Telellmgram : Persian community analysis in Telegram using LLMs and RAG

Omid Mollaei
omidmollaei@ut.ac.ir

1 Problem Statement

In the contemporary digital landscape, social media platforms have evolved into critical mirrors of real-world communities, serving as dynamic arenas for discourse, opinion-sharing, and collective expression. Among these platforms, Telegram has emerged as a dominant force, particularly in regions like Iran, where it boasts over **50 million active users**. Its encrypted, channel-based structure fosters open discussions on diverse topics—ranging from politics and social movements to entertainment and personal narratives—making it a rich yet underexplored source of public sentiment.

1.1 Benefits of Analyzing Telegram Content

The analysis of conversations and interactions on Telegram offers a wide range of advantages across different domains such as policy-making, business, and social research. By applying language models to Persian Telegram communities, we can uncover patterns that are otherwise difficult to capture through traditional methods. Below, the key benefits are outlined and discussed in detail:

- **Understanding Public Opinion**: Telegram provides an open environment where users freely share their ideas, concerns, and preferences without the formal restrictions of surveys or interviews. By analyzing these conversations, researchers can determine what people like or dislike in relation to products, government policies, or social events. For instance, sudden increases in complaints about rising prices or poor infrastructure can serve as early indicators of public dissatisfaction. Furthermore, detecting "hot topics" is possible when certain words, phrases, or hashtags begin to appear more frequently. This helps in identifying emerging controversies or public debates at their early stages, allowing decision-makers to respond before issues escalate.

- **Helping Make Better Decisions**: The insights derived from Telegram data are highly valuable for both public and private sectors. Governments, for example, can monitor which societal problems are most frequently discussed—such as unemployment, housing shortages, or transportation difficulties—and then allocate resources accordingly. Businesses also benefit from this type of analysis, as they can discover consumer needs and expectations by observing discussions in professional or hobby-related groups. Unlike traditional market research, which is often slow and costly, Telegram analysis provides immediate, real-world feedback, making decisions more evidence-based and responsive.

- **Improving Marketing**: Companies and organizations can use Telegram analysis to strengthen their marketing strategies. One key advantage is brand monitoring: by tracking how often their products or services are mentioned and evaluating the sentiment behind these mentions, businesses gain a clearer picture of their reputation among consumers. In addition, the analysis helps identify influential users whose posts are widely shared or commented upon. These individuals can serve as valuable partners for promotional campaigns, as they already command trust and attention within their communities. This approach is particularly effective in niche Persian-speaking markets, where word-of-mouth and peer recommendations often play a stronger role than conventional advertising.

- **Finding Key People and Groups**: Social

network analysis applied to Telegram content allows researchers to identify central figures who shape discussions. These "key nodes" are users whose posts are shared or referenced by many others, making them critical in the spread of information. Recognizing such individuals helps in mapping the flow of influence within communities. At the same time, analytical methods can also uncover coordinated or suspicious behavior, such as clusters of fake accounts posting identical content simultaneously. Detecting these patterns is vital for ensuring information integrity and for preventing manipulative campaigns within online spaces.

- **Discovering Why Things Happen**: Beyond describing what people are discussing, Telegram analysis can also reveal the underlying causes of social phenomena. For example, by examining conversations leading up to a protest, we can identify which issues—economic hardship, political grievances, or cultural debates—sparked collective action. Similarly, the analysis of false information, or "fake news," allows researchers to trace how misleading claims originate and spread across different groups. Observing how messages evolve as they are reposted highlights the mechanisms by which narratives are shaped, exaggerated, or reframed during circulation.
- **Predicting What Might Happen**: A forward-looking benefit of Telegram analysis is its predictive power. By carefully monitoring the language used in discussions about protests, strikes, or political movements, analysts can estimate the likelihood and potential scale of upcoming events. This is especially valuable for policy-makers who must prepare in advance for social mobilization. Similarly, trend forecasting becomes possible when examining how early adopters within online communities talk about new technologies, fashion items, or cultural practices. The ability to predict future developments based on current conversations can provide a competitive advantage in both governance and business.

In addition to these specific applications, the overall usefulness of Telegram analysis stems from several unique characteristics:

- It reflects **authentic, unfiltered opinions** that people share voluntarily, unlike the more artificial answers often given in formal surveys.
- It delivers **rapid insights**, producing results in real time rather than requiring the long cycles of traditional research.
- It enables the discovery of **hidden connections and patterns** that would be invisible without computational analysis, such as links between users, communities, or evolving narratives.

Overall, analyzing Telegram content with modern language models represents a powerful tool for social understanding, decision-making, and forecasting in Persian-speaking communities.

1.2 The Role of Large Language Models (LLMs)

Traditional approaches to language analysis, such as rule-based systems or earlier generations of machine learning models, often fall short when applied to real-world Telegram conversations. These methods were typically trained on clean, formal text such as news articles or academic writing, which makes them poorly equipped for the informal, dynamic, and often noisy language used in social media. Telegram messages in Persian-speaking communities present a variety of challenges: frequent code-switching between Persian and English, the use of slang and internet-specific expressions, spelling mistakes due to rapid typing, and complex ideas that unfold across multiple short messages rather than in neatly structured paragraphs. Moreover, cultural references, sarcasm, and humor are heavily context-dependent, and simple models fail to capture these nuances. For example, a traditional neural network may incorrectly interpret a sarcastic comment as positive sentiment simply because it contains positive words, while missing the ironic tone conveyed by the broader conversation.

Large Language Models (LLMs), such as GPT-4, represent a significant advancement over these earlier tools. Because they are trained on vast and diverse datasets that include informal internet conversations, online forums, and multilingual text, they possess a much richer understanding of how people actually communicate online.

They are able to process both formal and informal registers, adapt to code-switching within a single sentence, and interpret idiomatic expressions in context. This makes them particularly powerful for analyzing Telegram communities, where the language used is often a blend of Persian grammar, English loanwords, emojis, humor, and cultural shorthand. LLMs excel at maintaining context across long threads of conversation, correctly interpreting the shifting meanings of words, and identifying subtle emotional undertones. For instance, when encountering the phrase “ !” (great!), the model can distinguish whether it is an expression of genuine enthusiasm or sarcastic frustration by considering the surrounding dialogue, something most traditional tools would misclassify.

