indicates the rank of the hashtag when combining the top 50 of the most frequent hashtags used before and after Mahsa. The Monarchy community ranked 12th; it was 56th for the Islamic Republic and not on the top list for MEK. The distinct application of this hashtag by different groups can be seen in Figures 5.8, 5.11 and 5.12. Both the 'Islamic Republic' and 'Monarchy' groups harnessed this hashtag to drive their specific agendas, with the 'Islamic Republic' using it in a critical or denigrating context by using شاه چراغ (Shah Cherag), حجاب (Hijab), زن زندگی آگاهی (Women Life Awareness), زن عفت افتخار (Women Chaste Honor), زن زندگی شهادت (Women Life Martyrdom) and پرخورد قاطع (Decisive Response). It is also significant that both groups concurrently utilized the مرد میهن آبادی (Men Homeland Prosperity) hashtag with the زن زندگی آزادی (Women Life Freedom) hashtag. This suggests a conscious effort to create narratives that relate men's roles with discussions on women's freedoms.

The three ideological groups, namely the Islamic Republic, MEK, and Monarchy, demonstrate shared tendencies in their deployment of various propaganda techniques. Predominantly, these techniques include the Slogan, augmented by Flag Waving, Appeal To Authority, Loaded Language, Name Calling/Labeling, Dictatorship, *Reductio ad Hitlerum* or Guilt By Association,

---

<sup>3</sup>The killing of Shi'ite pilgrims came on the same day that Iranian security forces clashed with increasingly strident protesters marking the 40-day anniversary since the death of Mahsa Amini. Interior Minister Ahmad Vahidi blamed the protests sweeping Iran for paving the ground for the Shiraz attack. <https://www.reuters.com/world/middle-east/riot-police-deploy-iranian-cities-people-gather-aminis-memorial-2022-10-26/>

Glittering Generalities (Virtue), and the persuasion technique of Appeal to Time; here, the purpose behind this persuasion technique is clear: to influence opinion and elicit action.

For the Islamic Republic group:

- **لیک یا خامنه‌ای** (At your service, O Khamenei) encapsulates the techniques of Slogan, Flag Waving, and Appeal to Authority.
- Hashtags like **حرب الحسین یجعنا** (Love of Hussein brings us together), **ایران قوی** (Strong Iran), **ایران ما** (Our Iran), and **پرچم ایران بالاست** (Iran's Flag Is High) exemplify Flag Waving and Slogan.
- **مکتب امید** (School of Hope) and "Hero" represent Virtue and Slogan techniques.
- **سلبریتی‌های شورش الاغها** (Donkeys' Rebellion), "Great Satan" (referring to the USA), and **وزاری دوزاری** (Wanky Celebrities) serve as instances of Name Calling and Loaded Language.
- **حمایت انقلابی** (Revolutionary Support) and **برخورد قاطع** (Decisive Response) signal the Appeal to Time technique, indicating a moment for revolutionary action and a definitive reaction to protests. Here, the technique extends beyond mere persuasion.
- Finally, **قه کش را اعدام نکنید** (Do Not Execute The Machete-wielder) employs the Dictatorship and Guilt By Association (or Reductio ad Hitlerum) techniques. Used sarcastically, this slogan calls for the execution of protesters by associating them with the violent imagery of a machete-wielder and suggests no other alternative actions.<sup>4</sup>

For the Monarchy group:

- **جاوید شاه رمز پیروزی** (Everlasting King, Symbol of Victory) and **جاوید رضاشاه دوم** (Everlasting Reza Shah the Second) encompass techniques such as Slogan, Flag Waving, and Appeal To Authority.
- Hashtags like **اعتراضات سراسری** (Nationwide Strikes), **اعتراضات سراسری** (Nationwide Protest), and **بازگشت به مشروطه** (Return to Constitutionalism) fall under the Appeal To Time technique, signaling the time has come for these nationwide protests and strikes, and a return to a constitutional monarchy.
- **مرگ بر سه فاسد ملا چی مجاهد** (Death To Three Corrupt: Mullah, Leftist, Mujahed) employs several techniques: Loaded Language, Slogan, Name Calling/Labeling, Dictatorship, and Reductio ad Hitlerum or Guilt By Association. The Dictatorship element emerges from the slogan's call for a singular action against these three groups. Meanwhile, the Guilt By Association technique manifests by grouping the Mullahs, Leftists, and Mojahed (MEK) together, labeling them as corrupt.
- **زن زندگی آزادی** (Women Life Freedom) employs the "Virtue" or "Glittering Generalities" technique, emphasizing universally accepted positive values
- **پهلویسم تنها راه نجات** (Pahlavism, the Only Way to Salvation/Liberty) reflects the Dictatorship technique by positing a singular solution and eliminating other alternatives.

For the MEK group:

<sup>4</sup>This hashtag is a response to the hashtag **نه به اعدام** (No To Execution) and **اعدام نکنید** (Do Not Execute) used by hashtag activists.

- مرگ بر ستمگر چه شاه باشه چه رهبر (Death To The Oppressor Whether King Or Leader) is characterized by techniques like Slogan, Guilt By Association, Loaded Language, Dictatorship, and Labeling due to its clear and singular call for action and its designation of both the King and the Leader as oppressors. Furthermore, it employs Reductio ad Hitlerum by equating the King (Pahlavi) and the Leader (Khamenei), suggesting they are similarly oppressive.
- اعتصابات سراسری (Uprising Until Overthrow), قیام تا سرنگونی (Nationwide Strikes), IranProtests, and آتش جواب آتش (Fire Answers Fire) are indicative of the Slogan technique and also convey an Appeal To Time due to the emphasis on urgent action.
- درود بر مجاهد (Greetings to Mujahideen) and WeStand4FreeIran showcase Flag Waving, celebrating or expressing pride in one's group, and also the Slogan technique due to their repetitive and rallying nature.
- مسعود رجوی (Masoud Rajavi) and مریم رجوی (Maryam Rajavi) use the Appeal To Authority technique, invoking respect and reverence for figures of authority within the organization.
- آزادی آزادی آزادی (Freedom Freedom Freedom) is characterized by the Virtue technique, tapping into universally accepted values of freedom and aspiration for liberty.

Furthermore, the specific names mentioned by both the ‘Monarchy’ and ‘MEK’ groups function as Loaded Language. These names often refer to individuals who lost their lives in the aftermath of protests or were detained. By invoking these names, the groups aim to incite strong emotional reactions in their audience, stirring feelings of empathy, anger, and sorrow. Such usage serves to heighten the emotional appeal and galvanize support or action based on the elicited emotions.

Similarly, the ‘Islamic Republic’ also resorts to this strategy by emphasizing the victims of the Shah Cheragh terrorist attack. By highlighting the innocent lives lost in this tragedy, they aim to shift the narrative and engender sympathy and outrage towards the perpetrators, indirectly associating the tragedy with the larger protests and the Mahsa Amini movement. This tactic can be seen as a way to divert attention and channel the national sentiment of sorrow and anger toward their narrative, thereby consolidating their position and discrediting opposing voices.

Furthermore, the propaganda technique of Repetition is also prevalent in this context. Using hashtags multiple times, these groups aim to reinforce their message and make it more memorable. Repetition is a common strategy in propaganda because it helps embed the desired narrative in the minds of the audience, making it more likely for them to accept and internalize the information being presented.

Hashtag (Translation)	Unique User Count	Total Usage Count
مهسا_امینی (Mahsa Amini)	6038	130515
مهسا_امینی (Mahsa Imeni)	2088	10138
مهسا_امینی (MahsaAmini (Half Space))	1572	5032
مهسا_امینی (Mahsa Amiti)	59	3095
دروغ_ضرب_و_شتم_مهسا (Lie: Beating of Mahsa)	757	2553
مهسا_امینی_بهانه_است (Mahsa Amini is an Excuse)	387	2134
مهسا_امینی_رمز_فتنه (Mahsa Amini Symbol of Sedition)	258	1828
مهسا (Mahsa)	460	1003
مهسا_امینی (Mahsa Amini - different Unicode)	235	717
مهسا_امینی (Mahsa Amini - different Unicode)	236	499
مهسا_امینی (Mahsa's Amini)	239	487
مهسا_امینی (Mahsa Amini - different Unicode)	125	379
مهسا_امینی (Mahsa Amini - extra underscore)	183	369
دروغ_مهسا_امینی (Lie: Mahsa Amini)	14	364
فتنه_مهسا (Sedition Mahsa)	15	337
به_بهانه_مهسا (On the Excuse of Mahsa)	81	335
مهسا_امینی (Mahsa Amini - different Unicode)	116	331
فتنه_مهسا_امینی (Sedition Mahsa Amini)	40	301
لعنت_به_مهسا_امینی (Curse on Mahsa Amini)	33	258
مهسا_امینی_پیراهن_ Osman (Mahsa Amini - Osman's Shirt)	89	255

Table 5.4: Frequency and Unique User Count of Variant Hashtags of top 20 hashtags including مهسا (Mahsa) in the 'Islamic Republic' Community

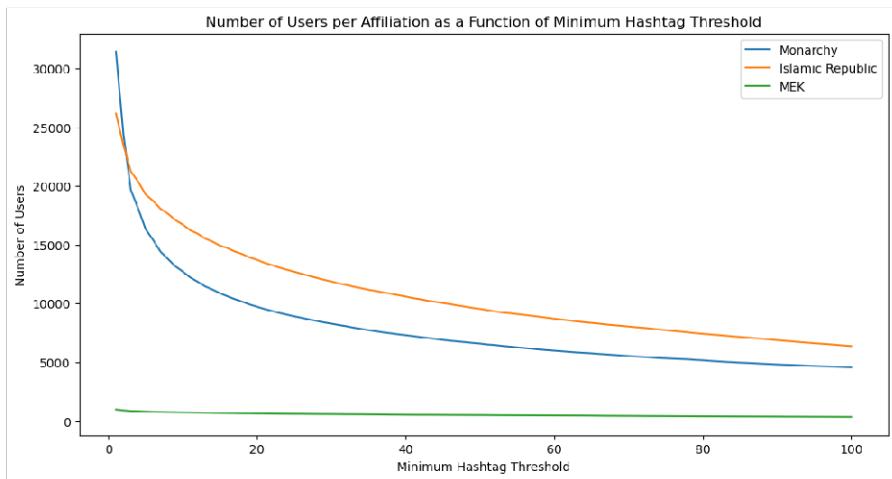


Figure 5.2: Effect of the minimum hashtag threshold on the sizes of ideological groups.

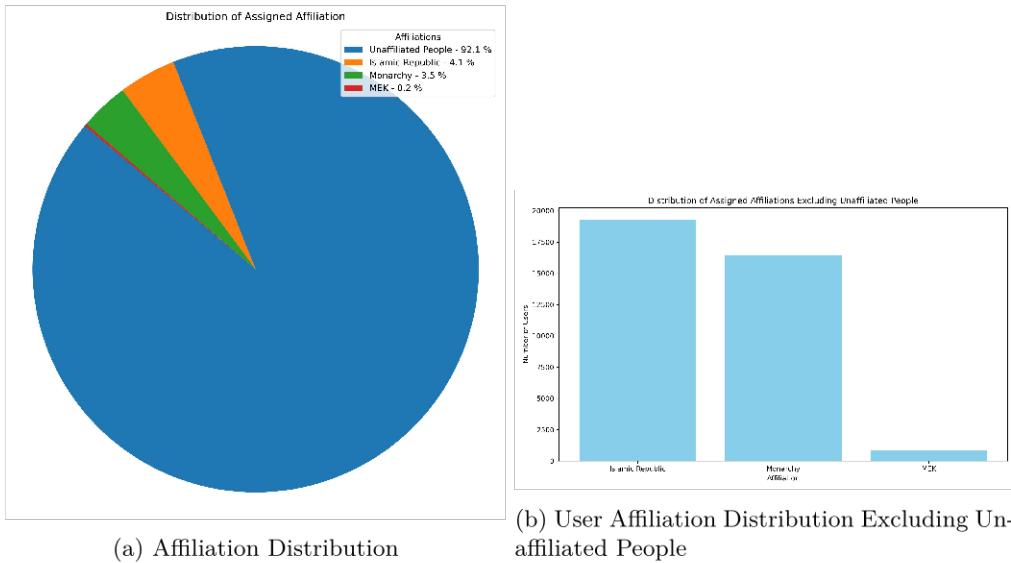


Figure 5.3: Overall Affiliation Distributions of all the users in the dataset with a threshold of 5 minimum hashtag usage. Unaffiliated people are not affiliated with the other presented affiliations.

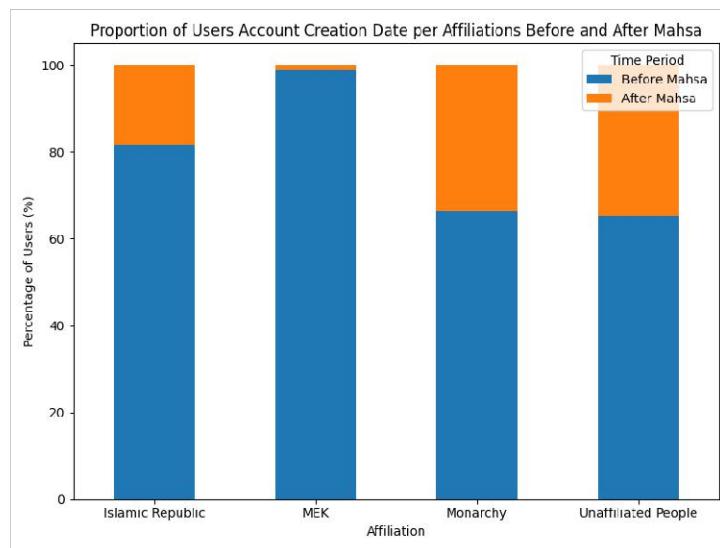


Figure 5.4: Stacked Percentage Bar Plot Illustrating Affiliation Portions Before and After Mahsa: Islamic Republic (Before: 81.54%, After: 18.46%), MEK (Before: 98.79%, After: 1.21%), Monarchy (Before: 66.44%, After: 33.56%), Unaffiliated (Before: 65.29%, After: 34.71%).

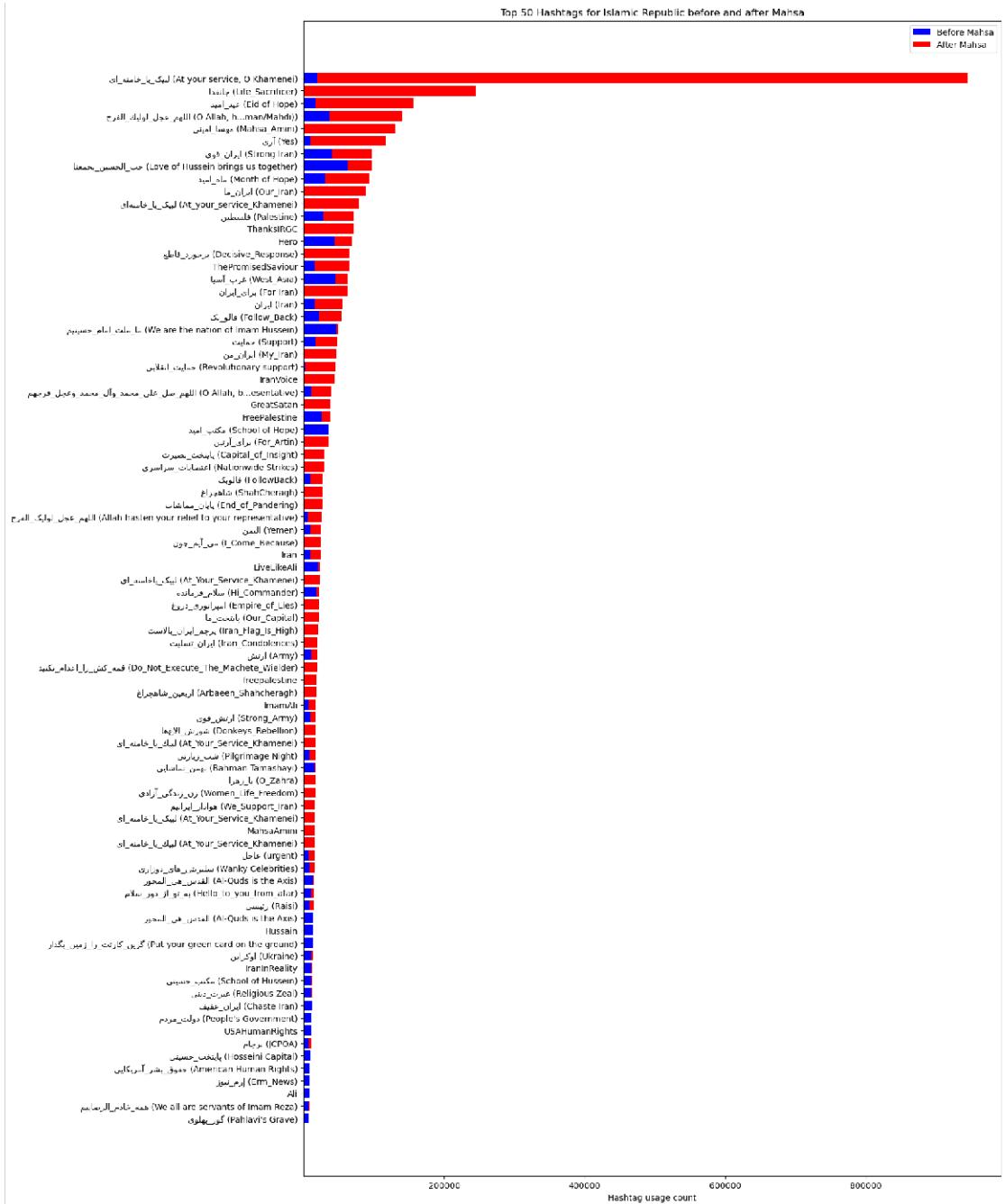


Figure 5.5: Top 50 most frequently used hashtags before and after Mahsa for the Islamic Republic community.

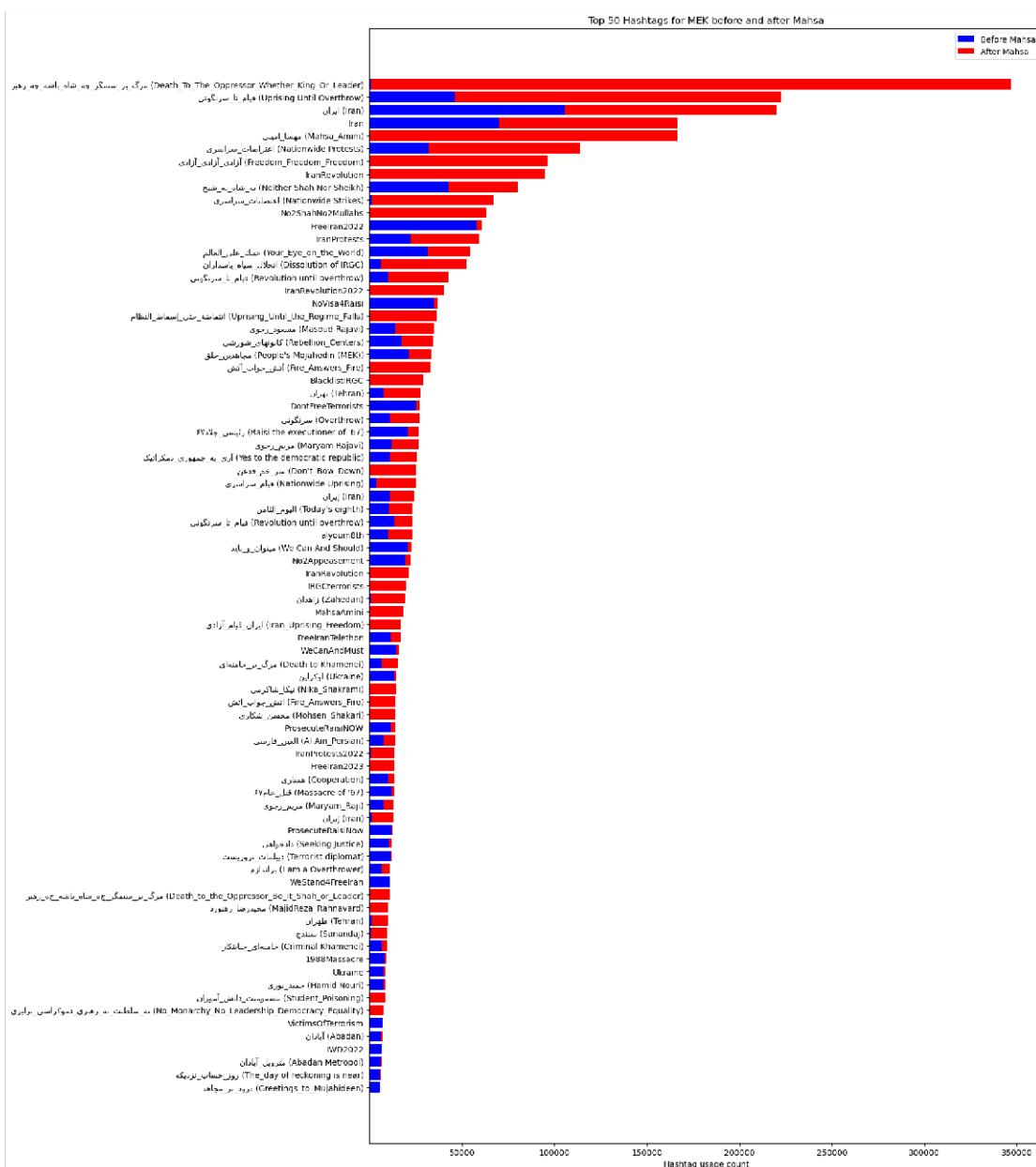


Figure 5.6: Top 50 most frequently used hashtags before and after Mahsa for the MEK community.

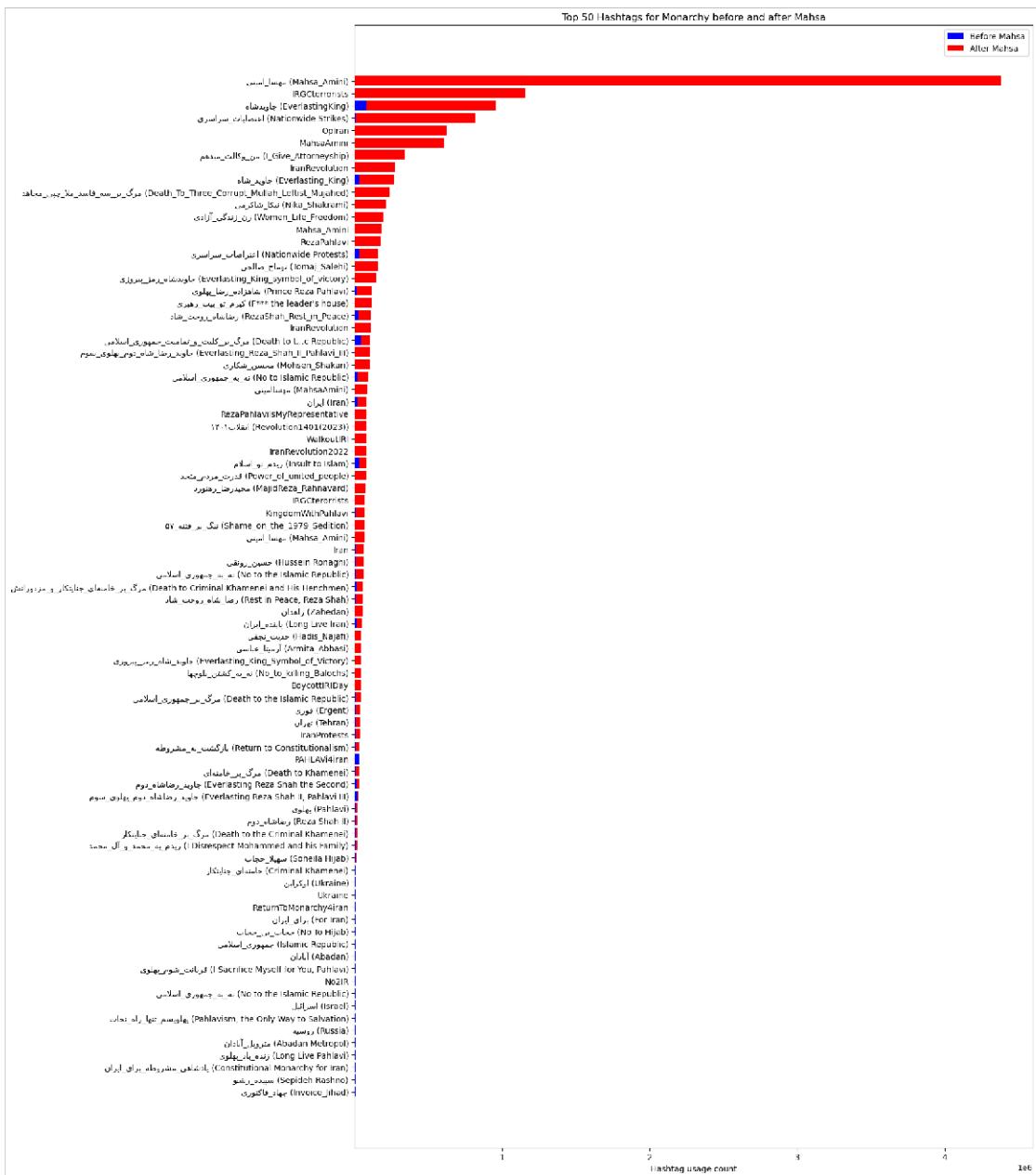


Figure 5.7: Top 50 most frequently used hashtags before and after Mahsa for the Monarchy community.

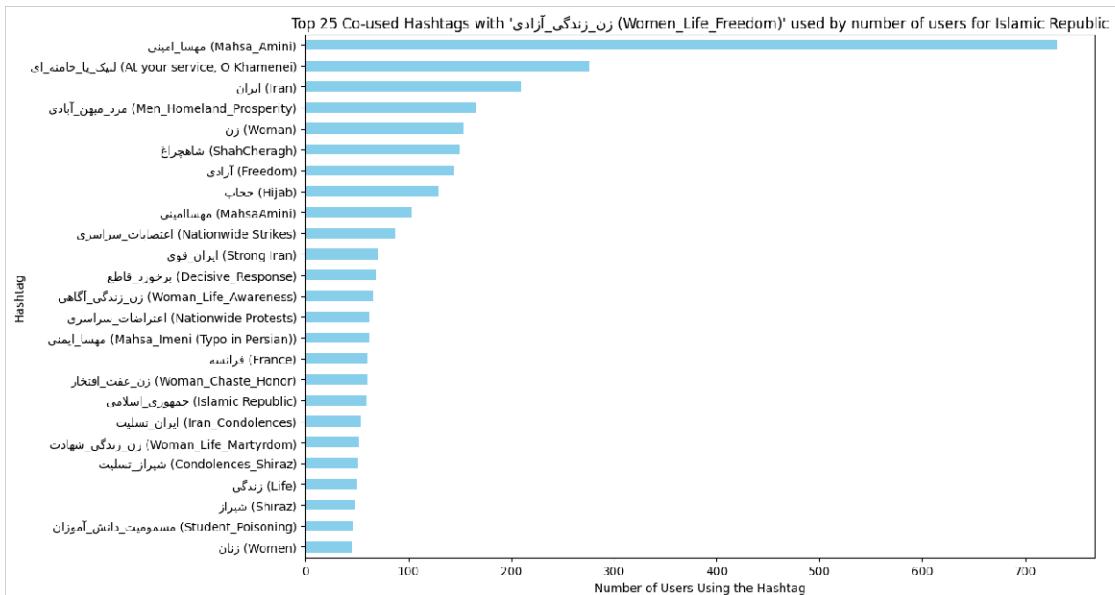


Figure 5.8: Top Co-used hashtags with زن\_زندگی\_آزادی (Women Life Freedom) hashtag within Islamic Republic community.

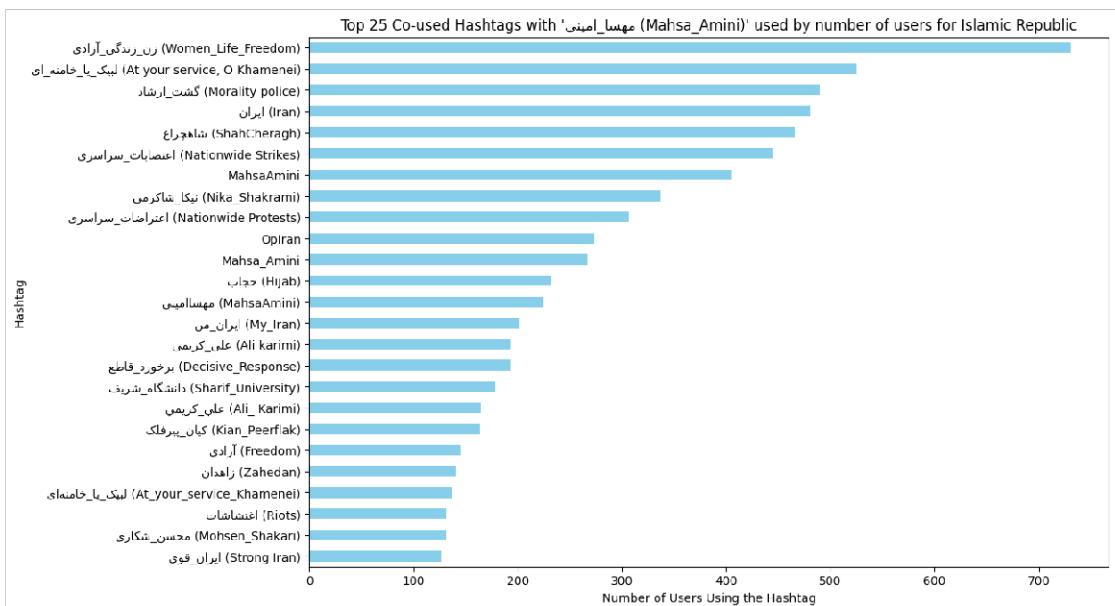


Figure 5.9: Top Co-used hashtags with مهسا\_امینی (Mahsa Amini) hashtag within Islamic Republic community.

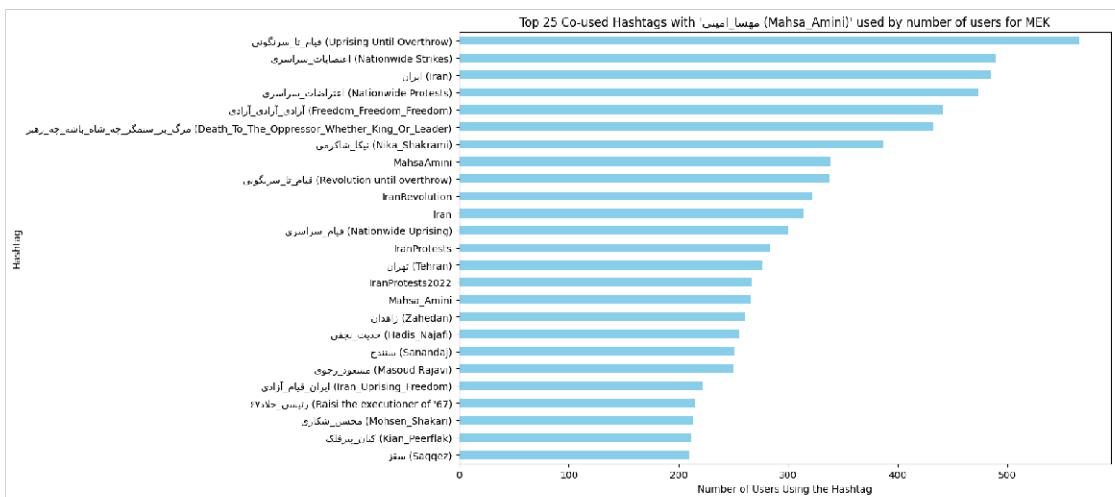


Figure 5.10: Top Co-used hashtags with مهسا امینی (Mahsa Amini) hashtag within MEK community.

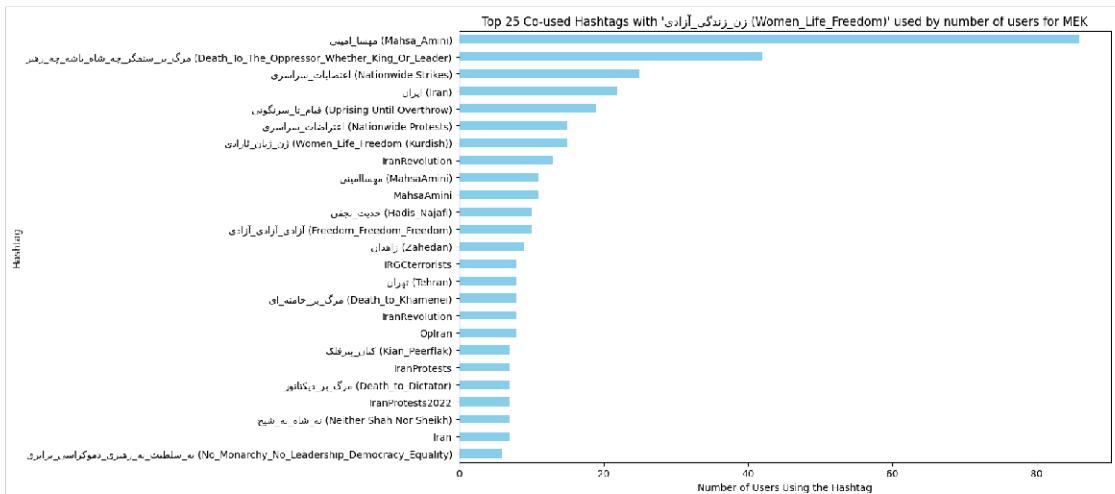


Figure 5.11: Top Co-used hashtags with زن زندگی آزادی (Women Life Freedom) hashtag within MEK community.

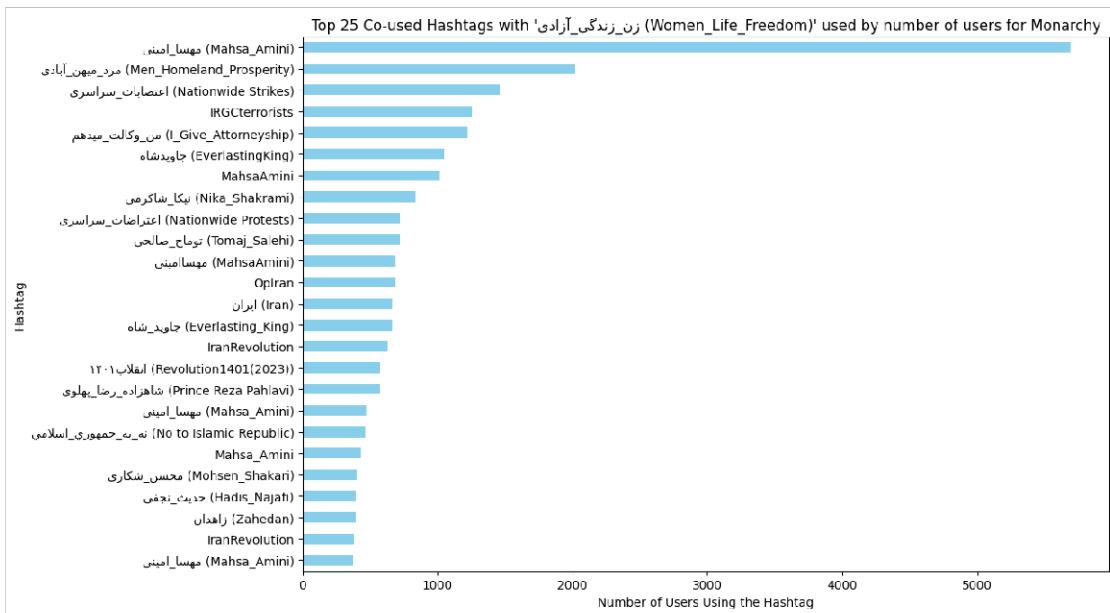


Figure 5.12: Top Co-used hashtags with زن\_زندگی\_آزادی (Women Life Freedom) hashtag within Monarchy community.

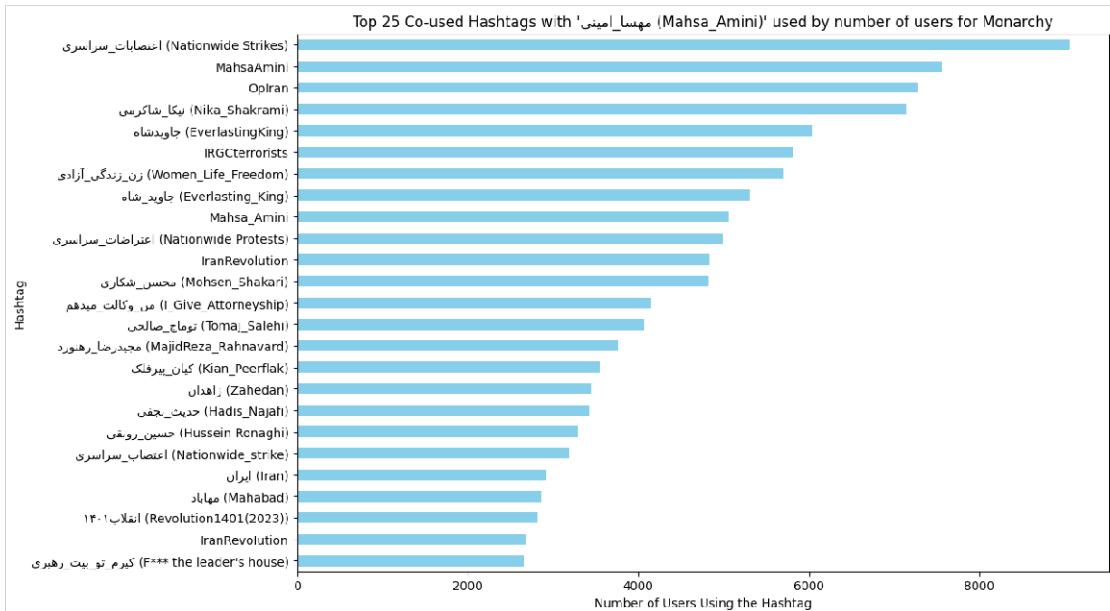


Figure 5.13: Top Co-used hashtags with مهسا\_امینی (Mahsa Amini) hashtag within Monarchy community.

## Addressing RQ2

*How do these ideologically conflicting groupings use hashtags, and how do they relate to propaganda techniques?*

The ideologically conflicting groups, namely the "Islamic Republic", "MEK", and "Monarchy", leverage hashtags on Twitter to propagate their respective narratives and ideologies, which can often be aligned to established propaganda techniques.

The Islamic Republic predominantly uses their hashtags to invoke strong religious sentiments aligned with the Shia ideology, often barbed with **Loaded Language** and **Labeling** to cast the USA negatively. Concurrently, they portray an image of national strength and insight to generate support. Followers display allegiance to Khamenei, the Supreme Leader.

MEK's hashtags mainly highlight their organizational ambitions, such as referencing the "National Council of Resistance" to harness nationalistic sentiments to generate support for the organization. Moreover, the **Appeal to Authority** is explicitly visible in their hashtags, indicating allegiance to their leader, Maryam Rajavi.

The Monarchy's hashtags heavily rely on nationalistic feelings to evoke sentiments of national unity and revolution, with a strong emphasis on an everlasting monarchy as the nation's spirit. Following the path of Pahlavism is projected as the singular solution, implying the **Dictatorship** technique. Moreover, their hashtags effortlessly integrate **Loaded Language**, **Dictatorship**, **Name-Calling**, **Labeling**, and **Guilt-by-Association** techniques to denigrate their adversaries, casting them as corrupt and needing absolute dismissal.

A noticeable shift to extremism and radicalization is observed across all groups after the Mahsa period, as indicated by the prolific usage of "death to" in their hashtags. This suggests a call to action that goes beyond mere dissent, leaning towards a more violent resolution to confront the adversaries decisively. Significantly, the Islamic Republic shows a readiness for punitive actions against the protesters by endorsing executions and decisive response.

The tragic incident involving Mahsa Amini provides an interesting case study on how these groups strategically leverage hashtags centered around prominent events to amplify their messages. Each group uses the Mahsa Amini hashtag in combination with other hashtags that reinforce their agendas, ideologies, and calls for action. For instance, the Islamic Republic uses this opportunity to blame the Mahsa Amini movement for a terrorist attack, deploying the **Guilt-by-Association** technique.

These groups strategically use various propaganda techniques such as **Slogans**, **Flag-Waving**, **Appeal to Authority**, **Loaded Language**, **Labeling/Name-Calling**, **Dictatorship**, **Reductio ad Hitlerum/Guilt by Association**, and **Glittering Generalities** (virtue) to influence opinions and elicit action. Using the names of innocent victims lost in protests or detained after protests as **Loaded Language** is a critical strategy used by the MEK and Monarchy groups to incite strong emotional reactions and stir empathy, anger, and sorrow in their audiences. This tactic is also observed in the Islamic Republic group as they highlight the victims of the Shah Cheragh terrorist attack. In essence, these strategies help every group to maximize their influence, secure support, and discredit opposing voices in line with their respective ideologies and aspirations.

### 5.3.4 Trend Analysis

The dominant Bi-grams and Tri-grams of all the groups over the period After Mahsa are presented in Figures 5.14 and 5.15. The 'MEK' network, which represented 0.2% of the dataset, was excluded from the following analysis due to its limited size. The technique of Repetition is evident in the dominant discourse analysis. By repeatedly using certain phrases, these groups reinforce their message and make it more memorable for the audience.

**Monarchy** As revealed in Figures 5.14 and 5.15, the Monarchy group strategically employs several propaganda techniques. The term تها راه نجات (The only way to salvation) encapsulates the **Dictatorship** technique, presenting a binary solution to a multifaceted issue. Furthermore, تجزیه طلب (Separationist) combines the **Appeal To Fear** and **Labeling** techniques, creating an undertone of imminent division or fragmentation and can be used to label opponents as separationists. The Monarchy's discourse also champions territorial integrity with تمایت ارضی (Territorial Integrity), an intertwining of the **Appeal To Values** and **Appeal To Fear** techniques. Their narrative invokes patriotism through the rallying cry, (I would sacrifice my life for Iran), representing the **Flag Waving**, **Loaded Language**, and **Slogan** techniques. Moreover, the use of خاندان ایران‌پهلوی (Iran Making Pahlavi Dynasty) evokes the **Slogan** and **Causal Oversimplification** technique, suggesting that the dynasty singularly “created” or “crafted” Iran. Their choice of پرچم شیر و خورشید (The flag of lion and sun) is a poignant example of **Flag Waving** and historical symbolism. Lastly, terms like خفه شو (Shut up) succinctly exemplify the **Thought-terminating cliché** technique, aiming to stifle conversation and opposition.

**Islamic Republic** Illustrated in Figures 5.14 and 5.15, the Islamic Republic's discourse brandishes its own array of tactics. The emotive phrase جانم فدای رهبر (I would sacrifice my life for the leader) serves as a hallmark **Slogan** and **Loaded Language**, fostering allegiance to the regime's leadership. The reverential term مقام معظم رهبری (The supreme leader) harnesses the **Appeal to Authority** technique, asserting the leadership's paramountcy. The Republic's narrative also employs **Name Calling/Labeling** with disparaging labels like وطن فروش (Homeland seller). Prominent in their discourse is consistent Shia ideological appeals, with icons like امام زمان (Imam of Time/Mahdi) and امام حسین (Imam Hossein), signaling a robust **Appeal to Values**. The veneration of figures like قاسم سلیمانی, the IRGC commander, fuses both the **Appeal to Values** and the **Slogan** techniques.

**People** The ‘Unaffiliated People’ group observed two distinct spikes in conversation trends, one at the end of September 2022 and another in mid-January 2023. Specifically, the term “Men Homeland Prosperity” (M.H.P) experienced a sharp spike on September 28, while “Women Life Freedom” showed a more gradual increase, as illustrated in Figure 5.15. In mid-January, we can observe a Bi-gram “Reza Pahlavi” spike as shown in Figures 5.14. Smaller timeframes around these peaks were analyzed to gain a deeper understanding of these trends.

**The Emergence of Men Homeland Prosperity** N-gram analyses between September 21 and October 5, 2022, as reflected in Figures 5.17 and 5.16, confirm a peak of M.H.P on September 28. This peak coincides with the release of the Grammy Award-winning song بارای (“Baraye” translated to “For”) by Sherwin Hajipour on the 28th of September. This song, coined as the anthem of the movement<sup>5</sup>, is crafted from tweets<sup>6</sup> containing بارای and has a verse of بارای مرد مهمن آبادی (For Men Homeland Prosperity). By observing the people’s Bi-grams in Figure 5.16, a trend of بارای (For) is evident in بارای روزی (For Mahsa), بارای ایران (For Iran), and بارای میسا (For a day). This originates from a trend that proliferated during the second week of the movement as people took to Twitter to voice their reasons for protesting. Figure 5.20 showcases the top 20 n-grams containing بارای before Sherwin Hajipour’s song release. As the song’s lyrics are composed of actual tweets, the origination of M.H.P can be traced by finding the first tweet containing

<sup>5</sup><https://time.com/6253188/iran-grammys-shervin-hajipour/>

<sup>6</sup><https://www.washingtonpost.com/world/2022/10/04/iran-protests-song-shervin-hajipour-arrested/>

M.H.P. The first 20 tweets are presented in Table D.1, offering a systematic examination of its inception before the song's introduction.

A comprehensive analysis of the first 20 tweets featuring M.H.P reveals that it originates from members of the monarchy community, as determined by previous affiliation label assignments. 0.75% of the first 20 tweets and users come from members of the monarchy group.

- The first mention of M.H.P came from an influencer affiliated with Monarchy. This tweet also incorporated the *Mahsa Amini* hashtag and the phrase **Women.Life.Freedom**. Subsequently, multiple users with minor modifications spread this tweet in short periods.
- A subsequent tweet deviating from the earlier pattern states: “The only way to liberty is through a Monarchic Iran”. This claim falls under the **Dictatorship** propaganda technique, suggesting a singular solution when others are plausible and eliminating other options. The M.H.P slogan frequently appeared alongside *Women Life Freedom* in many of these tweets.
- A different tweet from an unaffiliated user was documented, indicating the dissemination and adoption of the M.H.P slogan among broader audience segments.
- A subsequent tweet, distinct from the previously observed patterns, was posted by the initial disseminator of the M.H.P adaptation. The tweet states: “This slogan belongs to ... communists and feminists. It’s beautiful but incomplete. The complete one is: Women, life, freedom. Men, homeland, prosperity.” This tweet connected “Women Life Freedom” with communism, thus employing the **Guilt By Association** propaganda technique. By suggesting that “Women Life Freedom” is “incomplete” without incorporating a masculine and nationalistic perspective, it introduces its counterpart, M.H.P.
- Another subsequent tweet, following the patterns previously noted, incorporated the phrase “King Happiness Prosperity” in conjunction with both “Women Life Freedom” and M.H.P. This specific combination, aiming to resonate with universal virtues and aspirations, employed the **Glittering Generalities (Virtue)** propaganda technique.

The tri-gram “Men Homeland Prosperity” bears a striking resemblance to the hashtag **شاه میهن آبادی** which translates to “King Homeland Prosperity.” This specific hashtag is labeled in Table B.6 under the Monarchy affiliation.

Moreover, during this week, as depicted in Figure 5.16, the Islamic Republic’s Bi-grams feature the term **تجزیه طلب** (Separationist), which aligns with the **Appeal To Fear** propaganda technique frequently used by the Monarchy group.

**Peak of the Bi-gram “Reza Pahlavi” on January 18, 2023** Figures 5.18 and 5.19 display the dominant N-grams for all communities between January 15 and January 29, 2023. On January 18, the Bi-gram “Reza Pahlavi” surged in the ‘People’ group’s discussions, maintaining its top position for the subsequent seven weeks as shown in Figure 5.14. The spike in discourse around “Reza Pahlavi” correlates to a petition launched on Change.org titled “Prince Reza Pahlavi is my representative.”<sup>7</sup> This link becomes evident with the prevalent Tri-gram **وکالت رضا پهلوی** (Attorneyship, Reza Pahlavi) in the people’s discussions. An in-depth analysis of the N-grams indicates that the ‘People’ group’s dialogue subsequently reflected themes and propaganda terms associated with the Monarchy community, notably around terms like “Separationism”, and “Territorial Integrity” in the first week after the start of the petition.

<sup>7</sup>The petition’s description states: “Considering the revolution happening in Iran and given that Prince Reza Pahlavi asked the people of Iran to give him the power of attorney to lead this movement, this petition allows the people of Iran to express their opinion.” <https://www.change.org/p/prince-reza-pahlavi-is-my-representative-c0fab7a1-2d92-4e8d-93c2-5f894a6e439b>

### **Addressing RQ3**

*What are the propaganda strategies and techniques these groupings use dominantly over time in their discourse, and what strategies have exerted the most significant influence on public discourse?*

Based on the dominant trend analysis methodology, several dominant propaganda strategies and techniques employed by various groups in their discourse over time can be discerned:

The Monarchy group predominantly utilized techniques such as **Dictatorship** through binary solutions, **Appeal To Fear** and **Labeling** to imply the threat of fragmentation, **Appeal To Values** to champion territorial integrity, and **Flag Waving**, **Loaded Language**, and **Slogan** to stir patriotism. This was evident with terms such as پرچم شیر و تجزیه طلب, تها راه نجات, and خورشید. The Islamic Republic's discourse was characterized by **Slogan** and **Loaded Language** to foster allegiance, **Appeal to Authority** to emphasize the leadership's paramountcy, and **Name Calling/Labeling** with disparaging terms. This group also made a significant **Appeal to Values** by referencing Shia icons. Meanwhile, the Unaffiliated People' group showed spikes in their discussions around terms like "Men Homeland Prosperity" and "Reza Pahlavi", particularly influenced by notable events like the release of Sherwin Hajipour's song and the launch of a Change.org petition. Furthermore, their discourse reflected elements from the Monarchy's propaganda techniques, especially around terms like "Separationism" and "Territorial Integrity" post the petition launch. Overall, the strategies that exerted the most significant influence on public discourse, especially within the Unaffiliated People' group, seemed to be those employed by the Monarchy, notably the techniques of **Dictatorship**, **Appeal To Fear**, **Appeal To Values**, and **Flag Waving**.