For the TeleLLMgram project, all of the analysis is conducted using cloud-based LLM APIs. These services provide access to state-of-the-art models that can perform complex reasoning tasks, multi-turn summarization, and semantic analysis at scale. By relying on powerful cloud models, we can automatically sift through thousands of Telegram posts to detect meaningful patterns, summarize heated debates, and highlight unusual behaviors such as coordinated posting. This automation is critical: carrying out such tasks manually would not only be extremely time-consuming but also prone to human bias and error. The efficiency and accuracy provided by LLMs enable us to process Telegram data in ways that would otherwise be impossible.

The specific strengths of LLMs are particularly relevant for Persian-language Telegram communities. Persian online communication is highly dynamic and filled with cultural layers that cannot be easily decoded by generic sentiment analysis tools. Users frequently mix Persian syntax with English words written in Latin script, such as in the phrase “Awlie!”—a colloquial borrowing of the Persian word “” (excellent) blended with English-style spelling. An LLM trained on multilingual internet text can interpret this hybrid form naturally. Similarly, LLMs can capture the cultural resonance of political, social, or religious terminology that carries meanings beyond their literal definitions. For instance, references to specific events, slogans, or historical figures may evoke strong emotional reactions in Persian discourse, and LLMs are better equipped to recognize these nuances than simple keyword-based methods.

Another advantage is the scalability of cloud-based LLMs. Unlike human analysts, who would take weeks or months to process large datasets, LLMs can provide near real-time analysis of Telegram activity. This includes generating summaries of long discussions, identifying emerging trends before they become mainstream, and detecting coordinated misinformation campaigns. Beyond descriptive analysis, LLMs also offer predictive potential: by analyzing ongoing conversations, they can forecast possible future developments such as rising public dissatisfaction, the spread of rumors, or the likelihood of collective action.

To ensure reliability, our project incorporates mechanisms to validate the outputs of LLMs. We use multiple query formulations to cross-check interpretations, and we apply additional filtering techniques to reduce noise or irrelevant results. For cost efficiency, we carefully design prompts and batch queries to maximize the analytical value of each API call. These practical considerations make it possible to benefit from the advanced capabilities of LLMs while keeping the system both efficient and robust.

In summary, LLMs are not just useful tools but essential components of the TeleLLMgram project. Their ability to handle informal, multilingual, and culturally nuanced communication patterns makes them uniquely suited to analyzing Persian Telegram content. Without LLMs, much of the subtlety, richness, and predictive potential of these conversations would remain inaccessible to researchers and decision-makers.

1.3 Project Value and Practical Applications

The TeleLLMgram project demonstrates the potential of large language models to deliver actionable insights from Persian Telegram communities. Unlike traditional social media monitoring systems that focus on surface-level metrics such as follower counts or simple sentiment scores, our approach provides deep, context-aware analysis of organic conversations. This allows stakeholders across business, government, academia, and civil society to make decisions based on authentic public discourse rather than limited or biased samples. Importantly, this approach has the capacity to cover—and in many cases surpass—all previous methods of analysis such as surveys, polls, or small-scale qualitative studies, by capturing opinions and interactions in real time at a much larger

scale.

- **Economic and Business Applications**: Companies operating in consumer markets can use this system to better understand real customer experiences. For instance, brands such as Digikala can identify recurring frustrations about delivery times, packaging quality, or after-sales support by analyzing complaint clusters in shopping-related channels. Financial institutions like Mellat Bank may detect growing concerns about digital banking reliability by observing discussion trends in finance groups, enabling early interventions before customer dissatisfaction becomes widespread. Startups such as SnappFood can identify underserved areas by monitoring how often residents in certain districts request specific restaurants or food options, allowing them to expand services strategically. Unlike surveys, which are often limited by sampling bias, Telegram analysis captures unfiltered, organic discussions that reveal actual pain points and unmet needs.

- **Political and Public Sector Applications**: Public authorities can leverage this technology to improve governance and responsiveness. Municipal governments, for example, can prioritize infrastructure repairs by clustering citizen complaints about electricity blackouts or water shortages across neighborhood channels. Health organizations, such as the Ministry of Health, can detect shortages of critical medicines faster than official reporting mechanisms by monitoring pharmacy discussion groups. Education authorities can uncover systemic curriculum issues by analyzing teacher conversations about recurring classroom difficulties, highlighting areas where reform or additional resources are most needed. On the political side, policymakers can track shifts in public opinion toward new laws or government policies by observing how narratives form and spread across Telegram, providing a more immediate and representative picture than traditional polling methods.

- **Social and Cultural Insights**: Beyond formal institutions, this project provides tools to understand broader social dynamics. Civil society organizations can detect when local grievances are escalating into collective action, allowing for proactive engagement before tensions rise. Cultural researchers can examine how humor, slang, and memes spread across Telegram channels, offering insights into the evolving identity of Persian-speaking youth. Social activists may also benefit by identifying influential voices and communities that shape discussions around rights, freedoms, or social justice. This provides a richer and more grounded view of societal trends than older qualitative approaches, which were often restricted to small focus groups or limited datasets.

- **Research and Media Applications**: Academic institutions such as the University of Tehran can use the TeleLLMgram system to study how political narratives evolve differently in closed student networks versus public discussion groups, offering evidence-based perspectives on youth engagement. News organizations like Hamshahri can detect emerging stories by spotting sudden spikes in local conversations long before they are reported in mainstream outlets. Similarly, market research firms can enrich their traditional methodologies by incorporating large-scale, real-time analysis of consumer conversations, bridging the gap between qualitative insights and quantitative breadth.

- **Open Access and Wider Impact**: To ensure that these benefits extend beyond large corporations or state institutions, we have made analysis pipelines, anonymized datasets, and documentation openly available through GitHub: https://github.com/omid-mollaei/Telellmgram. This open approach empowers small businesses, independent journalists, and graduate students to conduct advanced social media research without expensive proprietary systems. A small retailer can monitor its brand reputation as effectively as a multinational company, while a graduate student can study political communication patterns at a scale that would previously have required institutional resources. I have plan to continue working on this project later and improve it and implement what I have in my mind.

Overall, the TeleLLMgram project illustrates how analyzing social media with LLMs can go beyond the limitations of traditional surveys, focus groups, or manual monitoring. It offers not only faster and more cost-efficient analysis but also richer, culturally sensitive insights into political, social, and economic realities. This positions LLM-based Telegram analysis as a transformative tool for both immediate decision-making and long-term research. **(Note : Since my dataset is large, i will just upload a sample of it to this github repository)**

1.4 Project Team and Development Process

The TeleLLMgram project was initially designed as a collaborative effort between two students, Omid Mollaei and M.H. Khodad. The intention was to divide responsibilities across different aspects of the system, such as data collection, preprocessing, model integration, and evaluation. However, during the course, M.H. Khodad decided to withdraw, which left the continuation of the project solely in the hands of Omid Mollaei.