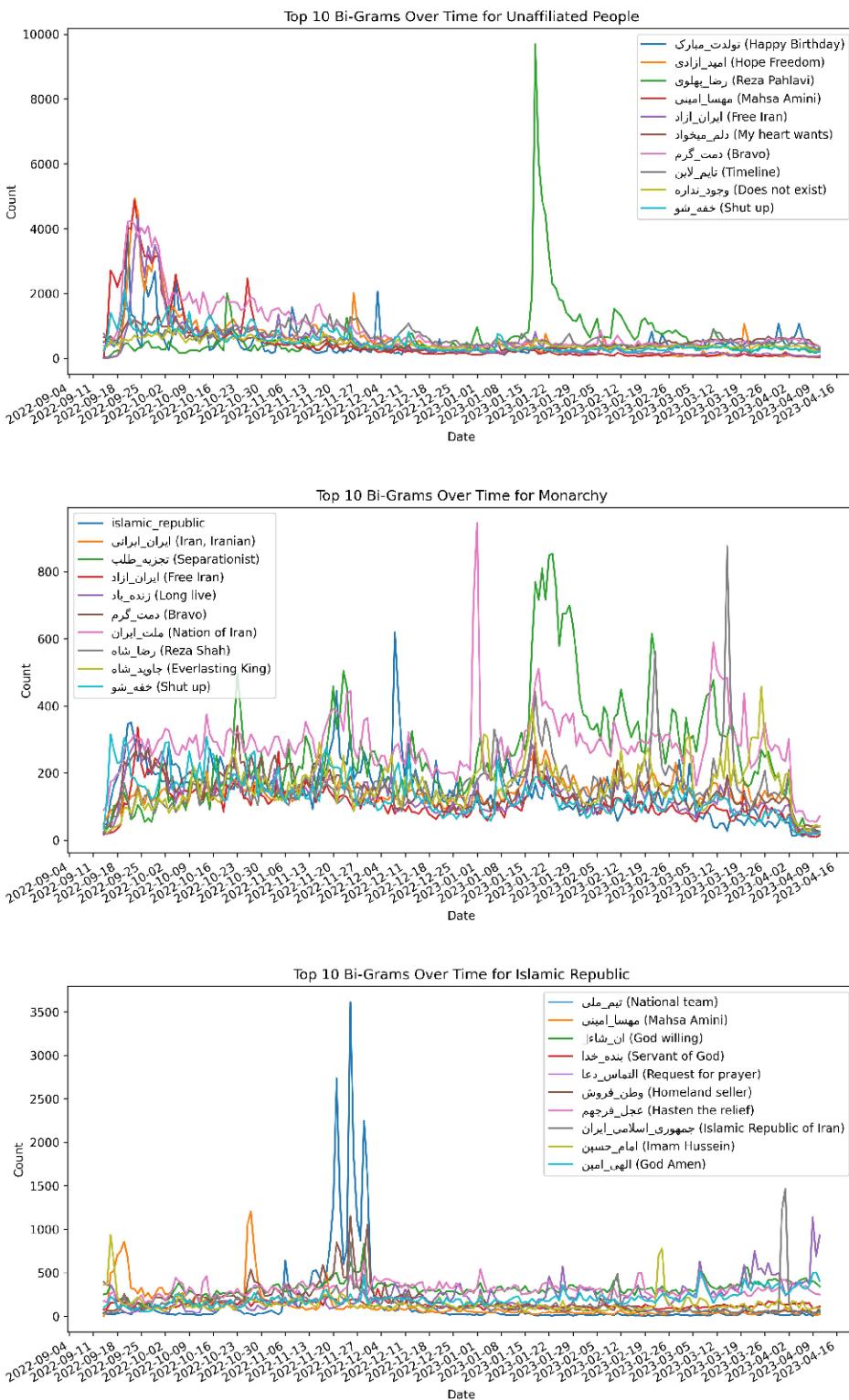


Figure 5.14: Top Bi-grams for Unaffiliated People, Monarchy and Islamic Republic after Mahsa. The Y-axis represents the number of times the Bi-Gram was used within each community.

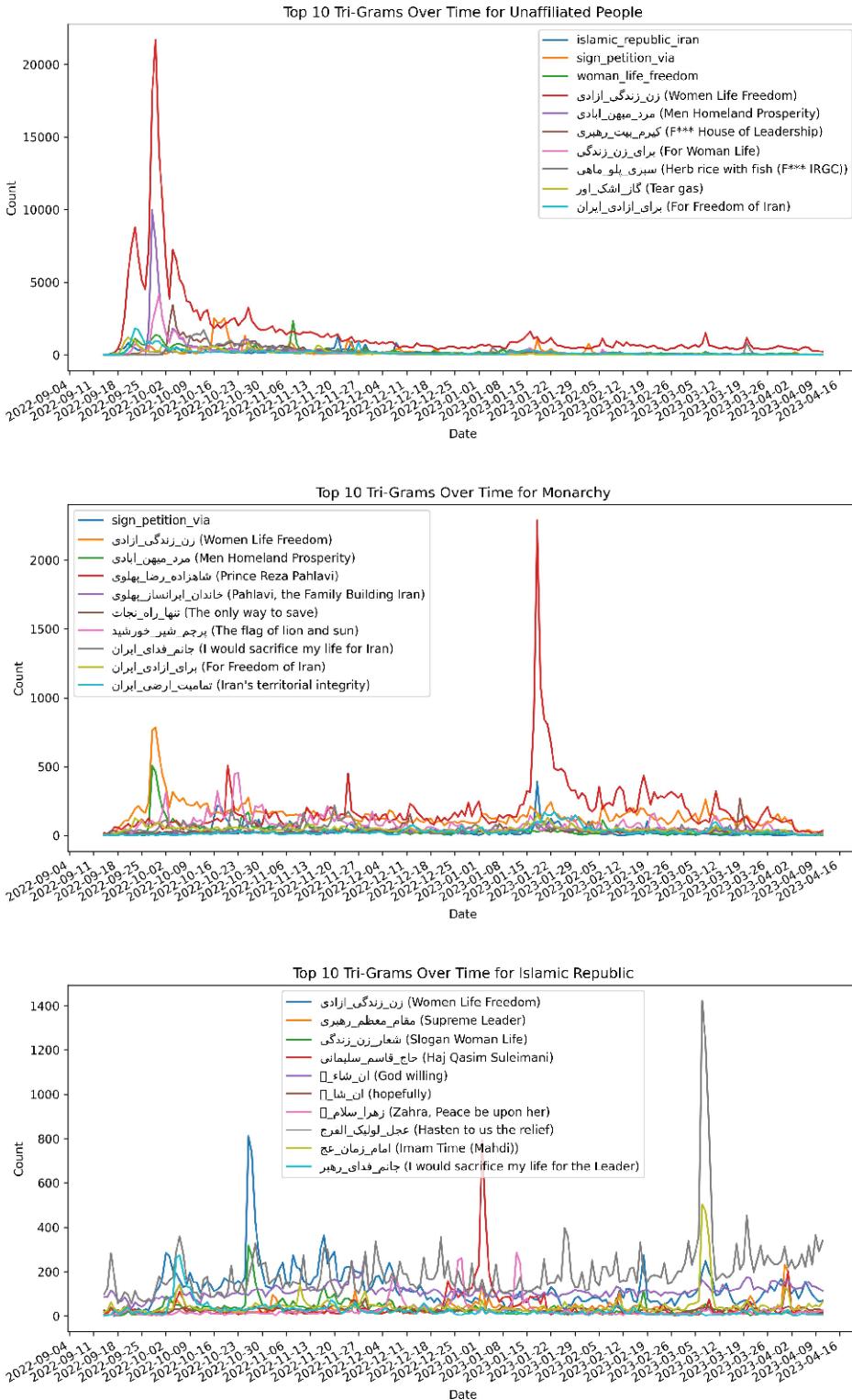


Figure 5.15: Top 10 Tri-grams for Unaffiliated People, Monarchy and Islamic Republic after Mahsa. The Y-axis represents the number of times the Tri-Gram was used within each community.

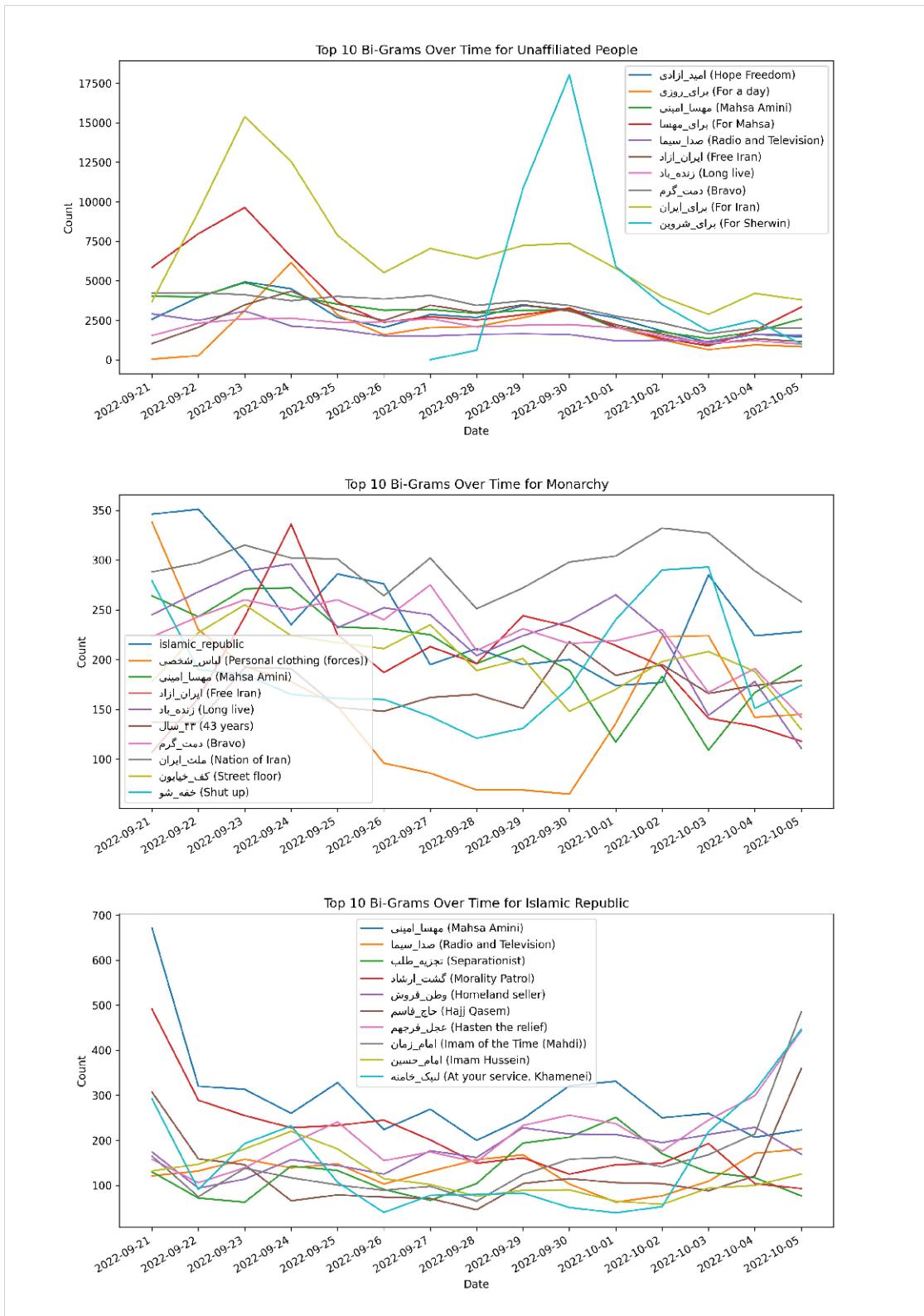


Figure 5.16: Top 10 Bi-Grams from 2022-09-21 to 2022-10-5. The Y-axis represents the number of times the Bi-Gram was used within each community.

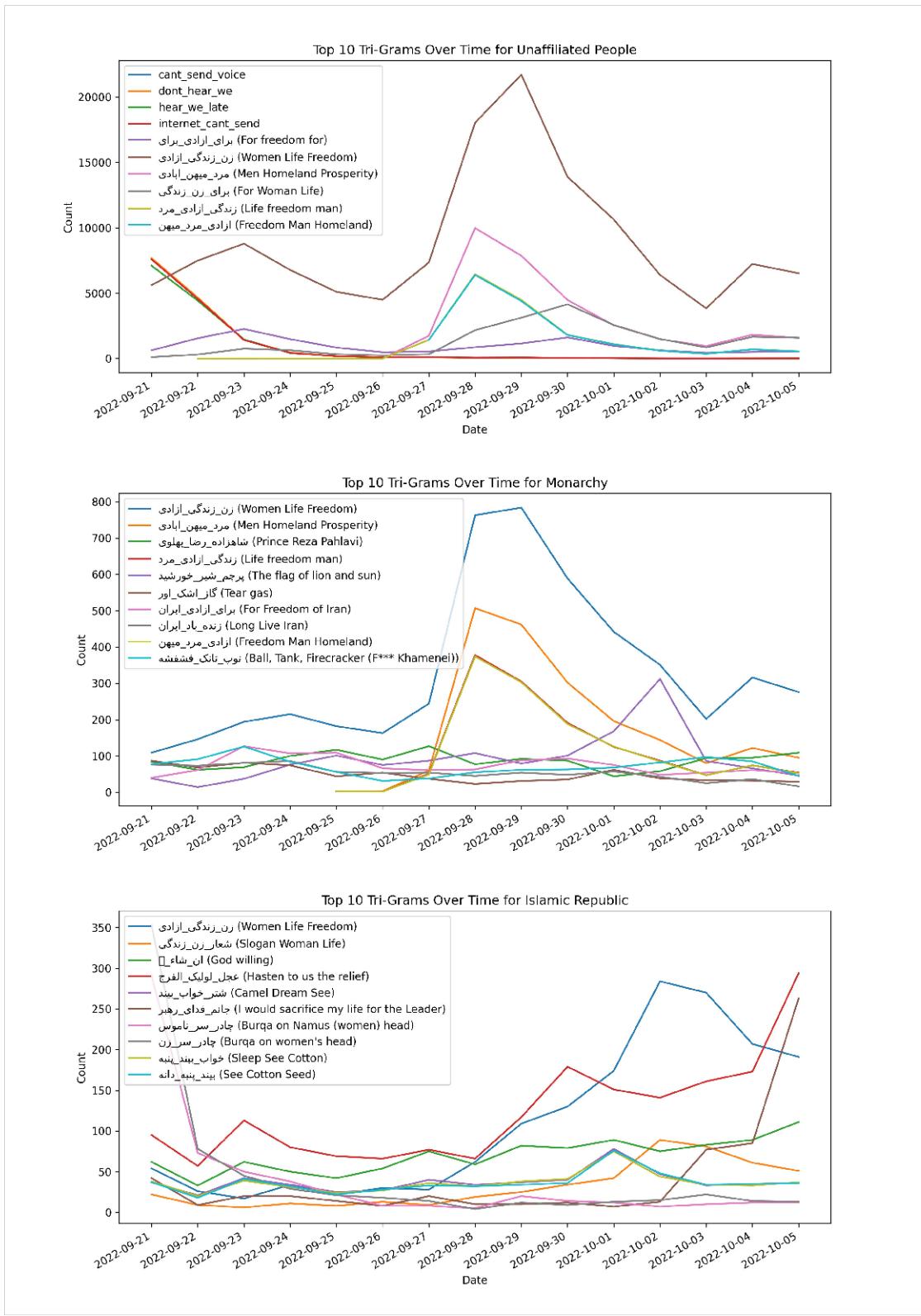


Figure 5.17: Top 10 Tri-Grams from 2022-09-21 to 2022-10-05. The Y-axis represents the number of times the Tri-Gram was used within each community.

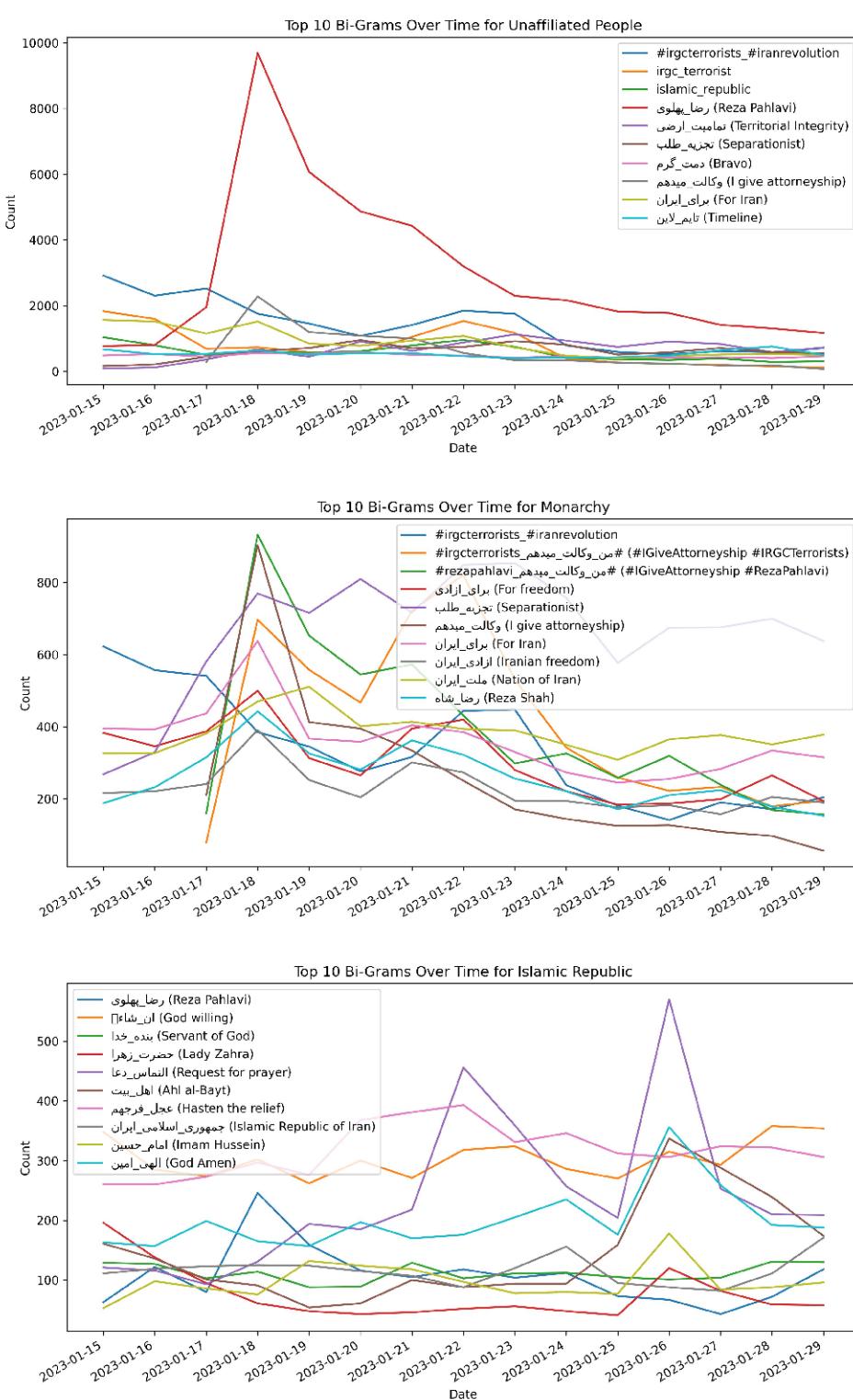


Figure 5.18: Top 10 Bi-Grams from 2023-01-15 to 2023-01-29. The Y-axis represents the number of times the Bi-Gram was used within each community.

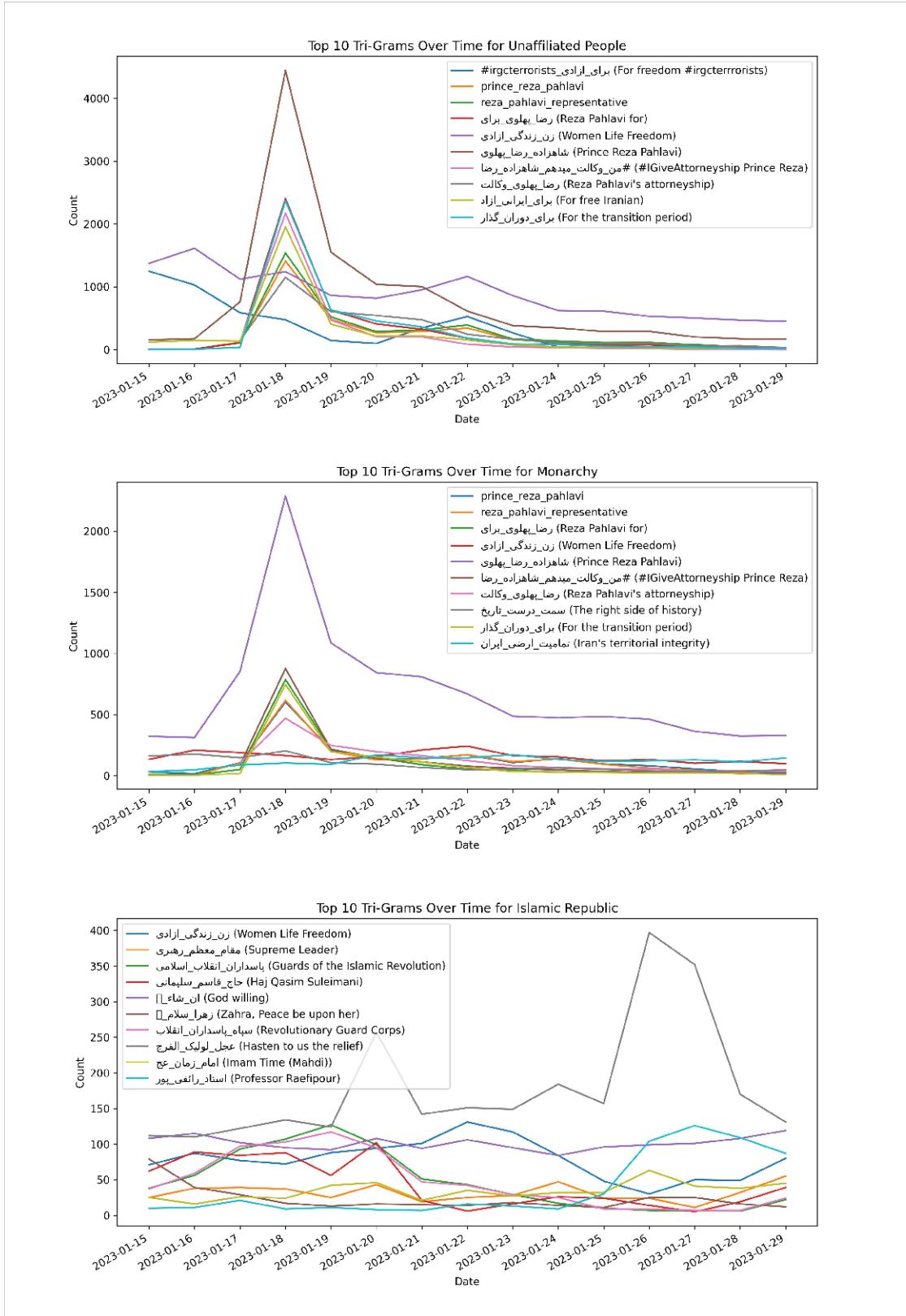


Figure 5.19: Top 10 Tri-Grams for 2023-01-15 to 2023-01-29. The Y-axis represents the number of times the Bi-Gram was used within each community.

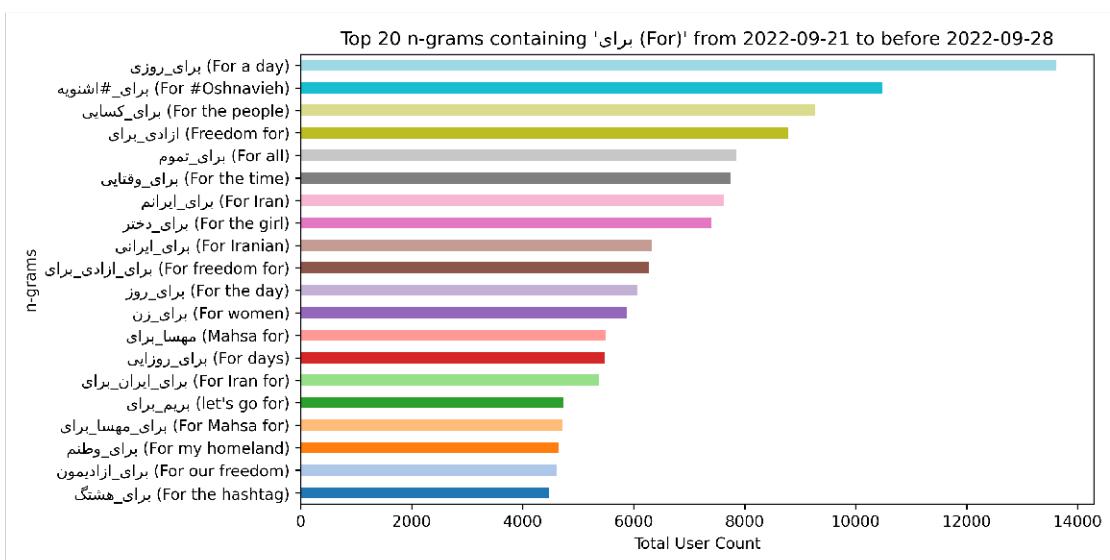


Figure 5.20: Top 20 most frequently used n-grams containing بُرای (For) prior to the release of the Baraye song by the number of unique users using them.

## 5.4 Deadends

### 5.4.1 Labeling Twitter Data

A novel annotation methodology was pursued to delve deeper into the nuances of networked and NLP propaganda detection. Unlike traditional labeling methods which often focus on merely identifying propaganda techniques, this methodology also aimed to discern the positive targets, negative targets, and the overarching objectives of the propaganda. A custom web application was developed for labeling the Twitter dataset, which communicated with OpenAI's ChatGPT API for semantic analysis and labeling assistance. The labeling process utilized the following algorithm:

1. **Initialization with Chain of Thought Prompt:** The ChatGPT model is initially provided with a "Chain of Thought" prompt. This prompt includes the context of the study, an explanation of the propaganda labels to be identified.
2. **Processing and Labeling Tweets:** Each tweet is fed to the ChatGPT model, which is requested to output a JSON file with the following fields:
  - **translation:** English translation of the tweet.
  - **explanation:** Explanation of any propaganda techniques identified and the predicted labels.
  - **labels:** Propaganda labels, if any.
  - **positive\_target:** Targets that the tweet supports or promotes.
  - **negative\_target:** Targets that the tweet undermines or criticizes.
  - **objectives:** The objectives or aims of the propaganda, if identifiable.

The ChatGPT API is called repeatedly until a valid JSON is returned. If the output is not a valid JSON, a prompt informing the model that the output was not valid and a request for a valid JSON is sent back.

3. **Automatic Re-evaluation by ChatGPT:** Once a valid JSON response is received, an automatic re-evaluation prompt is sent to ChatGPT. This prompt asks ChatGPT to self-criticize and revise any potential mistakes in the response.
4. **Human Validation and Revision:** In the web application, a human labeler reviews the ChatGPT model's response, including any revisions made during the automatic re-evaluation. The labeler has the option to validate the response as correct, send a prompt to the model to correct any issues, or manually correct the issues themselves.
5. **Storing the Validated Data:** Once the human labeler is satisfied with the labeling and the ChatGPT model has undergone re-evaluation, the validated data is stored in JSON format for further analysis.

The outlined methodology was ultimately not pursued further due to a critical flaw in its design. During a three-week period, more than 350 tweets were labeled using this approach. However, this process necessitates a consensus among multiple data labelers. The lack of resources posed significant challenges, causing the abandonment of this approach.

# Chapter 6

## Explaining Propaganda Techniques With LLMs

Persuasion and propaganda, while distinct in their goals and application intricacies, share overlapping techniques and detection mechanisms [88]. The words “propaganda” and “persuasion” are often used synonymously in academic writings and daily conversations [60]. While propaganda utilizes persuasive tactics, its objective sets it apart from mere persuasion. This chapter delves into the background of Natural Language Processing (NLP) and Large Language Models (LLMs). Subsequent sections introduce background information in explainable AI, Explainable AI with natural language, followed by related work in explainable propaganda detection and automated AI-generated explanations. Details regarding the dataset and the methodology for generating explanations using ChatGPT are presented. Results focus on the model’s accuracy in label prediction while it was trained on generating explanations as well as labels. Additionally, a methodology for the development and assessment of PropagandaTeacherGPT is proposed.

### 6.1 Background Information and Related Work

Building on the foundational concepts of Natural Language Processing (NLP) and Large Language Models (LLMs), this section delves deeper into the principles of Explainable AI, especially its integration with natural language. Further, a review is conducted on related work in the domain of explainable propaganda and the emerging realm of AI-generated explanations.

#### 6.1.1 Natural Language Processing

Natural Language Processing (NLP) is a critical component of artificial intelligence that interacts with computers and human language. It seeks to allow machines to interpret, generate, and respond to human language in a manner that is meaningful. [61]

##### Historical Development of NLP

- **1950s:** The inception of NLP was marked by foundational experiments like the Georgetown-IBM project, focused on machine translation. While revolutionary for its time, this approach produced rudimentary translations that lacked linguistic nuance. [56]

- **1960s-1970s:** The field transitioned to rule-based systems where linguists set explicit rules for language processing by computers. Although more structured, these systems struggled with the flexibility and variability inherent to human language. [126]
- **Late 20th century:** With advancements in computational power and the availability of larger datasets, NLP began adopting statistical methods such as Hidden Markov Models and Bayesian networks. These models operated by identifying and learning patterns from vast amounts of textual data, bypassing the need for rigid predefined rules. [61]
- **21st century:**
  - **Early 21st century:** This era saw the rise of statistical models and traditional machine learning models. Support Vector Machines (SVMs) and Naive Bayes were employed for tasks like text classification, sentiment analysis, and spam detection. Additionally, word embeddings, namely Word2Vec and GloVe, revolutionized how semantic relationships between words were captured.
  - **Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Networks:** With the resurgence of neural networks, RNNs and their advanced variant, LSTMs, became prominent for sequential tasks. They made significant strides in applications like machine translation, sentiment analysis, and text generation due to their ability to remember long-term dependencies.
  - **Transformers and Beyond:** The Transformer architecture, introduced by Vaswani et al. [119], heralded a new era in NLP. Models built on this architecture, such as BERT [38], GPT [90], and subsequent large language models [17], have revolutionized various NLP tasks. These models are characterized by their ability to learn from extensive amounts of data, and their proficiency in grasping linguistic nuances.

### 6.1.2 Large Language Models

Large Language Models are AI models trained on extensive text data, enabling them to generate human-like text based on the inputs they receive. These models are based on machine learning architecture known as transformer architecture [119], allowing them to consider a given text's context when generating new text. [17, 90]

One of the key features of LLMs is their ability to perform few-shot learning. This means they can learn to perform a new task based on only a few examples, a feature that has been demonstrated in various applications, such as translation, question answering, and summarization. [17]

The development of LLMs has been driven by increasing computational power and the availability of vast amounts of text data. The size of these models is often measured in parameters, which are the aspects of the model that are learned from the training data. LLMs can range from hundreds of millions to hundreds of billions of parameters. [90]

LLMs, such as GPT-3 [17], LLaMa[113], and GPT-4 [86], have shown impressive performance in various natural language understanding and generation tasks. However, they also present challenges, including the risk of generating inappropriate or biased content, their black-box nature that makes their decision-making process opaque [132], and the large computational resources required to train and run them.

Despite these challenges, LLMs represent a significant advancement in AI research and have a broad range of potential applications, from customer service chatbots to personal assistants, content creation, and beyond. [86]

### **LLaMA: Open and Efficient Foundation Language Models**

LLaMA, an acronym for Large Language Models, is an innovative collection of AI language models developed by Meta AI. The models range from 7 billion (7B) to 65 billion (65B) parameters, showcasing the current trend towards larger models in AI research. LLaMA models are trained on trillions of tokens, demonstrating the feasibility of training state-of-the-art models using only publicly available datasets, thereby avoiding the need for proprietary and inaccessible datasets. The research shows that LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla-70B and PaLM-540B. Meta AI has released all these models to the research community, emphasizing their commitment to open science and collaboration. [113]

The scaling of LLaMA models is inspired by the work of Hoffmann et al [54], which indicates that smaller models trained on more data often perform better than larger models given the same compute budget. This is particularly important when considering the inference budget, which becomes crucial when deploying a language model at scale. [113]

The LLaMA series of models are trained to achieve the best possible performance at various inference budgets, by training on more tokens than what is typically used. This approach demonstrates the potential for improving performance and efficiency in large language models. For instance, LLaMA-13B, despite being significantly smaller, outperforms GPT-3 on most benchmarks, and can be run on a single GPU, which facilitates access and study of LLMs. [113]

**LLaMA 2** Following the introduction of the LLaMA series, Meta AI announced an update version LLaMA 2. The pretrained models in this version were trained on an expansive dataset of 2 trillion tokens, marking a significant augmentation from the previous series. Moreover, the context length for LLaMA 2 models was doubled, allowing for enhanced comprehension of lengthier language sequences. [114]

LLaMA 2 was trained using a diverse mix of publicly accessible data, with the pretraining corpus witnessing an increase of 40% in size. An innovation in this iteration was the adoption of grouped-query attention [3]. Meta AI has released variants of LLaMA 2 with 7B, 13B, and 70B parameters for public usage. Furthermore, LLaMA 2-Chat, a specialized model variant of LLaMA 2, was introduced to cater to dialogue-based applications. This model underwent fine-tuning processes over 1 million human annotations to improve its performance in dialogue scenarios. The available variants of LLaMA 2-Chat encompass models with 7B, 13B, and 70B parameters. [114]

#### **6.1.3 Explainable AI**

Explainable AI (XAI) emphasizes understanding the decision-making processes of AI models [6]. This field aims for transparency and interpretability of machine learning models, facilitating human comprehension and trust. Advanced machine learning models, like neural networks, often operate as "black boxes", with inputs processed through multiple hidden layers to produce an output [50]. The intricate mechanisms within these models pose challenges for human understanding. XAI's importance is amplified by real-world implications. In sectors like healthcare, finance, and autonomous vehicles, comprehending AI decisions is essential for safety, regulatory compliance, and ethical considerations [20, 101].

XAI methodologies fall into two categories: Intrinsic and Post-hoc interpretability.

1. Intrinsic Interpretability Intrinsic interpretability involves the use of models that are inherently interpretable due to their simplicity or transparency. These models have a clear and understandable decision-making process, allowing for a straightforward interpretation

of the results. Examples of intrinsically interpretable models include linear regression and decision trees.

- **Linear regression:** In a linear regression model, each feature is assigned a weight or coefficient. The higher the absolute value of the coefficient, the more important the feature is for the prediction. This clearly explains how each feature impacts the model's output.
  - **Decision trees:** Decision trees split the data based on the value of certain features. The structure of the tree, with its splits and branches, gives a clear and intuitive understanding of how the model makes its predictions.
2. Post-hoc Interpretability Post-hoc interpretability, on the other hand, involves the application of techniques that explain the decisions of already trained models. This is particularly useful for complex models such as neural networks, where the decision-making process is opaque and difficult to interpret. Some of the popular techniques used for post-hoc interpretability include:
- **Feature Importance:** This technique assigns an importance score to each feature based on how much it contributes to the model's prediction. This can help to understand which features are most influential in the model's decisions. [97]
  - **Saliency Maps:** These are used primarily for image classification tasks. A saliency map highlights the areas in the input image that were most relevant for the model's prediction, giving a visual representation of what the model "sees". [105]
  - **Surrogate Models:** These are simpler, interpretable models (such as linear regression or decision trees) that are trained to mimic the predictions of the complex model. While they do not provide a perfect explanation, they offer an approximation of the model's decision-making process. [84]
  - **Counterfactual Explanations:** This method provides explanations by showing what changes in the input data would have led to a different prediction. This helps to understand the boundaries and decision rules of the model. [120]
  - **LIME (Local Interpretable Model-Agnostic Explanations):** LIME is a technique that explains the predictions of any classifier by learning an interpretable model locally around the prediction. [97]
  - **SHAP (SHapley Additive exPlanations):** SHAP is a unified measure of feature importance that assigns each feature an importance value for a particular prediction. It is based on game theory and involves calculating the contribution of each feature to the prediction for a specific instance. [75]

**Evaluation of Interpretability:** Evaluation of these explainability methods is also an important aspect. A common approach is human-grounded evaluations, where human evaluators judge the usefulness of explanations. This can be done through user studies or expert reviews. Another approach is functionally-grounded evaluations, which do not require human judgement. Here, the quality of an explanation is assessed by its ability to help achieve some functional requirement, like improving model performance when a particular feature indicated by the explanation is altered. [6, 72]

#### 6.1.4 Explaining AI in Natural Language

Explainable AI has seen tremendous growth recently due to the increasing need for transparency and accountability in AI systems. Although several methods have been proposed for enhancing the interpretability of machine learning models, there is an emergent trend towards generating natural language explanations of model decisions.

In this vein, Camburu et al. [19] have attempted to explain neural network predictions by providing textual explanations alongside predictions. This methodology presents an interpretable model that offers explanations in human-readable text for every prediction it makes. Another noteworthy contribution is the work by Hendricks et al. [52], where they proposed a model that generates natural language explanations for its predictions in the context of image classification.

Lei et al. [71] proposed a method to rationalize model decisions by selecting portions of input text that are critical to the prediction, along with a natural language justification. This allows users to understand how the model is processing the information.

Reiter [96] explores the challenges in natural language generation for Explainable AI. He highlights several complexities, including ensuring coherence in the logical structure of explanations, achieving precision without sacrificing conciseness, maintaining awareness of the user's background and the explanation's complexity, verifying that the explanation accurately reflects the model's behavior, and carefully navigating ethical considerations such as biases in the generated explanations. These challenges underscore the intricacy of developing Natural Language Generation methods that can truly make AI models transparent and understandable to end-users. Reiter's insights guide ongoing research in this domain and emphasize the need for robust methodologies that consider both technical and human-centered aspects.

Lastly, the work of Miller et al. [82] emphasizes the importance of both automated and human-generated explanations in building trust in AI systems, providing a framework that may be used in conjunction with natural language explanations.

These techniques highlight the potential for using AI models to produce understandable and intuitive explanations of their predictions. They represent a step towards making AI models more transparent and accountable, aligning them with human cognition and ethical considerations.

#### 6.1.5 Related Work in Explainable Propaganda Detection

Explainable propaganda detection represents an emerging direction in the field, focusing on identifying the presence of propaganda and elucidating the underlying techniques and mechanisms. Recent work has started combining traditional classification models with interpretability methods to reveal insights into specific propaganda techniques employed within a text.

**Linguistic and Game-Theoretic Approach to Propaganda Detection** Barfar [8] leveraged SHAP from cooperative game theory to interpret the decisions made by the LightGBM gradient boosting decision tree algorithm. The interpretation was framed around the idea of a cooperative game where each linguistic feature was treated as a player collaborating in the game. The Shapley value assigned a value to each feature (or player) based on its contribution to the prediction. Specifically, it calculated how much a particular feature, when added to different subsets of features, changes the prediction score. This gave a quantifiable measure of the importance of each feature in the overall prediction.

The interpretability arises from the ability to break down the prediction into the sum of contributions from individual features. By isolating the impact of each feature, the Shapley value helped elucidate how different linguistic aspects of an article contribute to its classification as propaganda. This granular understanding provides insights into what specific characteristics

of the text are most indicative of propaganda, translating the complex decision-making process of the model into terms that can be understood and analyzed.

**Interpretable Propaganda Detection Using Semantic and Syntactic Features** Yu et al. [131] proposed a method to detect propaganda techniques in sentences by leveraging a combination of semantical, syntactical, and structural features. The method combines:

- The relative position of the sentence within an article ( $f_{rp}$ ).
- Topic similarity and stance with respect to the article's title ( $f_{sim}$  and  $f_{stn}$ ).
- Syntactic structures from parsed sentences and sentiment indicators, as well as document-level propaganda predictions ( $f_{dp}$ ,  $f_{sent}$ , and  $f_{doc}$ ).

Central to the interpretability of their approach is a covariance matrix that maps the correlation between these features and 18 recognized propaganda techniques. This method offers a clear view of the relationships, allowing a quick determination of features strongly correlated with specific propaganda techniques. It explains why a particular sentence or article is flagged as propagandistic. When a sentence displays high values for features closely associated with a specific propaganda technique, it gives a reason for its classification. For instance, if the "Sentiment" feature ( $f_{stn}$ ) has high covariance with the "Loaded Language" technique, it indicates that sentences with strong sentiment are likely to be using the "Loaded Language" technique, and this feature can interpret the model's decision of detecting this technique.

### 6.1.6 Related Work in AI Generated Explanations

Bertaglia et al. [11] highlights the potential of using AI-generated explanations, particularly from ChatGPT, to enhance the annotation process, especially in the context of influencer marketing regulations. Recognizing the increasing global emphasis on transparency in influencer marketing, the study leverages AI capabilities to meet the challenges posed by the vastness of the influencer market.

The primary concern in automating the detection of sponsored content is the inconsistency in human data annotations. Low inter-annotator agreement can diminish the reliability of the models trained on this data. Bertaglia et al. proposed the incorporation of ChatGPT in the annotation process to furnish annotators with relevant features and concise explanations. This approach showcased a consistent improvement in inter-annotator agreement and the accuracy of annotations, as validated by their experiments.

From the their findings, two primary observations stands out:

1. **Annotation Quality Improvement:** The inclusion of explanations led to a notable increase in inter-annotator agreement metrics with a 17.20% in absolute agreement and a 3.45% improvement in detecting disclosed sponsored posts. This suggests that the introduction of AI-backed explanations not only facilitates more accurate identification of sponsored content but also stabilizes the annotation process, reducing disagreements between annotators.
2. **Positive User Experience:** Feedback from annotators reflected a generally positive experience. Most found the explanations to be both helpful and accurate, which in turn bolstered their confidence in making decisions. The consistent and dependable nature of these explanations can be instrumental in refining the data labelling process and supporting decision-making in contexts where regulatory compliance is essential.

However, despite its promising results, the study did acknowledge certain limitations. A primary concern is the potential for bias introduced by the explanations. If annotators disproportionately rely on specific cues presented by ChatGPT, it might inadvertently shape their annotations in a particular direction. For instance, the group that utilized both the traditional annotation setting and the ChatGPT-augmented setting displayed a significant alignment with GPT predictions compared to the proportion of detected sponsored posts. This could suggest an undue influence of the AI on human annotators.

## 6.2 Dataset

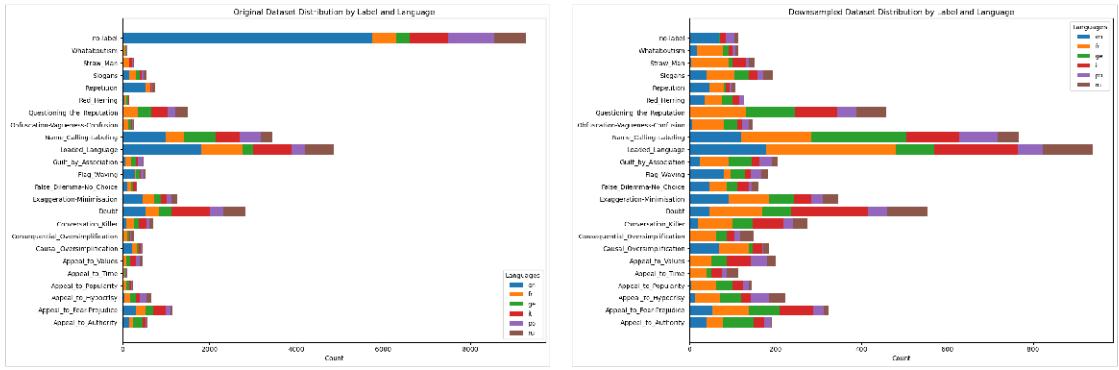
The research utilizes the SemEval 2023 Task 3 dataset [89] hereafter called PTD (Persuasion Technique Detection) dataset. Within this dataset, three distinct subtasks are offered, focusing on news articles in six languages: English, French, German, Italian, Polish, and Russian. For this study, the focus is primarily on the third subtask, Persuasion Techniques Detection.

The Persuasion Techniques Detection subtask involves identifying persuasion techniques in every paragraph of a news article. It is a multi-label task at the paragraph level, classifying 23 persuasion techniques. Notably, the methodology and annotations in the PTD dataset are extensions of a previous SemEval 2021 contest that concentrated on propaganda detection [30]. Propaganda is often marked by its intention to disseminate biased or skewed information. Persuasion, on the other hand, is a broader term encompassing all attempts to influence people's beliefs, attitudes, intentions, motivations, or behaviors. It's a neutral term that doesn't necessarily imply misinformation or malintent. While all propaganda can be classified as persuasive, not all persuasion falls under propaganda, the distinguishing factor frequently resides in the intent and the technique [88].

Below are the categorized persuasion techniques defined by Piskorski et al. [89]:

- **Attack on Reputation:** The argument targets the participant's personality, experience, or deeds to question or undermine credibility. This can also apply to groups, organizations, objects, or activities.
  - *Name Calling or Labelling:* Using specific labels to characterize an individual, group, object, or activity, often in a derogatory manner, but sometimes using terms the audience views positively.
  - *Guilt by Association:* Criticizing someone or something by linking them with another group, action, or idea that is viewed negatively by the audience.
  - *Casting Doubt:* Challenging the integrity or characteristics of a person or entity to cast suspicion on their overall trustworthiness or value.
  - *Appeal to Hypocrisy:* Attacking the subject by highlighting their inconsistencies or hypocritical behavior.
  - *Questioning the Reputation:* Targeting an entity by making negative assertions about them, particularly attacking their character and moral integrity, rather than focusing on the primary topic.
- **Justification:** Comprises a statement and an explanation or appeal, where the latter justifies or supports the statement.
  - *Flag Waving:* Endorsing a concept by emphasizing a group's pride or the benefits for that group.

- *Appeal to Authority*: Lending credence to a statement or idea by attributing it to a recognized authoritative figure or entity.
  - *Appeal to Popularity*: Supporting an argument by claiming it is the popular view or that the majority holds this perspective.
  - *Appeal to Values*: Giving importance to an idea by associating it with values cherished by the audience.
  - *Appeal to Fear, Prejudice*: Advancing or rejecting a notion based on eliciting fear or prejudice in the audience.
- **Simplification:** Excessively simplifies a problem, usually regarding the cause, consequence, or existence of choices.
  - *Causal Oversimplification*: Attributing a situation to a singular cause when multiple factors may be at play.
  - *False Dilemma or No Choice*: Presenting only two possible options when more exist. At its peak, it prescribes a specific course of action, negating all alternatives.
  - *Consequentialist Oversimplification*: Describing a chain reaction from an initial event that leads to increasingly unlikely outcomes, often in a dramatic manner.
- **Distraction:** Diverts focus away from the main topic or argument to distract the reader.
  - *Strawman*: Creating a distorted version of an opponent's argument to easily refute it, while avoiding the actual issue.
  - *Red Herring*: Introducing an unrelated or irrelevant topic to divert the audience's attention from the main discussion.
  - *Whataboutism*: Challenging an opponent's stance by pointing out their perceived hypocrisy, without addressing their original argument.
- **Call:** Not an argument, but encouragement to act or think in a particular way.
  - *Slogans*: A memorable phrase, often with emotional resonance, that can include categorization or generalizations.
  - *Conversation Killer*: Terms that stifle in-depth discussion or critical analysis on a subject.
  - *Appeal to Time*: An argument built around the notion that the moment is right for a specific action or change.
- **Manipulative Wording:** Specific language is used to impact the reader emotionally, containing non-neutral, confusing, exaggerating, or loaded words or phrases.
  - *Loaded Language*: Employing words with strong emotional connotations to sway audience opinion.
  - *Obfuscation, Intentional Vagueness, Confusion*: Using intentionally unclear or ambiguous language, allowing for varied audience interpretations.
  - *Exaggeration or Minimisation*: Amplifying or downplaying something beyond its actual scale or significance.
  - *Repetition*: Reiterating a phrase or idea with the intent of instilling it in the audience's mind.



(a) Original label distribution of PTD dataset.

(b) Downsampled PTD label distribution.

Figure 6.1: Original label distributions of the PTD dataset.

## 6.3 Methodology

This section delineates the methodology employed for this section of the study, encompassing four main facets: first, the process of augmenting the PTD dataset with explanations with ChatGPT; second, the curation of the training dataset; third the experimental configurations for model training; and finally, the method used for evaluation of the accuracy of the label prediction.

### 6.3.1 Augmenting PTD with GPT3.5 Explanations

The main objective of this section is to address the research gap in propaganda technique classification by developing a method that the general public can easily understand and is not language dependent. Unlike traditional approaches that rely on specific linguistic feature engineering, the goal is to devise a language-agnostic methodology.

The PTD dataset primarily evaluates propaganda from various articles. Since our initial focus was to fine-tune a model trained on PTD to analyze propaganda on Persian Twitter, each paragraph was considered as an independent input, necessitating special handling of specific labels of “Repetition”. Paragraphs containing the label “Repetition” were included only if the repetition occurred within the paragraph itself, or if a nearby paragraph (up to 2 surrounding paragraphs) contained the repetition. In cases where adjacent paragraphs exhibited repetition, they were concatenated as a singular paragraph. Otherwise, the paragraph was ignored.

The PTD comes pre-labeled, but to enhance its value for training machine learning models, GPT3.5 were used to generate explanations for each label. This approach functions as a method of data augmentation, improving the representational diversity of the explanations and potentially increasing the performance of models trained on this data. Additionally, by this type of data augmentation, the model learns to explain its predictions.

- **Contextualizing GPT3.5 with Chain of Thought Prompt:** The information is explained using two distinct prompts, one focusing on objectives and targets<sup>1</sup> and the other focusing more on general explanations.

<sup>1</sup>The emphasis on objectives and targets stems from the idea of fine-tuning a model trained on this task on annotations from the methodology detailed in the 5.4 section, where the annotations identified not only propaganda techniques but also their respective objectives and targets.

1. *Explanation-Centric Prompt:*

Instructions:

Read the provided JSON example containing the following fields: "id", "text", and "labels".  
Check if there are any propaganda labels present in the "labels" field.  
If there are propaganda labels:  
a. Generate an explanation for each propaganda technique present in the text.  
b. Formulate a comprehensive explanation of how each propaganda technique is employed in the text.  
c. Include the explanation for each propaganda label in the output JSON.  
If there are no propaganda labels:  
a. Analyze the content of the text to provide an in-depth explanation of the elements within the text that support the absence of propaganda.  
Return the output JSON object as the result.