As a result of this unexpected change, the scope of the project had to be adjusted. Several features and experiments originally planned for a larger team were either simplified or postponed. The focus was shifted toward implementing a core end-to-end pipeline that could reliably demonstrate the value of large language models for analyzing Persian Telegram content. While this meant that some of the broader goals—such as incorporating additional datasets, testing multiple model families in parallel, or developing more advanced visualization dashboards—could not be fully realized within the timeframe, the project nevertheless succeeded in producing a functional, practical system that highlights the key benefits of this approach.

In this sense, the project also demonstrates the challenges of research under constrained resources. Although the final implementation was smaller in scale than originally envisioned, the essential objectives were met, and the work provides a strong foundation for future extensions should additional collaborators or resources become available.

2 Related Work

2.1 Historical Phases of Social Media Analysis

Social media analysis has evolved through three key phases: early **structural network studies**, content-driven approaches, and, more recently, the integration of large language models (LLMs). The majority of early research concentrated on the *structure* of social networks—particularly Twitter—using graph-based metrics and community detection to study political discourse and information diffusion. For example, Conover et al. (2011) investigated political polarization through Twitter follower and retweet networks, while Pak and Paroubek (2010) demonstrated how tweets could be used for sentiment classification. Although influential, these works relied on publicly available Twitter APIs and English-language content, making them poorly suited for analyzing Persian-language platforms such as Telegram, which rose in prominence after Twitter restrictions in Iran (Vaziripour et al., 2018; Al-Rawi, 2022).

2.2 Persian Social Media NLP and Resource Development

Compared to English social media research, **Persian content analysis has been less studied**, and often without the rich structural data available from Twitter. Recent surveys track the evolution of Persian sentiment analysis, from lexicon-based methods to deep learning frameworks (Rajabi and Valavi, 2021; Heydari et al., 2024). Transformer-based resources such as ParsBERT (Farahani et al., 2020) and the more recent FarSSiBERT (Sadjadi et al., 2024) have advanced Persian NLP capabilities, particularly for informal social network texts. Entity linking in Persian social media has been addressed by ParsEL 1.0, which achieved strong performance on Telegram-sourced text (Asgari-Bidhendi et al., 2020). Topic detection methods using graph embeddings have also incorporated Persian Telegram datasets (Ranjbar-Khadivi et al., 2023), though these still primarily target textual features.

2.3 Retrieval-Augmented Generation

The introduction of Transformers (Vaswani et al., 2017) and the Retrieval-Augmented Generation (RAG) paradigm (Lewis et al., 2020) has enabled richer content understanding by combining retrieval methods (Izacard and Grave, 2020)

with generative models. However, **no prior work has applied RAG-based architectures to Persian Telegram content** in a way that integrates both multimodal features and network structural analysis.

2.4 How Our Work Differs

The TeleLLMgram project introduces several distinct contributions beyond existing research. While there have been numerous studies on social media analysis in Persian contexts, most prior work has focused on platforms like Twitter or relied on traditional natural language processing methods with limited capacity to capture cultural and linguistic nuances. Our work diverges from these earlier approaches in several important ways:

- **From Twitter to Telegram**: A significant portion of the academic literature in Persian social media analysis has concentrated on Twitter, mainly because of the availability of public data and established network analysis tools. However, Persian-speaking communities are far more active on Telegram, which functions as a central hub for political debates, commercial exchanges, cultural discussions, and everyday communication. By shifting the focus from Twitter to Telegram, our project fills a critical gap in the literature, bringing attention to a platform that has far greater influence in the region but has been comparatively understudied. This change of perspective enables a more accurate representation of Persian digital society.

- **RAG + LLM Pipeline**: Unlike previous studies that often rely on static classification or simple sentiment analysis, our project implements a retrieval-augmented generation (RAG) mechanism. Even though our implementation is intentionally lightweight, it allows the system to retrieve relevant contextual information from conversation archives and feed it into the large language model for more accurate and context-sensitive analysis. This ensures that the interpretations are not limited to isolated messages but are grounded in the broader discourse. In practice, this means that the system can answer higher-level questions, such as why a topic is gaining attention, rather than merely reporting that it is trending.

- **Network-Enhanced Interpretation**: Social media cannot be fully understood through textual analysis alone. Our project incorporates user and group network attributes into the interpretation process, linking patterns of content with patterns of interaction. For example, we do not simply identify which words or phrases are becoming more common; we also consider which accounts or groups are responsible for spreading them, and how central these actors are in their communities. This integration of content and network analysis produces richer insights into how information spreads and which voices carry the most influence in Persian Telegram ecosystems.

- **Multi-Purpose Usage**: One of the strengths of TeleLLMgram is its versatility. The same analysis pipeline can serve a wide variety of goals depending on the needs of the user. For instance, policymakers may apply it to understand public dissatisfaction and detect early signs of protest movements. Businesses can adapt it to monitor consumer sentiment and improve customer engagement. Journalists may use it to identify emerging news stories or investigate the spread of misinformation. Academic researchers can employ it to study cultural discourse, identity construction, or the evolution of political narratives. This flexibility distinguishes our work from many earlier studies, which were narrowly designed for a single purpose such as sentiment classification or fake-news detection. By contrast, our framework is adaptable, scalable, and open to being extended for multiple real-world applications.

In summary, our project differs from prior work not only in its platform focus but also in its methodological approach and flexibility. By analyzing Telegram rather than Twitter, combining retrieval-augmented generation with large language models, linking content with network structures, and enabling multi-purpose applications, TeleLLMgram offers a comprehensive and adaptable framework for understanding Persian online communities in ways that previous studies have not addressed.

3 Dataset

The dataset for this project has been constructed from Persian Telegram media, consisting of both channels and groups, using the official Telegram Desktop export feature. In the current workflow, all data are exported in `.json` format, which provides a structured and machine-readable representation of messages. This raw format is then transformed into tabular form for analysis.

The dataset presently includes **30 media sources** — a combination of channels and groups that were carefully selected for their relevance to the project's goals. The selection process emphasized diversity of content and viewpoints while avoiding sources that would introduce excessive bias or noise. In other words, only those media with meaningful and consistent activity were chosen, ensuring that the dataset is both representative and useful for downstream analysis.