2. *Objective-Centric Explanation Prompt:*

Instructions:

Read the provided JSON example containing the following fields: "id", "text", and "labels".  
Check if there are any propaganda labels present in the "labels" field.  
If there are propaganda labels:  
a. Generate an explanation for each propaganda technique present in the text.  
b. Formulate a comprehensive explanation or provide exemplification of how each propaganda technique is employed in the text, including the identified targets and objectives of the propaganda.  
c. Include the explanation, targets, and objectives for each propaganda label in the output JSON.  
If there are no propaganda labels:  
a. Analyze the content of the text to explain why no specific propaganda labels were identified.

Format the output as a JSON object with the following fields:  
a. "id": The identifier for the example.  
b. "labels": A list of identified labels of propaganda, if any.  
c. "explanation": A explanation section describing or exemplifying how each propaganda technique is employed in the text, including the identified targets and objectives of the propaganda, and the explanations. If there are no labels, provide an explanation and the rationale for the absence of propaganda techniques.

```

Return the output JSON object as the result.

Text: '''{text}'''

Your JSON Reply:

```

- **Multiple Explanations:** To ensure diversity and richness in the generated explanations, multiple explanations for the same paragraph were obtained using GPT3.5. Paragraphs from the PTD dataset were randomly selected for explanation generation.

By augmenting the PTD, we gain an additional layer of information for each label, offering further insight into how propaganda techniques manifest within the text. This data enrichment could also facilitate the development of explainable classification systems, where models offer rationales for their labeling decisions.

### 6.3.2 Curation of Training Data

For training, three types of prompts were used to optimize the model. The prompts are:

```

instruction_only_labels = """
Role: Propaganda Classifier
Task:
    Analyze text for propaganda techniques.
    Return a JSON object with "labels".
"""

instruction_explanation = """
Role: Propaganda Analyzer
Task:
    Analyze text for propaganda techniques.
    If found, create a comprehensive and in-depth explanation of the
        technique(s).
    If not, create a comprehensive and in-depth explanation of why
        the text is not propaganda.
    Return a JSON object with "labels" and "explanation".
"""

instruction_objective = """
Role: Propaganda Analysis
Task:
    Analyze text for propaganda techniques.
    If found, create an explanation including objectives and targets.
    If not, explain why the text is not propaganda.
    Return a JSON object with "labels" and "explanation".
"""

```

In the JSON structure, the ‘labels’ key had priority, reflecting its primary role in the classification task.

**Downsampling of Datasets** Addressing class imbalances in multi-lingual, multi-label datasets is especially challenging [51]. Given these complexities, a straightforward multi-class random downsampling approach was employed. Entries from each class were randomly selected to match the count of the least populous class. While this method brings each class closer to the size of the smallest one, it does not ensure an absolute balance across all classes.

### 6.3.3 Experimental Setting

The research utilized the chat-optimized variant of LLaMA 2 for fine-tuning. This training took place using one Tesla V100 - 32GB GPU. The context length of 1024 tokens was chosen for training efficiency as each individual input and output paragraph in the training set is less than 1024 tokens. The model was trained in batches of 16, with a learning rate of 2e-5.

### 6.3.4 Evaluation

The “instruction\_only\_labels” prompt is employed for evaluations. A temperature setting of 0, top probability of 1, and number of beans of 1 is utilized to sample from the transformer model in a more deterministic manner. Predictions are submitted to the organizers of SemEval 2023 [89] through their submission portal for evaluation.<sup>2</sup> The organizers present two surprise languages, Georgian and Greek, for the test set. In these evaluations, two key metrics are considered: F1 Micro and F1 Macro. The F1 score measures a test’s accuracy, considering precision and recall [107]. Specifically, F1 Micro computes the F1 score globally by aggregating true positives, false negatives, and false positives across multiple classes, providing a view of the overall performance across all classes. In contrast, F1 Macro calculates the F1 score individually for each class and then takes their average. While the model is designed for a dual task of prediction and generating explanation, it’s important to acknowledge that since these examples are generated using ChatGPT, they inherit the biases of the underlying model. Such outputs necessitate human validation to ensure validity and cannot be trusted standalone.

## 6.4 Results

This section is organized into two main parts. The initial segment provides an overview of the dataset derived from the explanation augmentation methodology. Subsequently, the outcomes from the LLaMA2 model, fine-tuned on this dataset, are presented in terms of prediction accuracy.

### 6.4.1 Explanation Augmented PTD (exAug-PTD) Dataset

The explanation augmented label distributions provide a visualization of the number of instances for each propaganda technique in the augmented PTD dataset, as depicted in Figure 6.2.

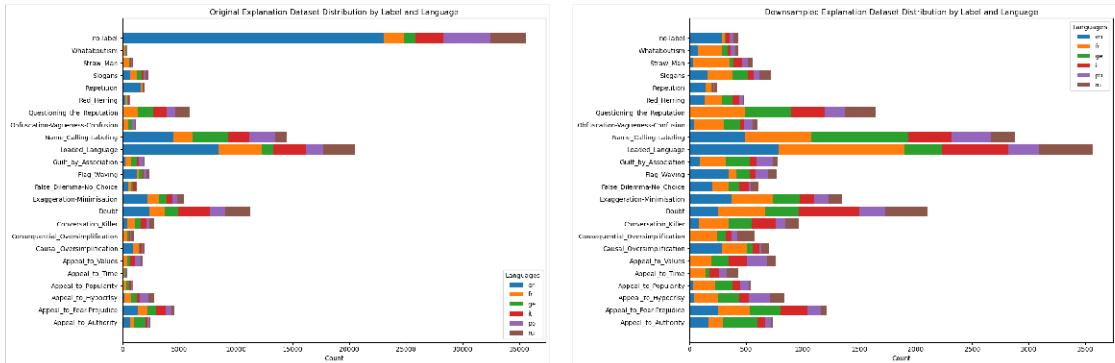
Below are randomly selected English samples of generated explanations from the augmented dataset. Each sample’s explanation is sampled twice.

1. **Input:** “Once again the media are blaming the messenger instead of dealing with Islamic violence.”

**Labels:** Whataboutism

---

<sup>2</sup>The results are accessible from <https://propaganda.math.unipd.it/semeval2023task3/leaderboard.php> under team name oalee.



(a) Explanation augmented label distribution of PTD dataset. (b) Downsampled explanation augmented label distribution.

Figure 6.2: Explanation augmented label distributions of the exAug-PTD dataset.

**Generated Explanation 1:** The text exhibits the propaganda technique of Whataboutism. Whataboutism is employed when the speaker deflects criticism or blames another party instead of addressing the main issue at hand. In this case, the author of the text accuses the media of blaming the messenger instead of addressing Islamic violence. By using Whataboutism, the author attempts to divert attention away from the topic of Islamic violence and shift the blame onto the media. This technique aims to manipulate the audience's perception and avoid addressing the actual issue.

**Generated Explanation 2:** The text contains the propaganda technique of 'Whataboutism.' This technique is employed by diverting attention from the main issue, which is the media's blame on the messenger, and shifting the focus to another issue, namely Islamic violence. By doing so, the text aims to undermine the legitimacy of the media's criticism and deflect attention away from the original topic.

2. **Input:** "Share their articles on your own social media accounts. If they can't get their work out there, we can help."

**Labels:** False Dilemma No Choice

**Generated Explanation 1:** The text contains the propaganda technique of False Dilemma, specifically the subtype of No Choice. The False Dilemma technique is employed by presenting a limited set of options and falsely suggesting that these are the only choices available, thereby excluding other alternatives. In this case, the text urges the reader to share the articles on their social media accounts as a means of support. By emphasizing this specific action, it creates the impression that not sharing the articles is equivalent to not helping. The objective of this propaganda technique is to manipulate the reader into feeling obligated to share the articles, without considering alternative ways of supporting the cause or evaluating the content's credibility. The target audience is encouraged to feel a sense of urgency and responsibility, potentially leading to uncritical sharing without fully assessing the information's accuracy or validity.

**Generated Explanation 2:** The text contains the propaganda technique known as False Dilemma-No Choice. This technique is employed through the statement 'Share their articles on your own social media accounts.' The technique presents a limited set of options, suggesting that the only choices available are to either share the articles or not. By framing

the situation as a dichotomy, it manipulates the reader into believing that there are no alternative courses of action. This technique aims to restrict critical thinking and coerce individuals into taking a specific action by eliminating the consideration of other possibilities or choices.

3. **Input:** "Then suddenly, those illegal aliens claimed they were being denied an opportunity to file for asylum or were not receiving due process. They filed a case in the Federal courts and were assigned to a Kritarch, Patti Saris, who apparently thinks she runs immigration policy, even after aliens have had their day in court. Saris thinks she can usurp the Constitution of the United States based on her whims regarding immigration policy."

**Labels:** Causal Oversimplification, Repetition

**Generated Explanation 1:** The text contains the propaganda techniques of Causal Oversimplification and Repetition.

Causal Oversimplification: This propaganda technique is employed when the text unfairly simplifies a complex issue by attributing the entire problem to a single cause. In this case, the text implies that the illegal aliens claimed they were being denied an opportunity to file for asylum or were not receiving due process, and then jumps to the conclusion that Judge Patti Saris thinks she runs immigration policy. This oversimplification ignores the multiple factors and legal complexities involved in immigration proceedings.

Repetition: The technique of repetition is used in the text to reinforce a particular message or idea. The phrase 'illegal aliens' is repeated multiple times, creating a negative association and potentially influencing the reader's perception. The repetition aims to evoke emotional responses and shape public opinion by continuously emphasizing the term 'illegal aliens' in a derogatory manner.

**Generated Explanation 2:** The text contains multiple propaganda techniques.

1. Causal Oversimplification: This technique is employed in the text through the statement 'who apparently thinks she runs immigration policy.' By portraying Judge Patti Saris as an individual who single-handedly controls immigration policy, the text oversimplifies the complex nature of the immigration system and attributes all decision-making power to her.

2. Repetition: The propaganda technique of repetition is used in the text with the repeated use of the phrase 'illegal aliens.' This repetition aims to reinforce a negative perception of immigrants by emphasizing their illegal status and potentially invoking fear or distrust.

Overall, the text employs these propaganda techniques to distort the reality of immigration policy and present a biased view that portrays immigrants in a negative light, simplifies complex issues, and reinforces stereotypes and prejudices.

4. **Input:** "Madagascar Outbreak: It Is 'Inevitable' The Plague Becomes Resistant To Drugs"

**Labels:** Appeal to Authority

**Generated Explanation 1:** The text contains the propaganda technique of Appeal to Authority. This technique is employed through the use of the phrase 'It Is 'Inevitable' The Plague Becomes Resistant To Drugs.' By invoking the authority of an unnamed source or expert who claims that the outcome of the plague becoming resistant to drugs is 'inevitable,' the text aims to persuade the audience to accept this claim without questioning it. The technique relies on the assumption that the audience will trust and defer to the authority figure mentioned, without providing concrete evidence or reasoning to support the claim. This appeal to authority can be manipulative by influencing the audience's beliefs and opinions based solely on the reputation or status of the authority figure.

**Generated Explanation 2:** The text contains the propaganda technique of Appeal to Authority. This technique involves using the opinion or endorsement of an authority figure to lend credibility and persuade the audience. In this case, the text implies that it is inevitable for the plague to become resistant to drugs based on an authoritative source or expert opinion. The target of this propaganda could be the general public or individuals concerned about the outbreak. The objective is to instill a sense of urgency and create a belief that immediate action or attention is necessary to address the issue of drug resistance in the context of the Madagascar outbreak.

5. **Input:** “Human beings are naturally ordered towards the truth, and we have to work at assuming a lie. This is why confidence tricksters can be successful, why lying works for getting what you want; people don’t see it coming. The first natural assumption is trust, at least at the basic level of expecting truth most of the time. We therefore instinctively see lying as a betrayal of trust.”

**Labels:** None

**Generated Explanation 1:** The provided text does not contain any propaganda labels. It appears to be a statement about human nature and the instinctive trust we place in others. The content focuses on the idea that human beings are naturally inclined towards truth and see lying as a betrayal of trust. It discusses how confidence tricksters and liars can be successful because people often don’t anticipate deception. The language used in the text is descriptive and aims to explain the phenomenon of trust and lying. Overall, the text does not employ manipulative techniques or attempt to influence opinions or beliefs, but rather provides an analysis of human behavior.

**Generated Explanation 2:** No propaganda techniques were identified in the text. The text primarily discusses human nature and the instinctive expectation of truth. It explores how lying can be successful due to people’s natural assumption of trust and the subsequent betrayal of that trust when lies are detected. The absence of propaganda labels in the text indicates that it focuses more on general observations and insights rather than employing specific propaganda techniques.

#### 6.4.2 Evaluation

In a handful of scenarios, the model exhibited tendencies to hallucinate labels such as Appeal\_to\_Happiness, Appeal\_to\_Uncertainty, Dilemma, Appeal\_to\_Futuro<sup>3</sup>, Appeal\_to\_Fear (the correct label is Appeal\_to\_Fear-Prejudice), and Appeal\_to\_Science. When such anomalies occurred, these hallucinated labels were removed. This occasional behavior can be attributed to the model’s token-sampling nature and the lack of mechanisms such as standalone tokens for techniques and reinforcement learning to ensure the model predicts only valid labels.

Figure 6.3 illustrates the training loss, with the noticeable dip corresponding to the completion of one epoch, indicating overfitting of the model. Figure 6.4 displays the F1 Scores at each checkpoint of the model. Tables 6.1 through 6.9 present the model’s accuracy in comparison to state-of-the-art (SoTA) models; the model of this research is presented as **ExLLaMA** and for each language, the checkpoint with the highest F1 Micro is used. From an accuracy standpoint, after 700 steps (equivalent to 11,200 samples), the model, on average, achieves its near-peak F1 Micro score and overall accuracy. This research serves as a proof of concept, showcasing that large language models can be optimized for the dual tasks of prediction and generating explanations and can attain top-tier accuracy even with a limited training set. This observation

---

<sup>3</sup>From the Italian test set.

aligns with the findings of Zhou et al. [134]. In their work, the authors demonstrated that with a carefully curated set of merely 1,000 samples, LLaMA could be fine-tuned and showcase robust performance comparable to models trained with much more data.

Acknowledging the limitation of AI-generated explanations, the subsequent section proposes a methodology for the development and assessment of PropagandaTeacherGPT using human annotations of explanations tailored for various audiences.

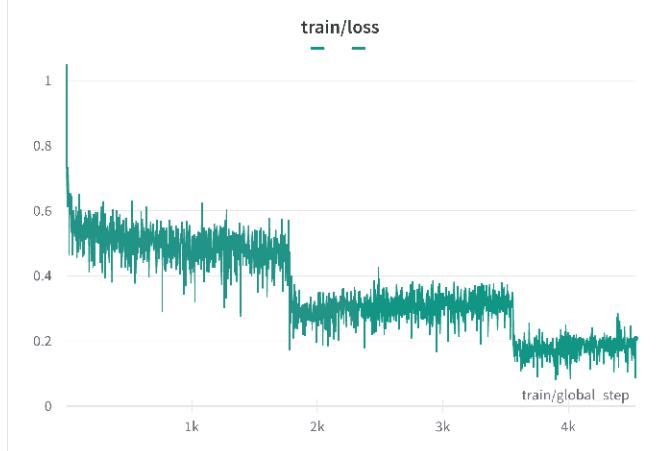


Figure 6.3: Train Loss.

Team	F1 Micro	F1 Macro
vera	0.44937	0.21698
<b>ExLLaMA</b>	0.41206	0.18622
KInIT	0.40515	0.17044
DSHacker	0.39029	0.15887
ACS	0.38628	0.17360
NAP	0.38148	0.17247
APatt	0.37562	0.12919
ReDASPersuasion	0.36553	0.09843
QCRI Team	0.36490	0.19847
NLUBot101	0.36058	0.19722
TeamAmpa	0.35951	0.18471

Table 6.1: Comparison of F1 Scores for English language.

Team	F1 Micro	F1 Macro
vera	0.59919	0.28122
KInIT	0.58544	0.28580
<b>ExLLaMA</b>	0.57122	0.27299
ReDASPersuasion	0.56486	0.17531
DSHacker	0.56480	0.23300
MarsEclipse	0.54888	0.23192
NAP	0.53879	0.26609
TeamAmpa	0.52597	0.27214
QCRI Team	0.51661	0.24156
SharoffAndLepekhin	0.51582	0.17610
SinaaAI	0.48048	0.17531

Table 6.2: Comparison of F1 Scores for Italian language.



Figure 6.4: F1 Scores of Training Checkpoints. The SoTA F1 Macro Score is set from the SoTA Micro F1 Score model.

Team	F1 Micro	F1 Macro
KInIT	0.43343	0.24205
vera	0.42635	0.25979
MarsEclipse	0.39550	0.22110
DSHacker	0.38588	0.20209
<b>ExLLaMA</b>	0.38246	0.23232
TeamAmpa	0.37781	0.22740
QCRITeam	0.36828	0.21594
ReDASPersuasion	0.35965	0.15988
ACS	0.33136	0.18251
NAP	0.32841	0.15809
NLUBot101	0.32298	0.20052

Table 6.3: Comparison of F1 Scores for Russian language.  
Table 6.4: Comparison of F1 Scores for French language.

Team	F1 Micro	F1 Macro
KInIT	0.49699	0.32169
vera	0.49238	0.34815
NAP	0.46869	0.32171
<b>ExLLaMA</b>	0.46401	0.28958
ReDASPersuasion	0.46311	0.26952
DSHacker	0.46249	0.25999
MarsEclipse	0.43601	0.25006
TeamAmpa	0.43442	0.30544
QCRITeam	0.42566	0.27344
NLUBot101	0.39580	0.25431
ACS	0.39364	0.25458

Team	F1 Micro	F1 Macro
KInIT	0.27928	0.16332
DSHacker	0.27827	0.13809
vera	0.27668	0.19865
QCRI Team	0.27261	0.15614
NAP	0.25803	0.16394
TeamAmpa	0.24098	0.19388
ACS	0.23529	0.09411
SinaaAI	0.23422	0.06936
MaChAmp	0.21465	0.12929
<b>ExLLaMA</b>	0.20398	0.04842
Riga	0.16391	0.03632

Table 6.5: Comparison of F1 Scores for Greek language.

Team	F1 Micro	F1 Macro
vera	0.44964	0.20228
KInIT	0.43105	0.24194
NAP	0.42180	0.24577
TeamAmpa	0.41164	0.26631
DSHacker	0.40753	0.15872
MarsEclipse	0.40604	0.17569
<b>ExLLaMA</b>	0.40296	0.19265
ReDASPersuasion	0.39677	0.16817
QCRI Team	0.38687	0.17241
ACS	0.37195	0.23262
APatt	0.36570	0.14969

Table 6.6: Comparison of F1 Scores for Polish language.

Team	F1 Micro	F1 Macro
vera	0.40674	0.19466
KInIT	0.39869	0.20301
<b>ExLLaMA</b>	0.39620	0.16316
TeamAmpa	0.38106	0.24366
DSHacker	0.37605	0.16803
QCRI Team	0.37191	0.17515
NAP	0.36996	0.18065
ACS	0.31730	0.13934
NLUBot101	0.30459	0.15092
SinaaAI	0.30169	0.13920
SharoffAndLepekhin	0.28149	0.07445

Table 6.7: Comparison of F1 Scores for Spanish language.

Team	F1 Micro	F1 Macro
vera	0.54174	0.25994
KInIT	0.53086	0.26292
MarsEclipse	0.52962	0.24197
DSHacker	0.52006	0.21069
NAP	0.50953	0.27195
ReDASPersuasion	0.50750	0.21182
QCRI Team	0.50207	0.24889
<b>ExLLaMA</b>	0.49758	0.25192
TeamAmpa	0.48811	0.28106
APatt	0.48375	0.17692
ACS	0.46133	0.24149

Table 6.8: Comparison of F1 Scores for German language.

Team	F1 Micro	F1 Macro
KInIT	0.46023	0.37995
vera	0.44800	0.32019
SinaaAI	0.43243	0.22689
DSHacker	0.42657	0.29866
QCRI Team	0.42368	0.29951
NAP	0.41261	0.30615
TeamAmpa	0.40816	0.25854
Riga	0.36190	0.20869
<b>ExLLaMA</b>	0.35032	0.23003
ACS	0.33544	0.24818
MaChAmp	0.30100	0.22056

Table 6.9: Comparison of F1 Scores for Georgian language.

## 6.5 A Methodology for PropagandaTeacherGPT

Explaining the decisions made by neural networks has emerged as a pivotal focus within modern AI research. While the methodology presented does not elucidate the inner workings of the model in the manner classical explainable AI does, it offers a pragmatic approach to explain model prediction. Moreover, even when AI models are interpretable, such as those utilizing game theory or other mathematical frameworks, they aren't necessarily explainable to the general public. These interpretations are often deeply entrenched in technical jargon and complex theoretical constructs. To make AI truly understandable, the explanations must be rendered in layman's terms, with the audience's background and knowledge level considered [96].

This research empirically demonstrated, as a proof of concept, that LLMs can be effectively trained on various objectives to deliver natural language explanations in conjunction with their predictions. Nonetheless, caution is warranted: AI-generated label explanations cannot be deemed reliable without proper validation, as they carry inherent biases of the AI model.

This section aims to detail a methodology for the development and assessment of 'PropagandaTeacherGPT'. The proposed methodology is structured as follows:

1. **Identification of Target Groups:** Initiate by distinguishing the primary audience segments. These segments should encompass academia, policymakers, and various subdivisions within the general public, which can be categorized by age brackets like young children, adolescents, adults, and seniors. Each group's distinct cognitive abilities and baseline knowledge will guide the crafting of age-appropriate and context-relevant explanations.
2. **Dataset Curation:** Compile a dataset prioritizing an equal distribution for each propaganda technique across multiple languages. The emphasis should not primarily be on the volume of the dataset but rather on achieving a balanced carefully curated representation of all techniques across various languages.
3. **Human Annotation:** Collaborate with domain experts to annotate and explain these techniques. Given the predetermined target groups:
  - Generate explanations that cater to varying depth levels, tailoring the content to the identified audience segments.
  - While the bulk of this procedure should be manual to encapsulate the intricacies of propaganda, AI can play a supplementary assistant role in shaping explanations suitable for the complexity level of each group.
  - Annotations should also identify the targets of the propaganda techniques. This includes:
    - *Positive Targets:* Those who stand to benefit or are portrayed favorably through the use of the propaganda technique.
    - *Negative Targets:* Those who are depicted in a negative light or stand to suffer due to the propaganda technique.
    - *Objective:* What is the objective of the propaganda within the text? This provides a neutral, unbiased description of the propaganda technique's intention.
4. **Fine-Tuning with Custom Prompts:** For distinct objectives, it's essential to curate specific prompts that guide the model in generating explanations tailored to various audience segments or tasks. These prompts not only signal the intended audience or the depth of complexity but also the nature of the response and task. For instance, a prompt might be structured as "Explain propaganda techniques with a child-friendly approach",

“Provide a detailed breakdown suitable for an expert in legal studies”, or “Identify targets and objective of propaganda for policy-makers”. Moreover, standalone tokens need to be introduced for different propaganda techniques. This then is used to fine-tune a pre-trained LLM. Subsequently, a reinforcement learning phase is implemented to ensure the model restricts its predictions to only valid labels.

5. **Evaluation Metric:** The evaluation is based on the cosine similarity metric, set to measure the semantic alignment of AI-generated explanations and predictions to a gold-standard human-annotated test set. To effectively apply this metric, it is essential to convert the textual data into numerical representations, commonly known as embeddings. One approach to obtain these embeddings is the Transformer-based Denoising AutoEncoder (TSDAE) [122]. While TSDAE has traditionally been used for creating sentence-level embeddings, which condense the semantic essence of a sentence into a compact vector, for the scope of this methodology, its utility in this research is extended to both sentence and paragraph levels. This ensures that individual propaganda targets<sup>4</sup> (at the sentence level) and broader objectives and explanations (at the paragraph level) are adequately captured. This adaptation ensures that the complex narratives of explanations, which often span several sentences, are adequately represented in the generated embeddings. Once both AI and human explanations, targets, or objectives are represented in this format, their cosine similarity is calculated. A higher cosine similarity value indicates closer semantic alignment, suggesting that the AI’s generated explanation closely mirrors the human expert’s interpretation. This approach provides a quantifiable metric to assess the fidelity and effectiveness of AI explanations and predictions.
6. **Bias Correction via Reinforcement Learning (RL):** Given the inherent biases that AI models, especially pre-trained ones, might have, it’s crucial to correct these biases to generate more neutral and accurate explanations. This correction can be achieved through reinforcement learning:
  - *Reward Mechanism:* Design a reward mechanism where the model receives positive feedback when it produces labels, objectives, targets, and explanations that align with human annotations and are less biased and a penalty for biased or misaligned predictions.
  - *Feedback Loop:* Introduce a continuous feedback loop where domain experts or crowd-sourced evaluators assess the generated explanations. Their feedback serves as a signal to adjust the model’s outputs, reducing bias iteratively.
  - *Fine-Tuning:* Use the feedback and rewards from the RL process to fine-tune the model’s parameters, guiding it towards generating better and unbiased explanations over time.

By adhering to this methodological blueprint, the endeavor is to yield explanations that harmonize with human explanation processes, ensuring clarity and minimizing biases while resonating effectively with a spectrum of audience profiles.

#### Addressing RQ4

*While ChatGPT-like tools can contribute to online manipulation by lowering the cost of creating content, how can we leverage it to counter computational propaganda?*

---

<sup>4</sup>Targets of propaganda cannot be easily defined in a finite set. Terms such as ‘Barack Obama’, ‘Barack Hossein Obama’, and ‘Obama’ refer to the same entity. Moreover, depending on the context and language, targets vary.