A major focus of the dataset preparation process was **text cleaning**. Telegram data are often messy, containing emojis, URLs, hashtags, or non-standard spellings. Considerable effort was invested in normalizing and cleaning the text to remove these distractions, so that the analysis would focus on the actual semantic content of the conversations. This step is critical in producing reliable results, since raw Telegram text is rarely suitable for direct computational analysis.

The overall pipeline for dataset construction follows a modular approach:

1. Export the target media from Telegram Desktop as raw `.json` files.
2. Use a modular Python-based workflow to parse and convert the raw files into structured `.csv` tables.
3. Store each media source as an independent CSV file, together with a single metadata file describing the media.

This modular design makes it easy to add or remove media without rewriting the pipeline. The use of CSV format offers additional benefits: it is lightweight, widely supported across data science tools, and human-readable. It can be opened directly in spreadsheet software for quick inspection or imported into analytical frameworks such as Python's `pandas` for more advanced processing. This flexibility ensures that the dataset remains accessible to both technical and non-technical users.

3.1 Data Preprocessing and Features

After export, each media source is processed automatically, with every message represented as a single row in the resulting CSV file. The following features were extracted and carefully designed to support meaningful analysis:

- **Message Key**: A unique identifier assigned to each message. This prevents ambiguity when handling large datasets, ensuring that every message can be distinctly referenced.
- **Message Text**: The raw textual content of the message. This forms the primary material for language model analysis and is the main source of semantic meaning in the dataset.
- **Message Clean Text**: A cleaned version of the message text with links, emojis, and other noise removed. This allows the analysis to focus on the linguistic content rather than formatting artifacts, improving both accuracy and interpretability.
- **Time**: The exact time a message was sent. Recording this feature makes it possible to study sequences of communication, detect bursts of activity, and link topics to specific times of day.
- **Date**: The date of the message. This complements the time feature and is essential for identifying long-term trends, correlating discussions with external events, and studying how topics evolve over days or weeks.
- **Reactions**: User reactions to the message (such as likes or dislikes). This provides valuable insight into how content resonates with the audience, offering a more direct measure of approval or disapproval than textual sentiment alone.
- **Hashtags**: Extracted hashtags from the message. These act as self-reported tags, giving a quick indication of the message's subject matter and enabling the detection of thematic clusters or trending topics.
- **User ID**: A unique identifier for the sender of the message (when available). This feature enables the study of individual user behaviors, participation patterns, and influence within groups or channels.

- **Sender Name** (for groups): The display name of the message sender. Similar to user ID, this helps in distinguishing participants, and in some cases may be leveraged for gender inference or sociolinguistic analysis.
- **Replies To** (for groups): The identifier of the message being replied to. This feature is essential for reconstructing conversational threads and analyzing dialog structures, allowing us to move beyond isolated messages and study interactive exchanges.

The final dataset thus consists of **30 CSV files**, one for each media source, together with a consolidated metadata file. Beyond its immediate use in this project, the dataset design is flexible, reproducible, and easily extendable, making it suitable for a wide range of future research and applied use cases.

4 Baseline

To contextualize the performance of the TeleLLMgram system, it is important to establish a baseline for comparison. The baseline represents a simple, straightforward method for analyzing Persian Telegram content without the advanced RAG or LLM-based techniques. By comparing the results of our pipelines against this baseline, we can better assess the value added by retrieval-augmented generation and large language models.

4.1 Baseline Design

For the baseline, the system relies on two primary components:

1. **Keyword Frequency Matching:** In this approach, the system identifies occurrences of query terms or keywords directly within messages. The top messages are selected purely based on keyword counts, without any semantic understanding or context. While computationally efficient, this method cannot capture nuances, synonyms, or implicit references, leading to low accuracy in real-world queries.
2. **Simple Statistical Summaries:** For pipelines that do not require retrieval, the baseline provides basic statistics such as message counts, most frequent words, or reaction distributions. These outputs are informative for general trends but do not allow for complex reasoning, contextual summarization, or multi-message integration.

4.2 Limitations of the Baseline

The baseline approach has several key weaknesses:

- **Poor Semantic Understanding:** Keyword matching fails to recognize context, synonyms, or implicit relationships between messages. For example, messages expressing sarcasm or discussing a topic indirectly are often missed.
- **Inability to Aggregate Long Contexts:** Complex queries that span multiple messages or threads cannot be answered accurately, as the baseline considers each message independently.
- **Low Accuracy for Complex Queries:** Questions requiring reasoning, summarization, or interpretation of user intent cannot be handled effectively, resulting in outputs that are often irrelevant or incomplete.
- **No Adaptation to User Intent:** The baseline cannot adjust its retrieval strategy based on the type of query, unlike RAG pipelines which use keyword scoring and relevance measures to select context-aware sentences.

4.3 Why Advanced Pipelines Improve Upon the Baseline

Compared to the baseline, the TeleLLMgram system demonstrates clear advantages:

- **Contextual Retrieval:** Our RAG-based pipelines go beyond simple keyword counting by scoring sentences according to relevance with respect to key concepts in the input query. This enables more accurate selection of messages for downstream LLM processing.
- **Multi-Message Understanding:** By chunking longer inputs and feeding them into the LLM, the system can interpret complex conversations across multiple messages, providing coherent and context-aware answers.
- **Semantic and Cultural Nuance:** LLMs can capture subtleties such as sarcasm, slang,

code-switching between Persian and English, and culturally specific references, which are completely missed by a simple baseline.

- **Flexible Pipelines:** Unlike the baseline, which only provides generic outputs, TeleLLMgram can generate specialized analyses for topics, trends, individuals, and time-based patterns, making it much more useful for decision-making and research.

In summary, the baseline provides a minimal, keyword-based reference point for Telegram content analysis. While it offers efficiency and simplicity, it lacks the semantic understanding, contextual awareness, and flexibility necessary for meaningful insights. The RAG and LLM pipelines in TeleLLMgram significantly surpass this baseline, demonstrating the effectiveness of combining retrieval-augmented generation with large language models for analyzing Persian Telegram communities.

5 Implementation

The implementation of the TeleLLMgram project centers around two key components: the dataset and the analysis pipelines. The dataset provides the raw material for the study, while the pipelines transform this material into actionable insights. This section outlines how the dataset was collected and prepared, and how the pipelines were designed to cover a wide range of analytical needs.