This research, as proof of concept, demonstrated Large Language Models' (LLMs) proficiency in predicting labels and generating explanations. Training the LLM "LLaMA 2" on GPT-generated explanations and three objectives after 11,200 samples yielded results comparable to the state-of-the-art in terms of label prediction accuracy. However, it's crucial to acknowledge that explanations crafted by LLMs, using ChatGPT, inherit model biases, raising concerns about explanation credibility. To address this, the study introduces a methodology for the development and assessment of PropagandaTeacherGPT. This model is designed to provide explanations for propaganda techniques while actively countering biases. The methodology encompasses tailored explanations for diverse audiences, careful dataset curation, human annotations of explanations, targets, and objectives, and fine-tuning with reinforcement learning. The assessment employs cosine similarity of explanation embeddings to measure semantic alignment, measuring explanations' alignment to human-annotated explanations in meaning. Furthermore, annotations, encompassing positive and negative targets along with objectives, could bridge NLP, network-centric analyses, and educational explainable AI.

## Chapter 7

# Discussion, Conclusion and Future Work

In this final chapter, the insights and findings from the investigation into the Iranian Twitter-sphere’s political dynamics are consolidated. The potential avenues for future exploration in this domain are also presented.

### 7.1 Discussion and Conclusion

This study proposed a method to determine political ideologies by analyzing selected Twitter hashtags using unsupervised clustering and subsequent labeling of a set of hashtags. Applying this methodology, within the Iranian Twittersphere, three main political groups were identified: the Islamic Republic, the Monarchy, and the MEK. Their respective shares in the dataset were 4.1%, 3.5%, and 0.2%.

The death of Mahsa Amini served as a catalyst for these communities, leading to increased activity and an intensified use of radical hashtags, MEK using “Death to oppressor whether king or leader” as their top hashtag, Monarchy using “Death to three corrupt, Mullahs, Leftists, Mojahedin (MEK)” and the Islamic Republic using “Decisive Reaction” in support of protester oppression and “Do not execute the machete wielder” sarcastically to support executions of protesters. Each group utilized the Mahsa Amini hashtags to promote their ideology and agenda. The Islamic Republic uses “At your service O Khamenei” and took a denigrating stance, blaming the movement for the Shah Cheragh terrorist attacks, notably employing the Reductio ad Hitlerum (Guilt By Association) propaganda technique. Confirming Kermani’s [63] observations, they further sought to dilute the movement’s online presence by deliberately misspelling the Mahsa Amini variation hashtag; the same tactic is also observed with OpIran hashtag, using ‘I’ instead of ‘I’. Conversely, both the MEK and Monarchy employ an “Appeal To Time” technique, co-opting the Mahsa Amini hashtag with those of nationwide protests and strikes and appealing that the time has come for this action, implying urgency. MEK distinctively couples the Mahsa Amini hashtag with hashtags like “Death to oppressor whether king or leader” and their leader, “Masoud Rajavi”. Monarchy, meanwhile, uses the hashtag with a conjunction of the “Everlasting King” hashtag and ‘I Will Give Attorneyship’ a campaign endorsing Prince Reza Pahlavi’s power of attorneyship for the people of Iran on change.org. In contrast, the “Women.Life.Freedom” hashtag saw varied usage across factions: it was a centerpiece for the Monarchy, used to a lesser degree by the Islamic Republic, and is interestingly missing from the MEK’s top 50 most frequently

used hashtags. While the Islamic Republic is using it to denigrate the hashtag, both use it to advocate for their agendas and ideologies similar to the Mahsa Amini hashtag. However, here the Islamic Republic introduces new hashtags such as ‘Women Life Martydom’, ‘Women Chaste Honor’, ‘Women Life Awareness’ to promote their ideology and denigrate the movement. Interestingly, both Islamic Republic and Monarchy use ‘Women Life Freedom’ hashtag in conjunction with ‘Men Homeland Prosperity’.

All groups leveraged emotionally charged hashtags: the Monarchy and MEK invoked victim-centric hashtags to evoke emotions, while the Islamic Republic frequently referenced the Shah Cheragh attack. Common propaganda techniques, such as Repetition, Reductio ad Hitlerum (Guilt By Association), Dictatorship, Appeal To Authority, Slogan, Flag Waving, Loaded Language, Name Calling/Labeling, and Virtue (Glistering Generalities), were pervasive across all groups in their hashtags.

The research also proposed an N-gram analysis method to uncover dominant discourse themes. This methodology illuminated the dominant discourse trends of the groups. The Monarchy leaned heavily into Dictatorship, Appeal To Fear, and Name Calling propaganda techniques. The Islamic Republic, on the other hand, employed Name Calling and the Appeal To Values, especially Shia-centered values. Remarkably, the monarchy network exhibited a tangible influence on public discourse. The slogan “Men Homeland Prosperity (M.H.P)” gained momentum due to Sherwin Hajipour’s Grammy award-winning song Baraye, composed of tweets. It was found that the primary propagator of M.H.P was from the Monarchy group, and 75% of the first 20 tweets contacting M.H.P came from monarchy users. Furthermore, the initial propagator employed the Reductio ad Hitlerum (Guilt By Association) propaganda technique by connecting Women Life Freedom to communism and suggesting it’s ‘incomplete’ without nationalistic and a masculine dimension, it introduces the counterpart ‘Men Homeland Prosperity’. This slogan became the most frequently used bi-gram after Mahsa and later spread to people’s chants in protests. Moreover, Reza Pahlavi emerged as a dominant bi-gram in discussions for seven weeks following the initiation of a change.org campaign. During the first week, public discourse mirrored the monarchy’s propaganda terms and techniques, especially the “Appeal To Fear” technique associated with separatism and territorial integrity.

In addition, this research explored the capabilities of LLMs in the domain of persuasion technique identification and explanation and proposed a data augmentation methodology using ChatGPT to generate explanations with varied objectives. This augmented data was then used to fine-tune LaMMA 2 for three distinct tasks: generating explanations using two different prompts and objectives and standalone label prediction. Empirical findings suggest that fewer data annotations with explanations are required to fine-tune large language models to achieve top-tier accuracy. Acknowledging the limitations and inherent biases in AI-generated explanations, a methodology was outlined for the development and assessment of PropagandaTeacherGPT, an educational AI reinforced with human explanation annotations tailored for various audiences.

## 7.2 Future Work

Drawing from the insights of Reiter [96], explanations should account for the user’s background and the required complexity of the explanation. There’s potential for Large Language Models (LLMs) to be tailored with distinct prompts, enabling them to address a range of complexities. An intriguing avenue for exploration encompasses the development of the outlined PropagandaTeacherGPT, labeling data adhering to a proposed methodology and categorizing annotation of propaganda by techniques, explanations tailored for various audiences, positive and negative targets, and objectives. Such an approach can potentially bridge network-centric, NLP-driven,

and explainable educational propaganda analyses.

Within the Iranian Twittersphere, should data be available, future research could delve into bot dynamics and their behavioral patterns. Expanding the dataset to incorporate aspects such as retweet networks and “like” patterns could enrich the analysis. A personal observation on Twitter brought to light certain accounts—suspected to be bots—demonstrated intriguing liking behavior spanning a wide opposite political spectrum, ranging from Khamenei, Iran’s supreme leader, to Reza Pahlavi and Israel’s foreign ministry. Such patterns might indicate sophisticated strategies in play. While the overt propaganda network of the Islamic Republic appears to exert minimal influence on the public, they might exploit bots and other strategies to intensify polarization and simultaneously bolster both Pro-Monarchy and Anti-Monarchy rhetoric to drive polarization further and redirect discussions.

With the potential return of the API, tools like Botometer [128] can be employed to scrutinize bot behavior within the dataset. Botometer, tailored to assess the automation likelihood of a Twitter account, can help distinguish between human and bot discourse. Additional research is necessary to further identify bots within the dataset, possibly enriched with external data. Recognizing their prevalence and patterns is crucial, given the significant role bots can play in molding narratives and potentially manipulating public discourse. The evolving nature of these bots in response to detection methods underscores the need for continual research in this domain.

Furthermore, a more granular analysis of user affiliations utilizing labeled data could unveil a spectrum of affiliations. Moreover, Labeling tweets in accordance with the propaganda techniques would enable its training and subsequent analysis.

# Bibliography

- [1] Ervand Abrahamian. *The Iranian Mojahedin*. Yale University Press, 1989.
- [2] Ervand Abrahamian. *Iran between two revolutions*. Princeton University Press, 2021.
- [3] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints, 2023.
- [4] Iuliia Alieva, Lynnette Hui Xian Ng, and Kathleen M Carley. Investigating the spread of russian disinformation about biolabs in ukraine on twitter using social network analysis. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1770–1775. IEEE, 2022.
- [5] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.
- [6] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Andreas Bennetot, Siham Tabik, Alicia Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.
- [7] Pooya Azadi and Mohsen B Mesgaran. The clash of ideologies on persian twitter, 2021.
- [8] Arash Barfar. A linguistic/game-theoretic approach to detection/explanation of propaganda. *Expert Systems with Applications*, 189:116069, 2022.
- [9] Alberto Barrón-Cedeno, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. Proppy: A system to unmask propaganda in online news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9847–9848, 2019.
- [10] W Lance Bennett and Alexandra Segerberg. The logic of connective action: Digital media and the personalization of contentious politics. *Information, communication & society*, 15(5):739–768, 2012.
- [11] Thales Bertaglia, Stefan Huber, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi. Closing the loop: Testing chatgpt to generate model explanations to improve human labelling of sponsored content on social media, 2023.
- [12] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

- [13] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Eric Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
- [14] Gillian Bolsover and Philip Howard. Computational propaganda and political big data: Moving toward a more critical research agenda, 2017.
- [15] Jordan Boyd-Graber, Yuening Hu, David Mimno, et al. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296, 2017.
- [16] Samantha Bradshaw and Philip Howard. Troops, trolls and troublemakers: A global inventory of organized social media manipulation. *Computational Propaganda Research Project*, 2017.
- [17] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
- [18] Chiyu Cai, Linjing Li, and Daniel Zeng. Detecting social bots by jointly modeling deep behavior and content information. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1995–1998, 2017.
- [19] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations, 2018.
- [20] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.
- [21] Manuel Castells. *Networks of outrage and hope: Social movements in the Internet age*. John Wiley & Sons, 2015.
- [22] C. Chai. Comparison of text preprocessing methods. *Natural Language Engineering*, 29:509–553, 2022.
- [23] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. Debot: Twitter bot detection via warped correlation. In *Icdm*, volume 18, pages 28–65, 2016.
- [24] Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. Transformers:“the end of history” for natural language processing? In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, pages 677–693. Springer, 2021.
- [25] Aditya Chetan, Brihi Joshi, Hridoy Sankar Dutta, and Tamoy Chakraborty. Corerank: Ranking to detect users involved in blackmarket-based collusive retweeting activities. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 330–338, 2019.

- [26] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 89–96, 2011.
- [27] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on world wide web companion*, pages 963–972, 2017.
- [28] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. Emergent properties, models, and laws of behavioral similarities within groups of twitter users. *Computer Communications*, 150:47–61, 2020.
- [29] Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 162–170, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [30] Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online), December 2020. International Committee for Computational Linguistics.
- [31] Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeno, and Preslav Nakov. Prta: A system to support the analysis of propaganda techniques in the news. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 287–293, 2020.
- [32] Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, EMNLP-IJCNLP 2019, Hong Kong, China, November 2019.
- [33] Lavinia Dan. Techniques for the translation of advertising slogans. In *Proceedings of the International Conference Literature, Discourse and Multicultural Dialogue, LDMD*, volume 15, pages 13–23, 2015.
- [34] Manijeh Daneshpour and Sima Hassandokht Firooz. Women, life, freedom: The new unveiling of feminism. *Journal of Feminist Family Therapy*, 34(3-4):390–394, 2022.
- [35] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web*, pages 273–274, 2016.
- [36] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.
- [37] Saeed Kamali Dehghan and Richard Norton-Taylor. Cia admits role in 1953 iranian coup. *The guardian*, 19, 2013.

- [38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [39] Mario Diani. The concept of social movement. *The sociological review*, 40(1):1–25, 1992.
- [40] Juan Echeverri  a, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Shi Zhou. Lobo: Evaluation of generalization deficiencies in twitter bot classifiers. In *Proceedings of the 34th annual computer security applications conference*, pages 137–146, 2018.
- [41] Charles A Fleming. Understanding propaganda from a general semantics perspective. *ETC.: A Review of General Semantics*, 52(1):3–13, 1995.
- [42] Institute for Propaganda Analysis. *How to detect propaganda*, volume I, chapter 2, pages 210–218. Publications of the Institute for Propaganda Analysis, New York, NY, USA, 1938.
- [43] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.
- [44] Floriana Gargiulo, Jacopo Bindi, and Andrea Apolloni. The topology of a discussion: the #occupy case. *PloS one*, 10(9):e0137191, 2015.
- [45] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.
- [46] Jean Goodwin and Raymie McKerrow. Accounting for the force of the appeal to authority. 2011.
- [47] Christian Grimme, Mike Preuss, Lena Adam, and Heike Trautmann. Social bots: Human-like by means of human control? *Big data*, 5(4):279–293, 2017.
- [48] Yupeng Gu, Ting Chen, Yizhou Sun, and Bingyu Wang. Ideology detection for twitter users via link analysis. In *Social, Cultural, and Behavioral Modeling: 10th International Conference, SBP-BRiMS 2017, Washington, DC, USA, July 5-8, 2017, Proceedings 10*, pages 262–268. Springer, 2017.
- [49] Stefano Guarino, Noemi Trino, Alessandro Celestini, Alessandro Chessa, and Gianni Rotta. Characterizing networks of propaganda on twitter: a case study. *Applied Network Science*, 5(1):1–22, 2020.
- [50] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.
- [51] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.
- [52] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations, 2016.
- [53] Renee Hobbs and Sandra McGee. Teaching about propaganda: An examination of the historical roots of media literacy. *Journal of Media Literacy Education*, 6(2):56–66, 2014.

- [54] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022.
- [55] John Hunter. *Brainwashing in a large group awareness training?: the classical conditioning hypothesis of brainwashing*. PhD thesis, 2015.
- [56] William John Hutchins. *Machine translation: past, present, future*. Ellis Horwood Chichester, 1986.
- [57] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, 2007.
- [58] Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang. Inferring lockstep behavior from connectivity pattern in large graphs. *Knowledge and Information Systems*, 48:399–428, 2016.
- [59] Garth S Jowett and Victoria O’donnell. *Propaganda & persuasion*. Sage publications, 2018.
- [60] Garth S. Jowett and Victoria O’Donnell. What is propaganda, and how does it differ from persuasion? In *Propaganda & Persuasion*, chapter 1, pages 1–48. Sage Publishing, 2012.
- [61] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*, volume 1. Pearson Education, 3rd edition, 2020.
- [62] John Kelly and Bruce Etling. Mapping iran s online public: Politics and culture in the persian blogosphere. *Berkman Center for Internet and Society and Internet & Democracy Project, Harvard Law School*, 2008.
- [63] Hossein Kermani. # mahaamini: Iranian twitter activism in times of computational propaganda. *Social Movement Studies*, pages 1–11, 2023.
- [64] Hossein Kermani and Niloofar Hooman. Hashtag feminism in a blocked context: The mechanisms of unfolding and disrupting# rape on persian twitter. *New Media & Society*, page 14614448221128827, 2022.
- [65] Hossein Kermani and Amirali Tafreshi. Walking with Bourdieu into Twitter communities: An analysis of networked publics struggling on power in Iranian twittersphere. *Information, Communication & Society*, 26(8):1653–1674, 2023.
- [66] Emad Khazraee. Mapping the political landscape of persian twitter: The case of 2013 presidential election. *Big Data & Society*, 6(1):2053951719835232, 2019.
- [67] Ramona Kreis. # refugeesnotwelcome: Anti-refugee discourse on twitter. *Discourse & Communication*, 11(5):498–514, 2017.
- [68] Sneha Kudugunta and Emilio Ferrara. Deep neural networks for bot detection. *Information Sciences*, 467:312–322, 2018.
- [69] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

- [70] Sangho Lee and Jong Kim. Early filtering of ephemeral malicious accounts on twitter. *Computer communications*, 54:48–57, 2014.
- [71] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions, 2016.
- [72] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery., 2018.
- [73] Shenghua Liu, Bryan Hooi, and Christos Faloutsos. Holoscope: Topology-and-spike aware fraud detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1539–1548, 2017.
- [74] Mattias Luber, Anton Thielmann, Christoph Weisser, and Benjamin Säfken. Community-detection via hashtag-graphs for semi-supervised nmf topic models. *arXiv preprint arXiv:2111.10401*, 2021.
- [75] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.
- [76] Ammar Maleki. Iranians' attitudes toward political systems: A 2022 survey report. *Netherlands: GAMAAN, the Group for Analyzing and Measuring Attitudes in Iran. Retrieved*, 1(10):2023, 2022.
- [77] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [78] Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. A survey on computational propaganda detection. *arXiv preprint arXiv:2007.08024*, 2020.
- [79] Hamid Mavani. *Religious authority and political thought in Twelver Shi’ism: From Ali to post-Khomeini*. Routledge, 2013.
- [80] Michele Mazza, Stefano Cresci, Marco Avvenuti, Walter Quattrociocchi, and Maurizio Tesconi. Rtbust: Exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th ACM conference on web science*, pages 183–192, 2019.
- [81] Clyde R. Miller. The techniques of propaganda, 1939. From "How to Detect and Analyze Propaganda," an address given at Town Hall.
- [82] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences, 2018.
- [83] Majid Mohammadi. Iranian women and the civil rights movement in iran: Feminism interacted. *Journal of International Women’s Studies*, 9(1):1–21, 2007.
- [84] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.
- [85] Mark EJ Newman. Detecting community structure in networks. *The European physical journal B*, 38:321–330, 2004.
- [86] OpenAI. Gpt-4 technical report, 2023.
- [87] Diogo Pacheco, Alessandro Flammini, and Filippo Menczer. Unveiling coordinated groups behind white helmets disinformation. In *Companion proceedings of the web conference 2020*, pages 611–616, 2020.

- [88] Jakub Piskorski, Nicolas Stefanovitch, Valerie-Anne Bausier, Nicolo Faggiani, Jens Linge, Sopho Kharazi, Nikolaos Nikolaidis, Giulia Teodori, Bertrand De Longueville, Brian Doherty, Jason Gonin, Camelia Ignat, Bonka Kotseva, Eleonora Mantica, Lorena Marcaletti, Enrico Rossi, Alessio Spadaro, Marco Verile, Giovanni Da San Martino, Firoj Alam, and Preslav Nakov. News categorization, framing and persuasion techniques: Annotation guidelines. Technical report, European Commission Joint Research Centre, Ispra (Italy), March 2023.
- [89] Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [90] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 2018.
- [91] Emad Rahamanian. The enemy narrative: How authoritarian regimes use twitter to oppress public opinion, evidence from flight ps752. *The Journal of Social Media in Society*, 10(2):344–372, 2021.
- [92] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. Deepfake detection: A systematic literature review. *IEEE access*, 10:25494–25513, 2022.
- [93] Francisco Rangel and Paolo Rosso. Overview of the 7th author profiling task at pan 2019: bots and gender profiling in twitter. *Working notes papers of the CLEF 2019 evaluation labs*, 2380:1–7, 2019.
- [94] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937, 2017.
- [95] Eshrag Ali Refaee, Basem Ahmed, and Motaz Saad. AraBEM at WANLP 2022 shared task: Propaganda detection in Arabic tweets. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 524–528, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.
- [96] Ehud Reiter. Natural language generation challenges for explainable ai, 2019.
- [97] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.
- [98] Monika L Richter. The kremlin’s platform for ‘useful idiots’ in the west: An overview of rt’s editorial strategy and evidence of impact. *European Values*, 2017.
- [99] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.
- [100] Elizabeth Rubin. The cult of rajavi. *The New York Times*, 2003.

- [101] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.
- [102] Mohsen Sayyadilarikandeh, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2725–2732, 2020.
- [103] Heidi Schuhbauer, Sebastian Schötteler, Johannes Niu, Bernhard Schiffer, and David Wolforth. A quantitative social network analysis of politicians’ tweets to explore political communication.
- [104] Alexander Shevtsov, Christos Tzagkarakis, Despoina Antonakaki, and Sotiris Ioannidis. Identification of twitter bots based on an explainable machine learning framework: The us 2020 elections case study, 2021.
- [105] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.
- [106] Hadi Sohrabi. New media, contentious politics, and political public sphere in iran. *Critical Arts*, 35(1):35–48, 2021.
- [107] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.
- [108] Maya Stewart and Ulrike Schultze. Producing solidarity in social media activism: The case of my stealthy freedom. *Information and organization*, 29(3):100251, 2019.
- [109] Cass Sunstein. *# Republic: Divided democracy in the age of social media*. Princeton university press, 2018.
- [110] Sara Tafakori. Digital feminism beyond nativism and empire: Affective territories of recognition and competing claims to suffering in iranian women’s campaigns. *Signs: Journal of Women in Culture and Society*, 47(1):47–80, 2021.
- [111] Gabriel H Teninbaum. Reductio ad hitlerum: Trumping the judicial nazi card. *Mich. St. L. Rev.*, page 541, 2009.
- [112] Robyn Torok. Symbiotic radicalisation strategies: Propaganda tools and neuro linguistic programming. In *Proceedings of the Australian Security and Intelligence Conference*, pages 58–65, Perth, Australia, 2015.
- [113] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.
- [114] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura,

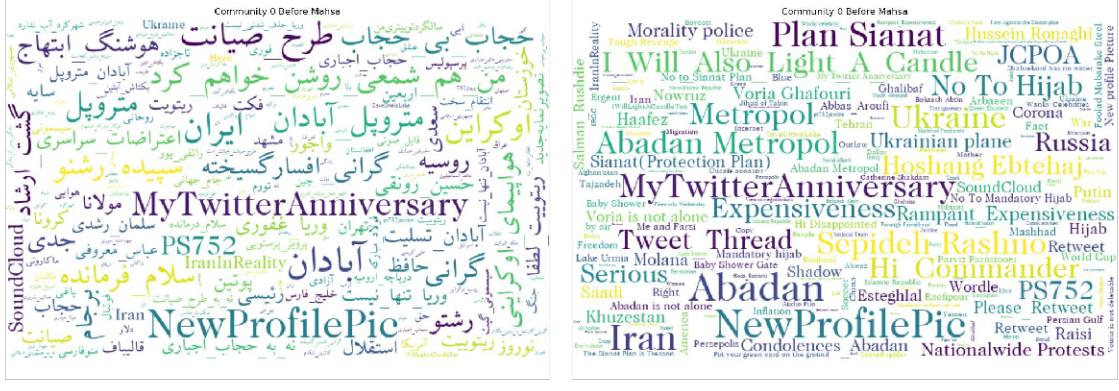
Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poultion, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Bin Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

- [115] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233, 2019.
- [116] Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*, 2018.
- [117] Zeynep Tufekci. *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press, 2017.
- [118] Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 280–289, 2017.
- [119] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.
- [120] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. In *2017*, 2017.
- [121] Douglas Walton. *Methods of argumentation*. Cambridge University Press, 2013.
- [122] Kexin Wang, Nils Reimers, and Iryna Gurevych. Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning, 2021.
- [123] Claire Wardle and Hossein Derakhshan. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg, 2017.
- [124] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*. Cambridge university press, 1994.
- [125] Anthony Weston. *A Rulebook for Arguments*. Hackett Publishing, 5th edition, 2018.
- [126] William A Woods. Transition network grammars for natural language analysis. *Communications of the ACM*, 13(10):591–606, 1970.
- [127] Samuel C Woolley and Philip Howard. Computational propaganda worldwide: Executive summary. 2017.
- [128] Kai-Cheng Yang, Emilio Ferrara, and Filippo Menczer. Botometer 101: Social bot practicum for computational social scientists. *Journal of Computational Social Science*, 5(2):1511–1528, 2022.

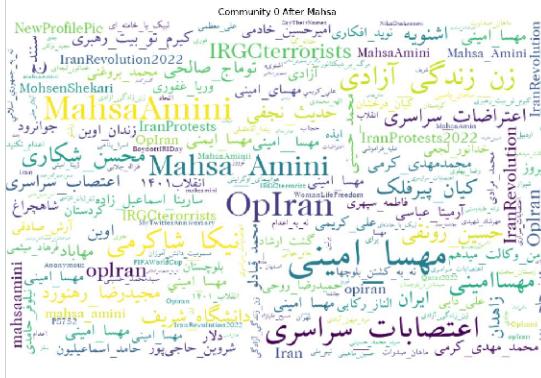
- [129] Kai-Cheng Yang, Onur Varol, Clayton A Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1):48–61, 2019.
- [130] Zhi Yang, Jilong Xue, Xiaoyong Yang, Xiao Wang, and Yafei Dai. Votetrust: Leveraging friend invitation graph to defend against social network sybils. *IEEE Transactions on dependable and secure computing*, 13(4):488–501, 2015.
- [131] Seunghak Yu, Giovanni Da San Martino, Mitra Mohtarami, James Glass, and Preslav Nakov. Interpretable propaganda detection in news articles, 2021.
- [132] Jianyi Zhang, Xu Ji, Zhangchi Zhao, Xiali Hei, and Kim-Kwang Raymond Choo. Ethical considerations and policy implications for large language models: Guiding responsible development and deployment, 2023.
- [133] Jinxue Zhang, Rui Zhang, Yanchao Zhang, and Guanhua Yan. The rise of social botnets: Attacks and countermeasures. *IEEE Transactions on Dependable and Secure Computing*, 15(6):1068–1082, 2016.
- [134] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. Lima: Less is more for alignment, 2023.

## **Appendix A**

### **Unsupervised Clustered Community Wordclouds**



(a) Hashtags in the original language for Community 0 (People) Before Mahsa



(c) Hashtags in the original language for Community 0 (People) after Mahsa

Figure A.1: Top Hashtags Used by Community 0 (People) in the Unsupervised Network Clusters: This word cloud represents the most frequently used hashtags by Community 0 (People) in periods Before Mahsa (January 2022 to September 14, 2022) and after Mahsa (September 14, 2022 to April 1, 2023). The size of each word in the word cloud is proportional to its usage frequency in the respective community.



Figure A.2: Top Hashtags Used by Community 1 (Islamic Republic) in the Unsupervised Network Clusters: This word cloud represents the most frequently used hashtags by Community 1 (Islamic Republic) in periods Before Mahsa (January 2022 to September 14, 2022) and after Mahsa (September 14, 2022 to April 1, 2023). The size of each word in the word cloud is proportional to its usage frequency in the respective community.

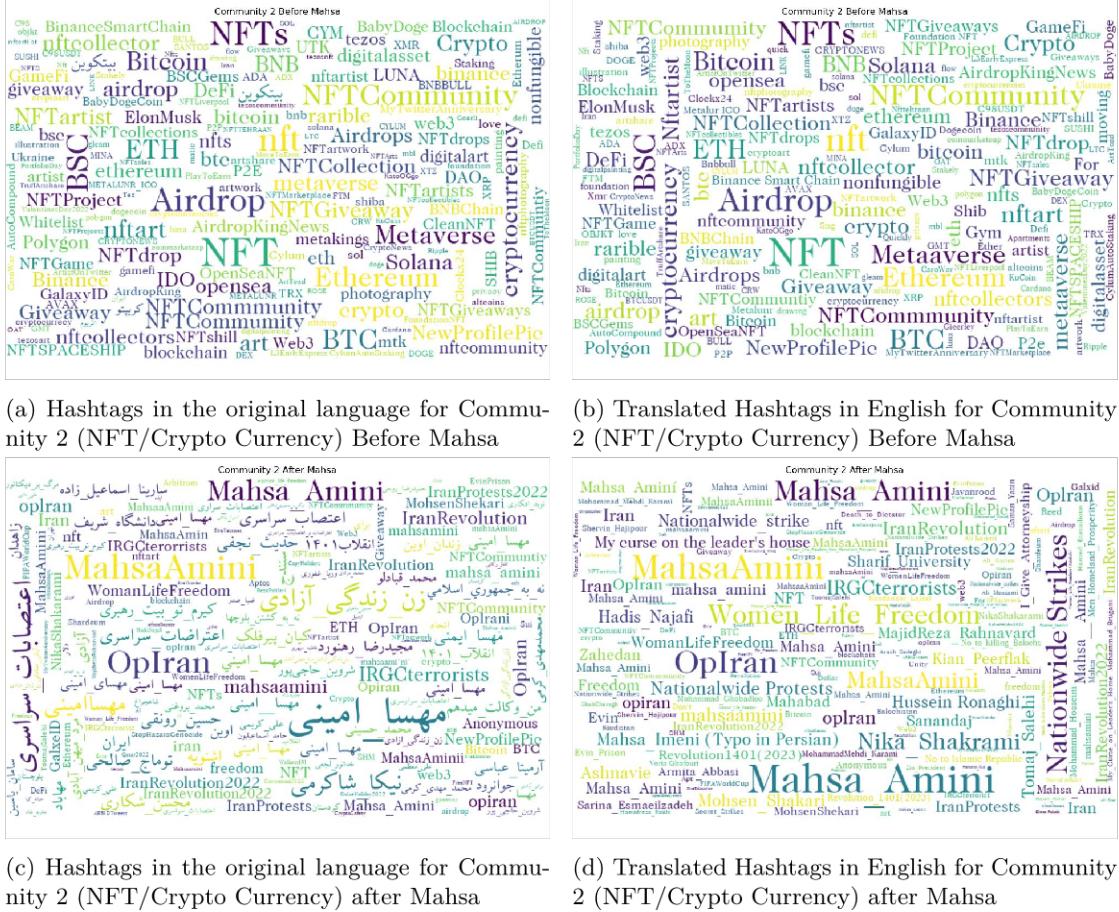


Figure A.3: Top Hashtags Used by Community 2 (NFT/Crypto Currency) in the Unsupervised Network Clusters: This word cloud represents the most frequently used hashtags by Community 2 (NFT/Crypto Currency) in periods Before Mahsa (January 2022 to September 14, 2022) and after Mahsa (September 14, 2022 to April 1, 2023). The size of each word in the word cloud is proportional to its usage frequency in the respective community.



Figure A.4: Top Hashtags Used by Community 3 (Monarchy) in the Unsupervised Network Clusters: This word cloud represents the most frequently used hashtags by Community 3 (Monarchy) in periods Before Mahsa (January 2022 to September 14, 2022) and after Mahsa (September 14, 2022 to April 1, 2023). The size of each word in the word cloud is proportional to its usage frequency in the respective community.

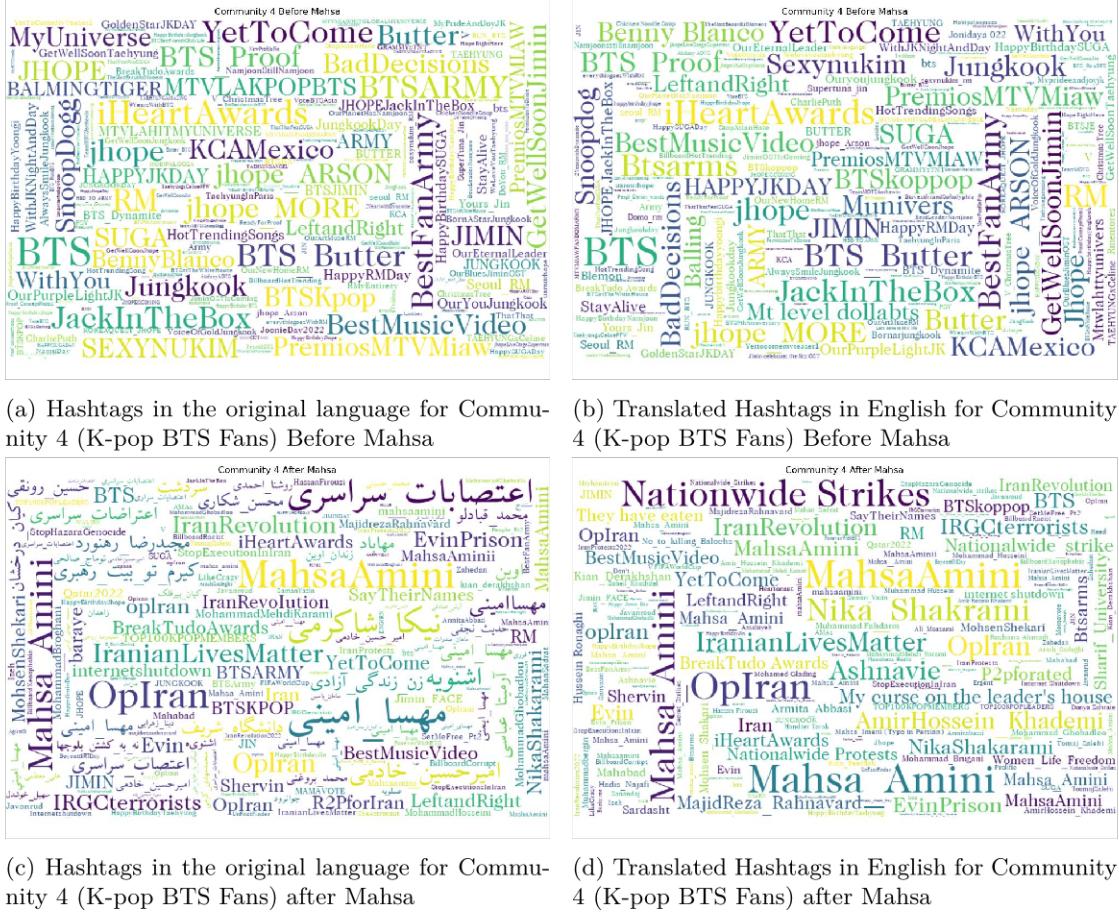


Figure A.5: Top Hashtags Used by Community 4 (K-pop BTS Fans) in the Unsupervised Network Clusters: This word cloud represents the most frequently used hashtags by Community 4 (K-pop BTS Fans) in periods Before Mahsa (January 2022 to September 14, 2022) and after Mahsa (September 14, 2022 to April 1, 2023). The size of each word in the word cloud is proportional to its usage frequency in the respective community.



(a) Hashtags in the original language for Community 5 (International/English Speaking Activists) Before Mahsa

(b) Translated Hashtags in English for Community 5 (International/English Speaking Activists) Before Mahsa



(c) Hashtags in the original language for Community 5 (International/English Speaking Activists) after Mahsa



(d) Translated Hashtags in English for Community 5 (International/English Speaking Activists) after Mahsa

Figure A.6: Top Hashtags Used by Community 5 (International/English Speaking Activists) in the Unsupervised Network Clusters: This word cloud represents the most frequently used hashtags by Community 5 (International/English Speaking Activists) in periods Before Mahsa (January 2022 to September 14, 2022) and after Mahsa (September 14, 2022 to April 1, 2023). The size of each word in the word cloud is proportional to its usage frequency in the respective community.



Figure A.7: Top Hashtags Used by Community 6 (MEK) in the Unsupervised Network Clusters: This word cloud represents the most frequently used hashtags by Community 6 (MEK) in periods Before Mahsa (January 2022 to September 14, 2022) and after Mahsa (September 14, 2022 to April 1, 2023). The size of each word in the word cloud is proportional to its usage frequency in the respective community.

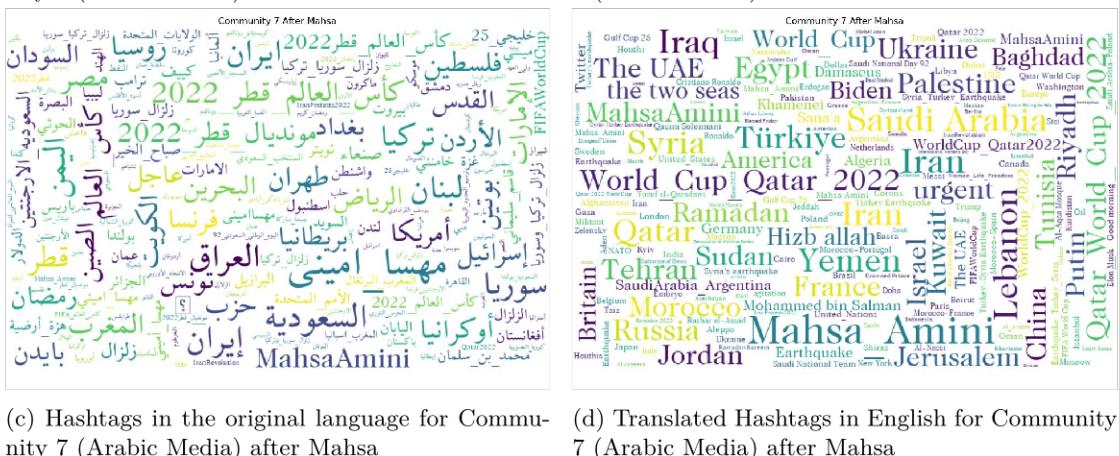
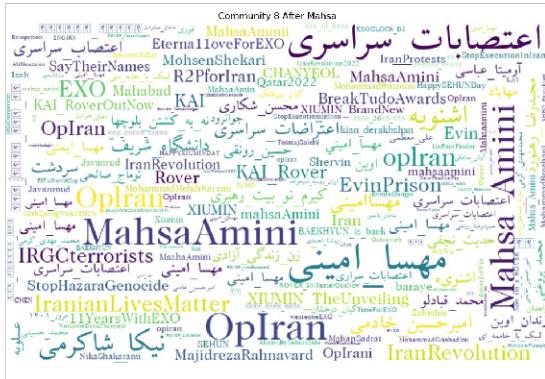


Figure A.8: Top Hashtags Used by Community 7 (Arabic Media) in the Unsupervised Network Clusters: This word cloud represents the most frequently used hashtags by Community 7 (Arabic Media) in periods Before Mahsa (January 2022 to September 14, 2022) and after Mahsa (September 14, 2022 to April 1, 2023). The size of each word in the word cloud is proportional to its usage frequency in the respective community.



(a) Hashtags in the original language for Community 8 (K-pop EXO Fans) Before Mahsa



(c) Hashtags in the original language for Community 8 (K-pop EXO Fans) after Mahsa

(b) Translated Hashtags in English for Community 8 (K-pop EXO Fans) Before Mahsa



(d) Translated Hashtags in English for Community 8 (K-pop EXO Fans) after Mahsa

Figure A.9: Top Hashtags Used by Community 8 (K-pop EXO Fans) in the Unsupervised Network Clusters: This word cloud represents the most frequently used hashtags by Community 8 (K-pop EXO Fans) in periods Before Mahsa (January 2022 to September 14, 2022) and after Mahsa (September 14, 2022 to April 1, 2023). The size of each word in the word cloud is proportional to its usage frequency in the respective community.

# Appendix B

## Labeled Hashtags

Hashtag(s)	Translation	Negative Towards
نه_سلطنت_نه_رہبری_دموکراسی_برابری	No monarchy, no leadership, democracy, equality	Islamic Republic, Monarchy
مرگ_بر_سمنگر_چه_شاہ_باشد_چه_رہبر	Death to the oppressor, be it a king or a leader	Islamic Republic, Monarchy
IRGCterrorists, IRGCTerrorists	IRGC (Islamic Republic Revolutionary Guards) terrorists	Islamic Republic

Table B.1: Hashtags and their negativity towards ideologies

Table B.2: Hashtags Affiliated With MEK

hashtag	translation
1. WeSupportMEK,, 2. MEKResistanceUnits,, 3. NCRIAlternative,, 4. IStandWithMaryamRajavi,, 5. PMOIVictory,, 6. OurChoiceMaryamRajavi,, 7. „دروود_بر_مجاهدین“ 8. FreeIran10PointPlan,, 9. شورای_می_مقاومت_ایران 10. شورای_می_ مقاومت 11. درود_بر_رجوی 12. FreeIranTelethon,, 13. کانون_شورشی 14. درود_بر_مجاهد 15. کانون_های_شورشی 16. کانونهای_شورشی_مسلح 17. درود_بر_کانونهای_شورشی	1. WeSupportMEK,, 2. MEKResistanceUnits,, 3. NCRIAlternative,, 4. I stand with Maryam Rajavi,, 5. PMOIVictory,, 6. OurChoiceMaryamRajavi,, 7. Greetings to the Mujahideen,, 8. Free Iran 10 Point Plan,, 9. National Council of Resistance,, 10. National Council of Resistance of Iran,, 11. Salute to Rajavi,, 12. Free Iran Telethon,, 13. Greetings to Mujahideen,, 14. Rebellion Center,, 15. Armed insurrectionist centers,, 16. Insurgent centers,, 17. Salutations to the Revolutionary Centers

Table B.3: Hashtags Affiliated With Islamic Republic

hashtag	translation
غیرتلی_سردار_ایران	1. Capital of Insight, 2. Iran Power, 3. Gherteli Commander
پاپخت_ بصیرت	4. Arbaeen Shahragh, 5. GreatSatan, 6. Commander of Hearts, 7. Raefipour is our voice,
فرمانده_دلا	8. Love of Hussein gathers us, 9. QudsDay, 10. FreePalaestine, 11. freepalestine, 12. I am also a soldier of Khamenei, 13. Husseini Generation, 14. School of Hussein, 15. House of Hussein Iran, 16. Only Haider is the commander of the faithful, 17. School of Sayyid Al Shahada (Imam Hussein), 18. Den of Corruption, 19. Oh Allah, hasten the reappearance of your Guardian, 20. Imam of the nation, 21. Islamic Republic is sacred, 22. Death to Al Saud, 23. ThePromisedSaviour, 24. Provincial soldier, 25. Yaaddi, 26. At your service, O Imam Khamenei, 27. Long live the Islamic Republic of Iran, 28. Hi Imam of my time, 29. Great Satan, 30. Savior of the world, 31. ExpelUSA, 32. Fadaianeh Mehdi, 33. Quds_Shirif, 34. Do Not Execute The Corrupt, 35. O Oh Allah Bless Muhammad and Accelerate the Relief of Your Representative, 36. AtYourService O Mahdi, 37. Oh God, hurry the wilderness of the vulva, 38. AtYourService O Khamenei, 39. Allah hasten your relief to your representative, 40. ThanksIRGC, 41. Jerusalem is the capital of the resistance, 42. The subversion of the stupid, 43. Homeland IS, 44. Death to Israel, 45. ThePromisedSavior, 46. Fajr Mubarak Decade, 47. O Allah, hasten the return of your representative, 48. Hello or Mehdi, 49. We are the nation of Imam Hussein, 50. Lie Beating and Insulting Mahsa, 51. Mahsa's beatings lie, 52. Mahsa Amini lie, 53. At Your Service Khamenei, 54. Mahsa sedition, 55. Al-Quds is the Axis, 56. Lovers of Qasem, 57. USAHumanRights, 58. Strong Army, 59. Imam of the hearts, 60. At your service, O Khamenei, 61. Revolutionary Imam, 62. Israeli puppet, 63. American Human Rights, 64. American human rights, 65. Iran's message
3. اقتدار_ایران	
4. GreatSatan	
5. اربعین_شاهچراغ	
6. Commander of Hearts	
7. Raefipour is our voice	
8. Love of Hussein gathers us	
9. QudsDay	
10. FreePalaestine	
11. freepalestine	
12. I am also a soldier of Khamenei	
13. Husseini Generation	
14. School of Hussein	
15. House of Hussein Iran	
16. Only Haider is the commander of the faithful	
17. School of Sayyid Al Shahada (Imam Hussein)	
18. Den of Corruption	
19. Oh Allah, hasten the reappearance of your Guardian	
20. Imam of the nation	
21. Islamic Republic is sacred	
22. Death to Al Saud	
23. ThePromisedSaviour	
24. Provincial soldier	
25. Yaaddi	
26. At your service, O Imam Khamenei	
27. Long live the Islamic Republic of Iran	
28. Hi Imam of my time	
29. Great Satan	
30. Savior of the world	
31. Expel USA	
32. Fadaianeh Mehdi	
33. Quds_Shirif	
34. Do Not Execute The Corrupt	
35. O Oh Allah Bless Muhammad and Accelerate the Relief of Your Representative	
36. AtYourService O Mahdi	
37. Oh God, hurry the wilderness of the vulva	
38. AtYourService O Khamenei	
39. Allah hasten your relief to your representative	
40. ThanksIRGC	
41. Jerusalem is the capital of the resistance	
42. The subversion of the stupid	
43. Homeland IS	
44. Death to Israel	
45. ThePromisedSavior	
46. Fajr Mubarak Decade	
47. O Allah, hasten the return of your representative	
48. Hello or Mehdi	
49. We are the nation of Imam Hussein	
50. Lie Beating and Insulting Mahsa	
51. Mahsa's beatings lie	
52. Mahsa Amini lie	
53. At Your Service Khamenei	
54. Mahsa sedition	
55. Al-Quds is the Axis	
56. Lovers of Qasem	
57. USAHumanRights	
58. Strong Army	
59. Imam of the hearts	
60. At your service, O Khamenei	
61. Revolutionary Imam	
62. Israeli puppet	
63. American human rights	
64. American human rights	
65. Iran's message	

Table B.4: Continuation of Hashtags Affiliated With Islamic Republic

hashtag	translation
66. رزمايش_اقتدار	66. LiveLikeAli,
68. حببي_حسين	67. Authority exercise,
69. ايران_عفيف	68. Chaste Iran,
70. ارش_فداي_ملت_ملت_فداي_ارش	69. Habibi Hasen,
71. فقط_حیدر_امیرالمؤمنین_است	70. Army,
72. زندگ_سگ_اسرائیلها	71. Israeli dog life,
حقوق_بشر_آمریکایی	72. Only Haidar Amir al -Momenin,
73. القدس_هي_المحور	73. Al-Quds is the Axis,
74. مرمدم_کار_پلیس	74. American Human Rights,
75. ایران_قوی	75. People next to the police,
76. مرمدم_کار_پلیس	76. Strong Iran,
77. اللهم_اعل_لویلک_الفرج	77. O Allah, hasten the return of your relief (Imam Zaman/Mahdi),
Khomeini_TheGreatLeader,	78. Khomeini TheGreatLeader,
79. کریم_أهل_البیت	79. Karim Ahl al -Bayt,
80. امیرالمؤمنین_نهضت_حسنی	80. Imam's line,
81. خط_امام	81. Hosseini Movement,
82. نهضت_حسنی	82. Imam Ali,
83. حب_الحسین_یجمعنا	83. Yemen disciplines the Emirates,
84. الین_یؤدب_الامارات	84. Love of Hussein brings us together,
85. امام_نمیخنی	85. Put your green card on the ground,
86. گرین_کارت_را_زمین_بگذار	86. Imam Khomeini,
87. شید_قاسم_سلیمانی	87. Martyr Haj Qasim Soleimani,
88. شید_حج_قاسم_سلیمانی	88. Martyr Qasim Suleimani,
89. اوکراین_تجربه_اعتماد_به_آمریکا	89. Martyr Sayyad khodayi,
90. شید_صیاد_خدایی	90. Ukraine's experience of trusting America,
91. اللهم_صل_علی_محمد_وال_محمد_واعل_فرجهم	91. O Allah, bless Muhammad and the family of Muhammad and hasten the return of your representative,
92. امام_زمان	92. Imam of the time,
93. یا_باب_الحواج	93. Death to the hypocrite,
94. مرگ_بر_منافق	94. O Bab Al-Hawaj,
95. امام_خامنه_ای	95. Put your green card on the ground,
96. گرین_کارت_را_زمین_بگذار	96. Imam Khamenei,
97. لبیک_یا_خامنه_ای	97. Thank you for my leader,
98. سلام_یا_مهدی	98. Hello, O Mahdi,
99. راهنی_پور_تنه_نیست	99. At your service Khamenei,
100. سردار_دله	100. Commander of Hearts,
101. رائفیپور_یا_نهایت	101. Raefipour Is Not Alone,
102. او_محمد_وال_محمد_واعل_فرجهم	102. Oh Allah Bless Muhammad and Accelerate the Relief of Your Representative,
103. روز_قدس	103. Oh God Hasten the Reappearance of your Representative,
104. اللهم_اعل_لویلک_الفرج	104. Qods Day

Table B.5: Hashtags Affiliated With Monarchy

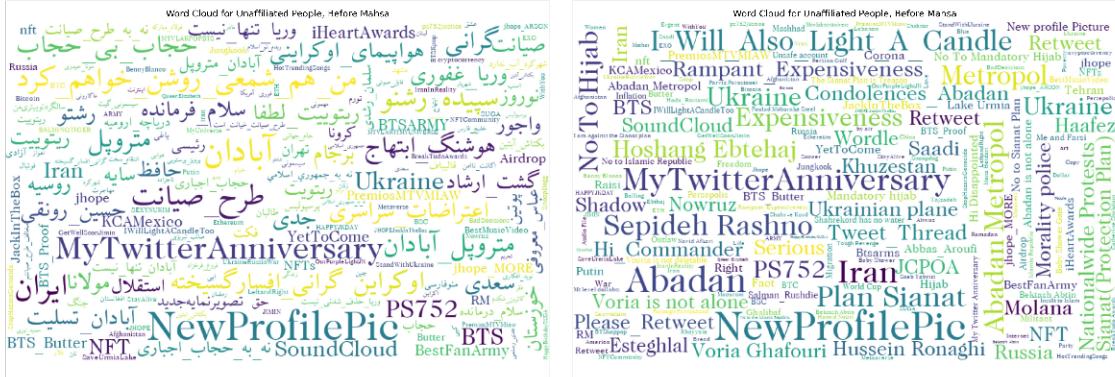
hashtag		translation
1. PAHLAVi4iran,	2. جاوید_رضا_شاه_دوم	1. PAHLAVi4iran, 2. Long live Reza Shah II, 3.
3. فرزندانپرچم	4. تاابدجاویدشاه	Children Of Flag, 4. Until eternity, Long live the King, 5. Iran, the land of Kings, 6. Pahlavi 4 Aran, 7. Referendum with Pahlavi, 8. Kingdom with Pahlavi, 9. Everlasting Reza Shah the Second, 10. Oh Shah of Iran, Return to Iran, 11. Everlasting Reza Shah II Pahlavi III, 12. His Royal Highness Reza Pahlavi, 13. Death to the three corrupt left-wing mullahs, Mujahideen, 14. Long live the King, the symbol of victory, 15. KingdomWithPahlavi, 16. One flag, one nation, 17. King Homeland Freedom, 18. Reza Pahlavi is my representative, 19. Say return, King, 20. Long Last Unified Iran, 21. Eternal Shah, 22. Now, 23. Reza Pahlavi for Iran, 24. Iran That Doesn't Have a Shah, Doesn't Have Accounts, 25. Death to the three corrupt left-wing mullahs, Mujahideen, 26. Long Live Pahlavi, 27. My lawyer is Prince Reza Pahlavi, 28. Shame on the three corrupt left-wing mullahs, Mujahideen, 29. Emperor Reza Pahlavi, 30. Our leader is Pahlavi, whoever denies it is foreign, 31. Shame on the sedition of 57, 32. KingRezaPahlavi, 33. Iran is ready, give the order, Prince, 34. Life, life, Pahlavi, 35. My vote is Pahlavi, 36. National Uprising, 37. We are a great nation, we will take back Iran, 38. Prince, where are you, come to our help, 39. From Cyrus to Pahlavi, 40. Rest in Peace, Shah, 41. We will take back Reza Shah's monarchy, 42. Shame on the sedition of 57, 43. Savior Of Constitution, 44. The prince is the symbol of victory, 45. Reza Pahlavi is my representative, 46. Eternal nation of Iran, 47. Death to the uprising of 57, 48. Javidshah (long live the / Everlasting king), 49. Long live the king, long live the king, long live the king, 50. Iranian leader Pahlavi, 51. RezaPahlaviIsMyRepresentative, 52. رضا_شاه_دوم_پهلوی_سوم
مرگ_بر_روح_الله_مصی_اسی_عبدالله	53. ننگ_بر_فنته_ایران	54. 57_شورش
اینترنیشنال_ایرانستیز	55. مرگ_بر_شورش	56. اینترنشنال_ایرانستیز
خاندان_ایرانساز_پهلوی	57. جاوید_شاه_رمن_پیروزی	57. LongLiveTheKing, 58. شاهزاده_رمن_پیروزی
رضاشاه_بانی_رفاه	58. ننگ_بر_شورش	59. بازگشت_به_مشروطه
رضاشاه_بنامه_تمامه	60. شاه_مین_ابادی	61. خاندان_ایرانساز_پهلوی
مشروطه_پادشاهی	62. نه_داس_نه_عمامه_57	63. رضاشاه_بانی_رفاه
پرچم_ملی_شیروخورشید	64. تماهه	65. RezaShah FounderOfProsperity, 66. National flag of Lion and Sun, 67. Constitutional kingdom