5.1 Dataset

5.1.1 Dataset Collection

The dataset was constructed using the Telegram Desktop export feature, which allows media sources (channels and groups) to be exported as structured `.json` files. The collection process was iterative: the dataset was updated five times over the course of the project. Each update required approximately two full days of work, and exporting each media source took an average of ten minutes, depending on its size and activity level.

The primary motivation for these repeated updates was to ensure that the dataset captured the evolving dynamics of Persian-language Telegram discussions, particularly those surrounding significant social and political events. A notable example is the war between Iran and Israel, which generated a surge of relevant content that was incorporated into the later versions of the dataset.

The final dataset includes 30 media sources, both groups and channels, chosen for their active participation and thematic relevance. Table 1 provides an overview of the selected media.

Media Name	Type
Israel Radio group	Group
War News	Channel
Farsi VOA	Channel
Political group discussion	Group
Fars News group	Group
Free Union of Iranian Workers	Channel
Important News group	Group
CheKhabar?	Channel
Free political group discussion	Group
Mizan	Channel
Voice of Israel	Channel
Persian Tweet : group	Group
DW Persian	Channel
Islamic world resistance news	Channel
Israel In Persian	Channel
Shargh News	Group
Latest News discussion	Group
BBC Persian	Channel
Important news group	Group
Mamlekate?	Channel
Persian Alarabie	Channel
Official news channel	Channel
Manoto TV	Channel
Saberin News	Channel
Free discussion	Group
Azadi	Channel
Iran International	Channel
Iran-Israel News	Channel
Iran workers media	Channel
Persian Tweeter	Channel

Table 1: Overview of the media sources included in the final dataset.

5.1.2 Dataset Preprocessing and Preparation

Each exported `.json` file was parsed and transformed into a structured `.csv` file. This conversion was achieved using a modular codebase that can flexibly adapt to new media with minimal changes. For every media source, one CSV file was produced, where each row corresponds to a single message. A separate metadata file was also generated, containing the ID, name, type, and file

path for each media source, allowing the dataset to be easily indexed and managed.

The preprocessing step was more than a simple format conversion: it involved advanced text cleaning to remove noise such as emojis, links, and irrelevant formatting, while preserving the semantic content of the conversations. This cleaning step was essential, as Telegram messages often contain noisy or mixed-content text that would otherwise compromise analysis quality.

The final dataset consists of 30 CSV files (one per media source) and one metadata file. This structure provides three key benefits:

- **Modularity:** New sources can be added or removed without disrupting the overall structure.
- **Compatibility:** CSV files are lightweight and supported by nearly all data science tools.
- **Transparency:** Each message can be inspected directly, making the dataset interpretable for both technical and non-technical users.

5.2 Pipelines

At the core of the TeleLLMgram project are its pipelines. A pipeline is defined as a structured sequence of steps that perform a specific type of analysis on the dataset. By abstracting different analytical needs into pipelines, the system can cover a wide variety of use cases while maintaining simplicity and modularity. This design allows a user to select an appropriate pipeline depending on their research question, and the system will automatically execute the required steps to deliver the result.

One of the main advantages of this design is that a relatively small number of pipelines can cover a wide range of analyses. Rather than building a separate tool for each specific query, the project defines general-purpose pipelines that can be flexibly applied to different contexts.

5.2.1 Specific Media Analysis

This pipeline focuses on analyzing the content of a single media source (channel or group) within a specified time period. It enables the study of how discussions evolve in that community, what topics dominate at different times, and how users engage with the media. Such analysis can be useful for media monitoring, brand reputation tracking, or understanding the narrative strategies of political groups.

5.2.2 Topic-Oriented Analysis

This pipeline extracts and analyzes conversations related to a specific topic across multiple media sources. It provides insight into how different groups discuss the same issue, what perspectives emerge, and how opinions are shaped in the online community. This is particularly valuable for social scientists, journalists, and policymakers who want to monitor public opinion on specific events or policies.

5.2.3 Time-Based Analysis

This pipeline examines user behavior over a given period of time, focusing on how discussions and reactions change in response to events. It is useful for studying the dynamics of public attention, identifying when interest in a topic peaks, and linking online discussions to offline events.

5.2.4 Trend Detection

Building on the time-based analysis pipeline, the trend-detection pipeline identifies emerging patterns or recurring themes in Telegram conversations. This can help detect early signals of social movements, viral content, or crises. It is particularly valuable for newsrooms, government agencies, and businesses that need to anticipate shifts in public discourse.

5.2.5 Individual Person Analysis

This pipeline analyzes the contributions of a specific user, based on the messages they have sent. It enables profiling of individual behavior, communication style, or influence within a group. In some cases, it may also support demographic inference (e.g., gender or role detection). Such analysis can be applied to influencer identification, community management, or security investigations.

5.2.6 Statistical Information

This pipeline generates basic statistical summaries for a given media source. It includes metrics such as the number of messages per day, distribution of reactions, or volume of replies. While simple, these statistics are highly informative for gaining a quick overview of media activity and for comparing different sources.

5.2.7 Example Pipeline Usage

Below are some example of various requested analysis their pipeline .

- **Pipeline:** Specific Media Analysis
  Example Request: *Please analyse Manoto TV during the war*
  Requires Retrieved Information: Yes
- **Pipeline:** Topic-Oriented Analysis
  Example Request: *What is the Iranian people opinion about war and its results*
  Requires Retrieved Information: Yes
- **Pipeline:** Time-Based Analysis
  Example Request: *During the 12-days war, what was the behaviour of people in telegram*
  Requires Retrieved Information: Yes
- **Pipeline:** Trend Detection
  Example Request: *What is the hottest/trend in Iran community last month*
  Requires Retrieved Information: Yes
- **Pipeline:** Individual Person Analysis
  Example Request: *Please analyse the Person with userid=acx123 in group=g123*
  Requires Retrieved Information: Yes
- **Pipeline:** Statistical Information
  Example Request: *No prompt*
  Requires Retrieved Information: No

Through this design, the TeleLLMgram system is able to provide a wide range of analyses while maintaining an elegant and compact architecture. By carefully defining the scope of each pipeline, the project achieves both flexibility and clarity, ensuring that complex research questions can be addressed without unnecessary complication.

5.3 Retrieval-Augmented Generation (RAG)

A central component of the TeleLLMgram analysis pipeline is the retrieval-augmented generation (RAG) mechanism, which allows the system to incorporate relevant contextual information from the dataset into the responses generated by the language model. Initially, we experimented with a semantic search-based approach, leveraging vector embeddings to retrieve sentences similar to the user query. While this method was conceptually appealing, in practice it proved to be both slow and insufficiently accurate. Many retrieved sentences were irrelevant or only loosely connected to the input prompt, which reduced the overall quality of the analysis.

To address this issue, we then implemented a keyword-based retrieval system using Whoosh. This method improved retrieval speed and achieved better accuracy than semantic search because it explicitly matched query terms to the text. However, the results were still suboptimal. The approach often retrieved sentences containing the keywords but lacking meaningful relevance to the user's intended context, and its performance could degrade on longer or more complex queries.

Finally, we developed a custom RAG logic tailored to the specific characteristics of Persian Telegram data. In this method, the system first identifies the five most important keywords from the input prompt. Then, each sentence in the dataset is scored based on the number of these keywords it contains. The top n sentences, ranked by this score, are returned as the retrieval results. This approach significantly improved accuracy over the previous methods. By directly scoring sentences based on keyword presence, it ensured that the retrieved information was both relevant and contextually aligned with the user's query, while remaining computationally efficient.

Despite these improvements, some limitations remain in our retrieval module. In particular, the current scoring approach does not fully capture semantic relationships or nuanced meanings beyond keyword overlap. We observed that a Multi-Hop RAG logic, which iteratively retrieves and refines context across multiple sentences or documents, could potentially provide superior accuracy and richer context integration. Unfortunately, due to time constraints, we were unable to implement this Multi-Hop strategy in the current version of TeleLLMgram. Nevertheless, the single-hop keyword-based RAG implemented here represents the most effective balance of speed, accuracy, and simplicity among the methods we tested, and forms a robust foundation for future enhancements.

5.4 Using Large Language Models (LLMs)

A key component of the TeleLLMgram project is the use of large language models (LLMs) to perform complex analysis and generate human-interpretable insights from Telegram data. To ensure accurate and reliable performance, we carefully designed task-specific prompts for each

pipeline. In some pipelines, the required input for the model is relatively long, containing multiple messages or a detailed context. In these cases, the input prompt is first broken into smaller, manageable chunks, which are then individually fed into the LLM. After processing each chunk, the outputs are aggregated and the model is prompted once more to generate a coherent final response that integrates all relevant information. This chunking strategy allows the model to handle longer contexts than it could in a single pass, while maintaining accuracy and coherence in the final output.

Initially, we experimented with both local and cloud-based LLMs. Using the local setup via Ollama, hardware limitations forced us to rely on a relatively small model. Unfortunately, the outputs produced by this model were insufficient in quality for our analysis needs, particularly for nuanced or context-heavy queries. Consequently, we decided to rely exclusively on cloud-based LLMs, which provide access to more powerful models capable of handling our analysis tasks effectively.

For the majority of experiments, we used the GPT-4o-mini model via `avalai.ir`, an Iranian service provider offering cloud-based LLM access. The average cost per API call was approximately 1,500 Toman, which was manageable given the improved quality and reliability of the results. This setup enabled us to efficiently perform high-quality text understanding, summarization, and context-aware analysis at scale, tasks that would have been impossible with local models under our hardware constraints.

Overall, the design choices in prompt engineering, chunked input handling, and selective cloud-based LLM usage were critical to the success of the TeleLLMgram pipelines, ensuring both scalability and accuracy in the analysis of Persian Telegram content.

5.5 What I Have Done vs. What I Proposed

The TeleLLMgram project evolved significantly from its original proposal due to practical constraints and new insights gained during development. Below, we summarize the main differences between the initial plan and the final implementation, along with the reasoning behind each change:

- **Dataset Expansion:** Originally, the project was planned with a smaller dataset consisting of only a few media sources. During implementation, the dataset was extended to include 30 media sources, encompassing both channels and groups with diverse content. This expansion was necessary to capture a more representative sample of Persian Telegram discussions, particularly in light of major social and political events, such as the war between Iran and Israel. The larger dataset allowed for richer analysis and more reliable results, even though it required additional time for collection and preprocessing.

- **Pipeline Adjustments:** The initial design included seven distinct analysis pipelines. During development, one pipeline was removed to focus on the most essential and feasible analyses given time and resource constraints. This decision allowed for greater depth and quality in the remaining pipelines while maintaining a modular and flexible architecture.

- **Simplification of RAG Logic:** The original plan envisioned a complex, multi-hop retrieval-augmented generation (RAG) system capable of iteratively retrieving and refining context. Due to time limitations and implementation complexity, the RAG mechanism was simplified to a single-hop keyword-based scoring approach. Despite its simplicity, this method achieved higher accuracy and efficiency than the initial semantic search and Whoosh-based keyword methods, while remaining compatible with the project's pipelines.

- **Local LLM Usage Cancelled:** The initial proposal included the use of a local LLM for analysis to improve privacy and control. However, hardware limitations forced the use of a smaller local model, which produced unsatisfactory results. Consequently, all analysis was performed using a cloud-based LLM (GPT-4o-mini via `avalai.ir`), ensuring high-quality output despite the cost per API call.

- **Graphical Interface Simplification:** A complex graphical interface was initially planned to allow users to interact with the pipelines and visualize results in an intuitive way. Due to limited time and resources, this component was simplified to a basic interface, sufficient for demonstrat-

ing pipeline functionality and conducting experiments, but without the advanced features originally envisioned.

- **Keyword and Phrase Extraction Simplification:** Initially, the plan involved extracting keywords and key phrases from messages and feeding these condensed inputs to the LLM in most pipelines. In practice, this approach proved to be both slow and inaccurate, as it sometimes omitted important context and subtle nuances. To improve both speed and accuracy, the pipeline was modified to feed entire messages directly into the model, ensuring that no critical information was lost and that the analysis remained contextually rich.

Overall, while some aspects of the project were simplified or removed compared to the original proposal, these changes were necessary to ensure that the core objectives—reliable analysis of Persian Telegram content using LLMs and RAG—could be achieved within the available time and resources. Moreover, the final implementation represents a practical and functional system that can serve as a strong foundation for future extensions, including the addition of more pipelines, more complex RAG mechanisms, or a fully-featured graphical interface.

6 Error Analysis

As previously described, we decided to complement automated evaluation with a human-centered assessment in order to obtain a more reliable picture of the system's performance. While automated metrics such as semantic similarity or retrieval precision can provide useful signals, they often fail to capture cultural nuances, user expectations, and subjective satisfaction — all of which are essential when analyzing Persian Telegram discourse. For this reason, a human-based evaluation was designed and carried out.