Table B.6: Continuation of Hashtags Affiliated With Monarchy

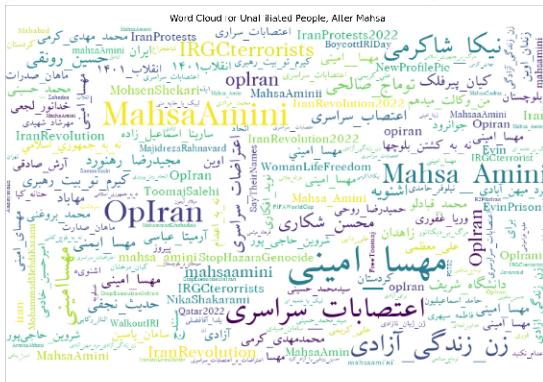
hashtag	translation
66. شاهنشاه_آریامهر 67. جاویدشاه، قانون_اساسی_مشروطه 68. Long live the King, 67. King of Kings Aryamehr,	
69. شاه_رضا_پهلوی 69. Constitutional Constitution, 70. من_به_شاھزاده_رضاپهلوی_وکالت_میدهم 70. I appoint Prince Reza Pahlavi as my representative, 71. God, the King of Homeland, 72.	
71. خدا_شah_میهن 71. PrinceRezaPahlavi, 73. RezaShahTheSecond, 74. جاویدشاه_مز_پیروزی 72. RezaPahlavilsMyRepresentative, 74. Everlasting King symbol of victory, 75. EverlastingKing, 76. I appoint the prince as my representative, 77. The Left Never Understood, 78. Reza Shah, 79. Reza Pahlavi is my representative, 80. National Union of Iran, 81. Reza Shah II, 82. I give Reza Pahlavi a lawyer, 83. Iranian National Revolution, 84. The woman, 85. The National Revolution of Iranians, 86. Salute to the nation-building Pahlavi dynasty, 87. National resurgence with Pahlavi, 88. Long live Savak, 89. LeaderOnlyPahlavi, 90. National Kian, 91. Sedition57, 92. Nation, 93. Revolutionary Bus, 94. Death To Three Corrupt Mullah Leftist Mujahed, 95. Iranstage, 96. Constitutional kingdom, 97. Rest in peace, Reza Shah, 98. Leader Only Pahlavi, 99. Ariamehr, 100. Death To Three Corrupt Mullah Leftist Mujahed, 101. Everlasting Reza Shah II, Pahlavi III, 102. Empress, the pride of Iran, 103. Pahlavi 4 Aran, 104. Constitutional Monarchy for Iran, 105. ReturnToMonarchy4iran, 106. Pahlavi4iran, 107. Return_2_Pahlavi_4_iran, 108. ۱۰۲. شهبانو_نفر_ایران، ۱۰۳. pahlavi4iran، ۱۰۴. پادشاهی_مشروطه_برای_ایران، ۱۰۵. Pahlavi4iran، ۱۰۶. ReturnToMonarchy4iran، ۱۰۷. Return_2_Pahlavi_4_iran، ۱۰۸. ۱۰۹. جاوید_خاندان_ایرانساز_پهلوی، ۱۱۰. Pahlavi4iran، ۱۱۱. جاوید_شah، ۱۱۲. ۱۱۳. پایانده_ایران، ۱۱۴. PAHLAVI4iran، ۱۱۵. PAHLAVI4iran، ۱۱۶. شah_برگد_شah، ۱۱۷. ۱۱۸. فته_۲۲_بهمن، ۱۱۹. ۱۱۸. فته_۲۲_بهمن، ۱۲۰. ارتیاج_سرخ_و_سیاه، ۱۲۱. ۱۲۰. رضاشاه_بزرگ، ۱۲۲. فته_شah_بزرگ، ۱۲۳. ۱۲۲. رضا_شah_بزرگ، ۱۲۴. نوروز_۱۳۵۸_شاهنشاهی، ۱۲۵. پادشاهی_پهلوی، ۱۲۶. یک_ملت_یک_برچم، ۱۲۷. مصدق_خائن، ۱۲۸. رضاشاه_بزرگ، ۱۲۹. ۱۲۷. قیام_ملی، ۱۳۰. رضا_شah_روحت_شah، ۱۳۱. ۱۳۰. دوران_طلای_پهلوی، ۱۳۲. شah_روحت_ایران_رفت، ۱۳۳. فته_۲۲_بهمن، ۱۳۴. پدر_ایران_نوین، ۱۳۵. زنده_باد_خاندان_ایرانساز_پهلوی، ۱۳۶. پهلویسم_تہا_راه_نجات_ایران، ۱۳۷. انقلاب_شah_و_مردم، ۱۳۸. رضاشاه_دوم_پهلوی_سوم، ۱۳۹. جاویدرضاشاه_دوم.	

## **Appendix C**

### **Algorithmically Affiliation Assigned Community Wordclouds**

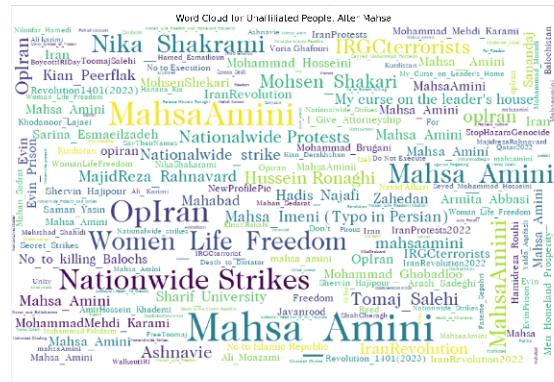


(a) Hashtags in original language before Mahsa (January 2022 to September 14th, 2022)



(c) Hashtags in original language after Mahsa (September 14th, 2022 to April 16th, 2023)

(b) Hashtags in English before Mahsa (January 2022 to September 14th, 2022)



(d) Hashtags in English after Mahsa (September 14th, 2022 to April 16th, 2023)

Figure C.1: Top Hashtags Used by Unaffiliated People Community With Deterministic Ideological Affiliation Assignment: These word clouds represent the most frequently used hashtags by the "Unaffiliated People" community, before and after Mahsa. The word size is proportional to number of users using the hashtag. They provide insight into the key topics and sentiments prevalent within this group during these time periods. "I will also light a candle" is a hashtag used to show solidarity with the victims of the Ukrainian flight PS752.



(a) Hashtags in original language before Mahsa (January 2022 to September 14th, 2022)



(c) Hashtags in original language after Mahsa (September 14th, 2022 to April 16th, 2023)



(b) Hashtags in English before Mahsa (January 2022 to September 14th, 2022)

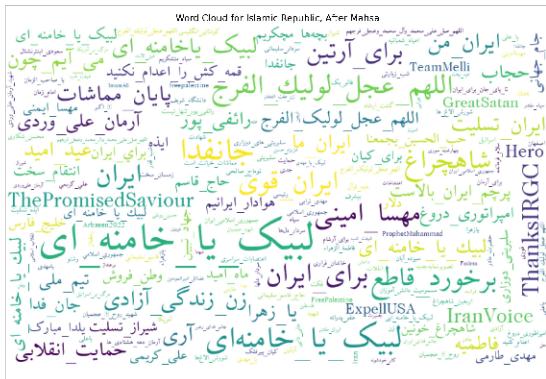


(d) Hashtags in English after Mahsa (September 14th, 2022 to April 16th, 2023)

Figure C.2: Top Hashtags Used by Monarchy Community With Deterministic Ideological Affiliation Assignment: These word clouds represent the most frequently used hashtags by the "Monarchy" community, before and after Mahsa. The word size is proportional to number of users using the hashtag. They provide insight into the key topics and sentiments prevalent within this group during these time periods.

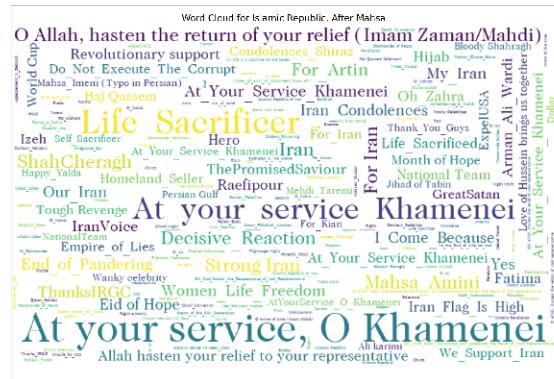


(a) Hashtags in original language before Mahsa (January 2022 to September 14th, 2022)



(c) Hashtags in original language after Mahsa (September 14th, 2022 to April 16th, 2023)

(b) Hashtags in English before Mahsa (January 2022 to September 14th, 2022)



(d) Hashtags in English after Mahsa (September 14th, 2022 to April 16th, 2023)

Figure C.3: Top Hashtags Used by Islamic Republic Community With Deterministic Ideological Affiliation Assignment: These word clouds represent the most frequently used hashtags by the "Islamic Republic" community, before and after Mahsa. The word size is proportional to number of users using the hashtag. They provide insight into the key topics and sentiments prevalent within this group during these time periods.



Figure C.4: Top Hashtags Used by MEK Community With Deterministic Ideological Affiliation Assignment: These word clouds represent the most frequently used hashtags by the "MEK" community, before and after Mahsa's influence. They provide insight into the key topics and sentiments prevalent within this group during these periods.

## Appendix D

### First 20 M.H.P Tweets

Table D.1: First 20 tweets containing مرد میهن آبادی (Men Homeland Prosperity) trigram.

Tweet Information	Tweet and Translation
<b>Date Tweeted:</b> 2022-09-25 18:07:26 <b>Username:</b> Pa*** <b>Affiliation:</b> Monarchy <b>Followers:</b> 13258 <b>Date Joined:</b> 2018-08-22 <b>Total Tweets:</b> 38103	زن، زندگی، آزادی مرد، میهن، آبادی #انقلاب_ایران #مهسا_امینی #IranRevolution <i>( Women, life, freedom Men, homeland, prosperity #Iran_Revolution #Mahsa_Amini )</i>
<b>Date Tweeted:</b> 2022-09-25 18:13:48 <b>Username:</b> Nf*** <b>Affiliation:</b> Monarchy <b>Followers:</b> 347 <b>Date Joined:</b> 2018-06-10 <b>Total Tweets:</b> 14177	زن، زندگی، آزادی مرد، میهن، آبادی #مهسا_امینی <i>( Women, life, freedom Men, homeland, prosperity )</i>
<b>Date Tweeted:</b> 2022-09-25 18:22:53 <b>Username:</b> Om*** <b>Affiliation:</b> Monarchy <b>Followers:</b> 98 <b>Date Joined:</b> 2015-05-26 <b>Total Tweets:</b> 50	زن زندگی آزادی مرد میهن آبادی #مهسا_امینی <i>( Women, life, freedom Men, homeland, prosperity #Mahsa_Amini )</i>

Table D.1: First 20 tweets containing مرد میهن آبادی (Men Homeland Prosperity) trigram. (Continued)

Tweet Information	Tweet and Translation
<b>Date Tweeted:</b> 2022-09-25 18:47:44 <b>Username:</b> Ca*** <b>Affiliation:</b> Monarchy <b>Followers:</b> 408 <b>Date Joined:</b> 2018-03-20 <b>Total Tweets:</b> 1343	<p>چقدر قشنگ @Pa*** :rose::rose::rose:          تنها راه رهای، ایران پادشاهی :red_heart::sun::lion::white_heart::green_heart:          زن زندگی آزادی          مرد میهن آبادی          تنها راه رهایی          ایران پادشاهی</p> <p>( @Pa*** How beautiful:rose::rose::rose:          The only way to liberty, Monarchic          Iran:red_heart::sun::lion::white_heart::green_heart:          Women, life, freedom          Men, homeland, prosperity          The only way to liberty          Monarchic Iran )</p>
<b>Date Tweeted:</b> 2022-09-25 18:49:18 <b>Username:</b> Ca*** <b>Affiliation:</b> Monarchy <b>Followers:</b> 408 <b>Date Joined:</b> 2018-03-20 <b>Total Tweets:</b> 1343	<p>چقدر قشنگ :rose::rose::rose:          زن زندگی آزادی          مرد میهن آبادی          تنها راه رهایی          ایران پادشاهی :green_heart::white_heart::lion::sun::red_heart:          #Mahsa_Amini #جاویدشاه#</p> <p>( How beautiful:rose::rose::rose:          Women, life, freedom          Men, homeland, prosperity          The only way to liberty          Monarchic Iran          :green_heart::white_heart::lion::sun::red_heart:          #Mahsa_Amini          #EverlastingKing )</p>
<b>Date Tweeted:</b> 2022-09-25 19:32:43 <b>Username:</b> Ca*** <b>Affiliation:</b> Monarchy <b>Followers:</b> 408 <b>Date Joined:</b> 2018-03-20 <b>Total Tweets:</b> 1343	<p>زن زندگی آزادی          مرد میهن آبادی          تنها راه رهایی          ایران پادشاهی #Mahsa_Amini #جاویدشاه#</p> <p>( Women, life, freedom          Men, homeland, prosperity          The only way to liberty          Monarchic Iran          #Mahsa_Amini          #EverlastingKing )</p>

Table D.1: First 20 tweets containing مرد میهن آبادی (Men Homeland Prosperity) trigram. (Continued)

Tweet Information	Tweet and Translation
<b>Date Tweeted:</b> 2022-09-25 19:38:05 <b>Username:</b> li*** <b>Affiliation:</b> Monarchy <b>Followers:</b> 139 <b>Date Joined:</b> 2012-01-24 <b>Total Tweets:</b> 198	زن زندگی آزادی مرد میهن آبادی تنها راه رهایی ایران پادشاهی #مهسا_امینی #جاوید_شاه #شاہنشاہ_روحت_شاد  <i>( Women, life, freedom Men, homeland, prosperity The only way to liberty Monarchic Iran #Mahsa_Amini #Everlasting_King #King_MayYourSoulRestInPeace )</i>
<b>Date Tweeted:</b> 2022-09-25 20:13:36 <b>Username:</b> Ca*** <b>Affiliation:</b> Monarchy <b>Followers:</b> 408 <b>Date Joined:</b> 2018-03-20 <b>Total Tweets:</b> 1343	زن زندگی آزادی مرد میهن آبادی تنها راه رهایی ایران پادشاهی #مهسا_امینی #جاویدشاه  <i>( Women, life, freedom Men, homeland, prosperity The only way to liberty Monarchic Iran #Mahsa_Amini #EverlastingKing )</i>
<b>Date Tweeted:</b> 2022-09-25 20:29:48 <b>Username:</b> Ca*** <b>Affiliation:</b> Monarchy <b>Followers:</b> 408 <b>Date Joined:</b> 2018-03-20 <b>Total Tweets:</b> 1343	زن زندگی آزادی @ar*** @Me*** مرد میهن آبادی تنها راه رهایی ایران پادشاهی #مهسا_امینی #جاویدشاه  <i>( @Me*** @ar*** Women, life, freedom Men, homeland, prosperity The only way to liberty Monarchic Iran #Mahsa_Amini #EverlastingKing )</i>

Table D.1: First 20 tweets containing مرد میهن آبادی (Men Homeland Prosperity) trigram. (Continued)

Tweet Information	Tweet and Translation
<b>Date Tweeted:</b> 2022-09-26 01:21:56 <b>Username:</b> Si*** <b>Affiliation:</b> Unaffiliated People <b>Followers:</b> 4347 <b>Date Joined:</b> 2009-09-07 <b>Total Tweets:</b> 1829	مرد، میهن، آبادی زن، زندگی، آزادی #مسا_امینی #MahsaAmini #OpIran ( <i>Men, homeland, prosperity</i> <i>Women, life, freedom</i> <i>#Mahsa_Amini</i> <i>#MahsaAmini</i> <i>#OpIran</i> )
<b>Date Tweeted:</b> 2022-09-26 01:48:36 <b>Username:</b> Pa*** <b>Affiliation:</b> Unaffiliated People <b>Followers:</b> 525 <b>Date Joined:</b> 2016-12-27 <b>Total Tweets:</b> 4600	@Si*** @چه شعار خوبی، مفهوم تساوی رو در خودش داره: و میشه حق جایجا هم بشه که هیچ تفاوتی نداشته باشه و اهمیت عدالت و برابری رو نشون بدhe مرد، میهن، آبادی زن، زندگی، آزادی زن، میهن، آبادی مرد، زندگی، آزادی #مسا_امینی #Mahsa_Amini ( @Si*** <i>What a good slogan, it has the concept of equality in it :OK_hand: and it can even be swapped without any difference, showing the importance of justice and equality</i> <i>Men, homeland, prosperity</i> <i>Women, life, freedom</i> <i>Women, homeland, prosperity</i> <i>Men, life, freedom</i> <i>#Mahsa_Amini</i> <i>#MahsaAmini</i> )
<b>Date Tweeted:</b> 2022-09-26 05:10:26 <b>Username:</b> ir*** <b>Affiliation:</b> Monarchy <b>Followers:</b> 501 <b>Date Joined:</b> 2021-01-12 <b>Total Tweets:</b> 907	مرد، میهن، آبادی زن، زندگی، آزادی #مسا_امینی #MahsaAmini #OpIran ( <i>Men, homeland, prosperity</i> <i>Women, life, freedom</i> <i>#Mahsa_Amini</i> <i>#MahsaAmini</i> <i>#OpIran</i> )

Table D.1: First 20 tweets containing مرد میهن آبادی (Men Homeland Prosperity) trigram. (Continued)

Tweet Information	Tweet and Translation
<b>Date Tweeted:</b> 2022-09-26 16:39:46 <b>Username:</b> Pa*** <b>Affiliation:</b> Monarchy <b>Followers:</b> 13258 <b>Date Joined:</b> 2018-08-22 <b>Total Tweets:</b> 38103	<p>این شعار مال عبدالله اوجلان ه. مال کوئیست ها و فیلیست هاست. زیباست اما ناقصه. کاملش اینه: زن زندگی آزادی مرد میهن آبادی</p> <p>( @Ir*** @Ba*** This slogan belongs to Abdullah Öcalan. It's for the communists and feminists. It's beautiful but incomplete. The complete one is: Women, life, freedom Men, homeland, prosperity )</p>
<b>Date Tweeted:</b> 2022-09-26 16:50:49 <b>Username:</b> th*** <b>Affiliation:</b> Monarchy <b>Followers:</b> 1667 <b>Date Joined:</b> 2015-01-27 <b>Total Tweets:</b> 18394	<p>زن زندگی آزادی شاه شادی آبادی</p> <p>زن زندگی آزادی مرد میهن آبادی</p> <p>#IranRevolution #MahsaAmini #مهسا_امینی</p> <p>( Women, life, freedom King, happiness, prosperity Women, life, freedom Men, homeland, prosperity #Mahsa_Amini #MahsaAmini #IranRevolution )</p>
<b>Date Tweeted:</b> 2022-09-26 23:28:17 <b>Username:</b> Nf*** <b>Affiliation:</b> Monarchy <b>Followers:</b> 347 <b>Date Joined:</b> 2018-06-10 <b>Total Tweets:</b> 14177	<p>زن زندگی آزادی مرد میهن آبادی</p> <p>( @Be*** @Em*** @di*** @da*** Women, life, freedom Men, homeland, prosperity )</p>

Table D.1: First 20 tweets containing مرد میهن آبادی (Men Homeland Prosperity) trigram. (Continued)

Tweet Information	Tweet and Translation
<b>Date Tweeted:</b> 2022-09-26 23:37:51 <b>Username:</b> Ti*** <b>Affiliation:</b> Monarchy <b>Followers:</b> 584 <b>Date Joined:</b> 2022-05-26 <b>Total Tweets:</b> 1354	<p>چه زیبا گفت...</p> <p>زن، زندگی، آزادی مرد، میهن، آبادی</p> <p>عاشق این همیستگی ایرانم. چهل و سه سال جاعش تلاش کرد تفرقه بندازه و حکومت کنه ولی آخرش مردم متحد تراز قبل تو دهنش زدن. در یک کلام جانم فدای ایران ما، نه جمهوری اسلامی اونا...</p> <p>#خامنه‌ای_جنایتکار #مهسا_امینی #OpIran #Anonymous</p> <p>( <i>How beautifully said... Women, life, freedom Men, homeland, prosperity</i> <i>I love this unity of Iran. For forty-three years, the regime tried to sow division and rule, but in the end, the united people prevailed. In one word, my life is for our Iran, not their Islamic Republic...</i> #Khamenei_Criminal #Mahsa_Amini #OpIran #Anonymous )</p>
<b>Date Tweeted:</b> 2022-09-26 23:45:11 <b>Username:</b> Av*** <b>Affiliation:</b> Monarchy <b>Followers:</b> 5291 <b>Date Joined:</b> 2022-05-18 <b>Total Tweets:</b> 1955	<p>زن، زندگی، آزادی مرد، میهن، آبادی</p> <p>#OpIran #مهسا_امینی</p> <p>( <i>Women, life, freedom Men, homeland, prosperity</i> #OpIran #Mahsa_Amini )</p>
<b>Date Tweeted:</b> 2022-09-27 00:24:21 <b>Username:</b> nc*** <b>Affiliation:</b> Unaffiliated People <b>Followers:</b> 662 <b>Date Joined:</b> 2015-09-13 <b>Total Tweets:</b> 2339	<p>زن، زندگی، آزادی مرد، میهن، آبادی</p> <p>کبر تو جمهوری اسلامی #مهسا_امینی برای #اعتصابات_سراسری</p> <p>( <i>Women, life, freedom Men, homeland, prosperity</i> <i>Down with the Islamic Republic</i> #Mahsa_Amini for #NationwideStrike )</p>

Table D.1: First 20 tweets containing مرد میهن آبادی (Men Homeland Prosperity) trigram. (Continued)

Tweet Information	Tweet and Translation
<b>Date Tweeted:</b> 2022-09-27 00:26:33 <b>Username:</b> nc*** <b>Affiliation:</b> Unaffiliated People <b>Followers:</b> 662 <b>Date Joined:</b> 2015-09-13 <b>Total Tweets:</b> 2339	<p style="text-align: right;">زَن، زَنْدَگَى، آزادِى @Tr***</p> <p style="text-align: right;">مرد، میهن، آبادِى</p> <p style="text-align: right;">کیر تو جمهوری اسلامی</p> <p style="text-align: right;">#مهسا_امینی برای #اعتصابات_سراسُری</p> <p style="text-align: right;">( @Tr*** Women, life, freedom Men, homeland, prosperity F*** the Islamic Republic #Mahsa_Amini for #NationwideStrike )</p>
<b>Date Tweeted:</b> 2022-09-27 00:27:25 <b>Username:</b> ne*** <b>Affiliation:</b> Unaffiliated People <b>Followers:</b> 662 <b>Date Joined:</b> 2015-09-13 <b>Total Tweets:</b> 2339	<p style="text-align: right;">زَن، زَنْدَگَى، آزادِى ... @on*** @nz***</p> <p style="text-align: right;">مرد، میهن، آبادِى ...</p> <p style="text-align: right;">کیر تو جمهوری اسلامی .</p> <p style="text-align: right;">#مهسا_امینی برای #اعتصابات_سراسُری</p> <p style="text-align: right;">( @nz*** @on*** Women, life, freedom Men, homeland, prosperity F*** the Islamic Republic #Mahsa_Amini for #NationwideStrike )</p>