6.1 Evaluation Methodology

Ten participants were recruited to test the system across all pipeline types. Each participant was asked to provide a variety of prompts covering different scenarios, after which they received outputs generated by the LLM-based pipelines. In order to ensure fairness and balance, participants were not restricted to only judging their own outputs. Instead, they were randomly assigned a mixed set of responses collected from multiple pipelines and different queries. This design avoided personal bias and ensured that all pipelines were assessed by multiple evaluators.

The rating scale ranged from 1 (*very poor*) to 10 (*excellent*). Participants were asked to consider four key aspects of quality during their evaluation:

- **Relevance:** Does the response properly address the user's query?
- **Accuracy:** Are the facts and interpretations supported by the retrieved messages?
- **Coherence:** Is the answer logically consistent and easy to follow?
- **Depth:** Does the response go beyond surface-level statements to provide useful insights?

The results were aggregated into average scores per pipeline To further visualize the outcomes, we plotted both pipeline-level averages and individual rating distributions.

6.2 Evaluation Results

The results clearly demonstrate the effectiveness of using LLMs in this project. Figure 1 presents the mean scores for each pipeline, while Figure 2 shows the full distribution of all human scores across pipelines.

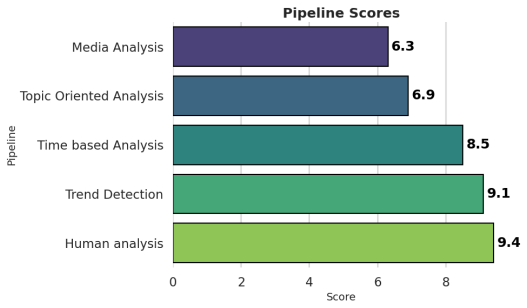

Figure 1: Average human evaluation scores for each pipeline (1 = very poor, 10 = excellent).

6.3 Discussion

Overall, most participants expressed satisfaction with the responses generated by the system. The majority of ratings fell within the 7–9.5 range, indicating that evaluators generally found the outputs relevant, coherent, and insightful. However, performance varied slightly across pipelines.

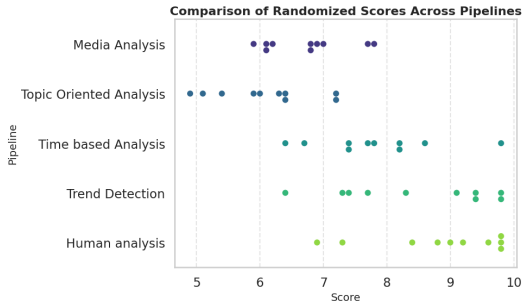

Figure 2: Distribution of individual scores across all pipelines, showing both high overall satisfaction and some pipeline-specific variation.

- **Strong Performance:** Pipelines focusing on *topic-oriented analysis*, *time-based analysis*, and *trend detection* consistently achieved high scores, reflecting the ability of the LLM to summarize evolving discussions and capture user sentiment over time.

- **Weaker Performance:** The *specific media analysis* pipeline obtained the lowest average score. Closer inspection revealed two main issues: (1) users tended to judge the outputs based on their own prior beliefs about the media, which introduced subjectivity, and (2) retrieval errors occasionally led to incomplete or misleading inputs for the LLM. Both factors reduced perceived accuracy.

Despite these limitations, the evaluation highlights the value of LLM-based analysis for Persian Telegram. The combination of high overall scores and positive qualitative feedback confirms that the system is capable of generating culturally-aware insights that satisfy end users. Future work will aim to reduce retrieval errors and improve the fairness of evaluations for pipelines that deal with highly sensitive or polarizing media.

7 Conclusion and Future Directions

7.1 Conclusion

The TeleLLMgram project set out to investigate how large language models (LLMs) and retrieval-augmented generation (RAG) techniques can be applied to analyze Persian Telegram communities in a systematic and scalable manner. Unlike traditional survey-based or keyword-driven approaches, our system demonstrates the capacity of LLMs to provide context-aware, nuanced, and human-like interpretations of large volumes of user-generated content. By building a modular dataset of 30 carefully selected Telegram channels and groups, designing multiple pipelines for distinct analytical tasks, and evaluating system performance through human-centered assessments, this project has established both the feasibility and the value of such an approach.

One of the most important contributions of this work lies in its shift of focus from platforms like Twitter—which have been extensively studied in prior literature—to Telegram, which is the dominant social networking medium for Persian speakers. This choice not only fills a gap in the research landscape but also provides a more accurate representation of how information, opinions, and narratives circulate among Persian communities. The project's implementation of six functional pipelines covering media-specific analysis, topic-oriented analysis, time-based analysis, trend detection, individual-level analysis, and statistical summaries illustrates how diverse research and application needs can be addressed within a unified framework.

Another major finding is that while complex architectures such as multi-hop RAG or advanced keyword extraction were initially envisioned, practical constraints—including time, computational resources, and the shift from a two-person team to a single developer—required simplifying the system. Despite these simplifications, the results demonstrate that even relatively lightweight RAG logics, when combined with strong cloud-based LLMs, can deliver highly satisfying outputs. Human evaluations revealed that most responses scored between 7 and 9.5 out of 10, showing that the system was generally successful in meeting user expectations across dimensions of relevance, accuracy, coherence, and depth. Although performance varied by pipeline—with topic and trend analysis outperforming media-specific analysis—the overall reception was positive, confirming that LLM-based systems can significantly outperform traditional baselines such as keyword matching or statistical summaries.

In summary, TeleLLMgram illustrates the transformative potential of combining RAG with LLMs for social media analysis in low-resource languages and under-studied platforms. It demonstrates that even under constrained resources, a carefully engineered system can provide insights

with real-world applicability to businesses, policymakers, journalists, academics, and civil society actors. Most importantly, it provides an extensible foundation on which more advanced techniques can be layered, pointing the way toward a new generation of culturally aware, real-time, and multi-purpose social media analysis systems.

7.2 Future Directions

While the present version of TeleLLMgram represents a solid and functional proof-of-concept, there are several clear directions in which the project can be extended and enhanced in future work. These opportunities span methodological, technical, and application-oriented domains:

- **Advanced Retrieval Mechanisms:** A key limitation of the current system is its reliance on a single-hop keyword-based retrieval approach. Future work should explore more sophisticated retrieval strategies, including multi-hop RAG, semantic vector-based retrieval using transformer embeddings, and hybrid systems that combine keyword scoring with dense semantic matching. Such improvements could reduce retrieval errors, capture subtler relationships between messages, and increase overall accuracy, particularly for complex or multi-faceted queries.

- **Integration of Multimodal Data:** Many Telegram conversations include not only text but also images, voice messages, videos, and stickers. Extending TeleLLMgram to incorporate multimodal analysis would allow for richer and more realistic interpretations of community behavior. For instance, memes and images could be classified for cultural significance, or voice messages could be transcribed and analyzed alongside text to reveal additional layers of meaning. This would make the system more comprehensive and aligned with the actual communication practices of Telegram users.

- **Enhanced Graphical Interface and Visualization:** The current interface was intentionally kept simple due to time constraints. In the future, developing a more interactive and user-friendly dashboard could greatly expand accessibility for non-technical users. Features such as dynamic filtering, real-time trend visualization, sentiment heatmaps, and network graphs of user interactions would make the insights not only more understandable but also actionable for journalists, businesses, and policymakers.

- **Scaling to Larger Datasets:** While the dataset currently includes 30 media sources, future versions should incorporate hundreds or even thousands of Telegram channels and groups. This scaling would require distributed storage, indexing, and retrieval systems, but it would enable far more representative analysis of Persian online discourse. It would also open the possibility of longitudinal studies that track narratives and sentiments over months or years, rather than weeks.

- **Integration with Real-Time Monitoring:** At present, the dataset is static and manually updated. Future work could implement continuous or near-real-time data collection pipelines that update the dataset automatically as new Telegram messages are posted. Real-time analysis would significantly enhance the system's value for applications such as crisis monitoring, misinformation detection, or early-warning systems for protests or policy changes.

- **Improved Evaluation Methods:** While human-based evaluation provided valuable insights into user satisfaction, future work should combine this with more formal automated metrics, such as BLEU, ROUGE, or BERTScore, adapted for Persian. Additionally, larger-scale user studies involving experts (e.g., journalists, sociologists, or business analysts) could provide deeper insights into the practical utility of the system in real-world decision-making contexts.

- **Incorporation of User and Network Features:** A major strength of Telegram is its networked nature, where the spread of information depends heavily on user interactions and group structures. Future enhancements could focus on modeling these networks explicitly, using graph-based representations to identify influencers, communities, and diffusion patterns. By combining textual analysis with network dynamics, the system could

provide richer insights into how information flows and how opinions are shaped.

- **Domain-Specific Fine-Tuning and Customization:** Although the current system relies on general-purpose LLMs, future work could involve fine-tuning models specifically on Persian Telegram data. This would allow the system to better capture linguistic peculiarities, slang, and culturally specific references. Additionally, developing domain-specific modules—for example, tailored to finance, health, or political discourse—would enable more precise and actionable analyses in those sectors.

- **Ethical, Privacy, and Policy Considerations:** As the system expands in scale and sophistication, it will be important to address the ethical challenges of analyzing social media data. Future research should develop privacy-preserving techniques, such as anonymization or federated learning, to ensure compliance with data protection standards. Additionally, frameworks for responsible use must be established to prevent misuse of the system for surveillance or censorship, ensuring that it remains a tool for empowering research and public understanding rather than suppressing free expression.

- **Broader Cross-Lingual and Cross-Platform Extensions:** While this project focuses on Persian Telegram, the framework can be extended to other languages and platforms. Applying similar techniques to Arabic, Turkish, or Kurdish communities on Telegram, or adapting the approach for WhatsApp and Instagram, would provide valuable comparative insights. Such extensions would also demonstrate the generalizability and robustness of the system beyond its initial focus.

In conclusion, TeleLLMgram represents both a working system and a stepping stone. It shows that meaningful, large-scale, and context-sensitive analysis of Persian Telegram is possible using state-of-the-art NLP methods, even under constraints. The project's simplifications should not be seen as limitations, but rather as pragmatic design choices that enabled the delivery of a functional prototype within a short timeframe. Building upon this foundation, future work can expand the dataset, refine retrieval and generation methods, and integrate advanced visualization and network analysis. Ultimately, this line of research has the potential to transform how Persian digital communities are studied, understood, and engaged with, paving the way for more transparent, data-driven, and culturally sensitive insights into one of the most dynamic social media ecosystems in the world.

References

Al-Rawi, A. (2022). News loopholing: Telegram news as portable alternative media. *Journal of Computational Social Science*, 5(1):949–968.

Asgari-Bidhendi, M., Fakhrian, F., and Minaei-Bidgoli, B. (2020). Parsel 1.0: Unsupervised entity linking in persian social media texts. arXiv preprint arXiv:2004.10816.

Conover, M. D., Ratkiewicz, J., Francisco, M. J., Gonçalves, B., Flammini, A., and Menczer, F. (2011). Political polarization on twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 89–96.

Farahani, M., Gharachorloo, M., Farahani, M., and Manthouri, M. (2020). Parsbert: Transformer-based model for persian language understanding. *arXiv preprint*, abs/2005.12515.

Heydari, M., Khazeni, M., and Soltanshahi, M. A. (2024). Deep learning-based sentiment analysis in persian language. arXiv preprint arXiv:2403.11069.

Izacard, G. and Grave, É. (2020). Leveraging passage retrieval with generative models. arXiv preprint arXiv:2007.01282.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474. See also arXiv:2005.11401.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta.

Rajabi, Z. and Valavi, M. (2021). A survey on sentiment analysis in persian: A comprehensive system perspective covering challenges and advances in resources, and methods. arXiv preprint arXiv:2104.14751.

Ranjbar-Khadivi, M., Akbarpour, S., Feizi-Derakhshi, M.-R., and Anari, B. (2023). Persian topic detection based on human word association and graph embedding. arXiv preprint arXiv:2302.09775.

Sadjadi, S. M., Rajabi, Z., Rabiei, L., and Moin, M.-S. (2024). Farssibert: A novel transformer-based model for semantic similarity measurement of persian social networks informal texts. arXiv preprint arXiv:2407.19173.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gómez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Vaziripour, E., Farahbakhsh, R., O'Neill, M., Wu, J., Seamons, K., and Zappala, D. (2018). A survey of the privacy preferences and practices of iranian users of telegram. In *Proceedings of the Workshop on Usable Security (USEC)*. Workshop on Usable Security (USEC) – PDF available from NDSS/USEC web site.