The First International Conference on
Machine Learning and Knowledge Discovery (MLKD 2024)
Amirkabir University of Technology, December 18-19, 2024

Analyzing Persian Twitter Sentiments on the Arbaeen Walk: A Comparative Study of LDA and BERTopic with the Arbaeen Tweets Dataset

Zeynab Didehkhani*

Abstract

As the world becomes increasingly interconnected, social media platforms like Twitter play a pivotal role in shaping public discourse. One significant topic that has emerged in recent years is the Arbaeen walk, a religious and political movement that has attracted global attention. This paper has two main objectives: first, to construct the "Arbaeen Tweets dataset," curated specifically for sentiment analysis; and second, to analyze Persian tweets related to the Arbaeen walk from 2021-2022. By employing natural language processing (NLP) techniques—namely Latent Dirichlet Allocation (LDA) and BERTopic—we aim to uncover prevalent themes and sentiments, referred to as "topics" in topic modeling. The "Arbaeen Tweets dataset" consists of 2,622 colloquial Persian texts from Twitter, labeled for sentiment analysis. Our findings indicate that while LDA demonstrates slightly superior quantitative performance, BERTopic offers greater coherence in topic interpretation.

Keywords: Text Mining, Topic Modeling, LDA, BERTopic, Dataset construction, Sentiment Analysis, Arbaeen, Foot-pilgrimage, Twitter

1 Introduction

Arbaeen Walk: A Historical perspective
From the literal point of view, Arbaeen means forty. The Arbaeen pilgrimage is an Islamic event, on the 20th of the second lunar month(Safar) commemorating forty days after the martyrdom of Imam Hussain (AS) in Karbala, Iraq, on the 10th of Muharram(the first lunar month) which is called Ashura. The Arbaeen walk has deep historical roots, tracing back to when, the companion of Prophet Muhammad(PBUH& HP) and his disciple, Jabir bin Abdullah Ansari and Atiya Awfi, respectively, left Madinah on foot in 1281 AH and reached Karbala on the morning of the first Arbaeen, and according to what is evident in authentic Shiite sources, on the day Arbaeen Lady Zainab(pbuh), grand daughter of Prophet Muhammad(PBUH& HP) and sister of Imam Hussain(AS), and Imam Sajjad (AS), son of Imam Hussain(AS), along with eighty four people entered Karbala and after talking with Jabir Abdullah Ansari, they visited the holy grave of Imam (pbuh) and the Arbaeen pilgrimage started from there. In the Shia traditions, Arbaeen pilgrimage is mentioned as a sign of faith and being a believer. Since about a thousand years ago, the Shiites fulfilled this duty with great difficulty due to the political conditions of those days. But over the last decade, the pilgrimage has attracted millions of pilgrims from different backgrounds, including not only Shia but also Sunni Muslims, Christians, and followers of other religions from various countries around the world, making it the largest annual gatherings globally.The population of Arbaeen pilgrims in Karbala is comparable to the number of pilgrims at the House of God in Mecca during Hajj.

The subject of foot-pilgrimage has been researched and studied in fields such as tourism and tourism management, as evidenced in papers such as [1, 2, 3, 4, 5].In these papers bulk of the focus is on the Santiago de Camino and/or walking trails associated to religious values other than the Islamic foot-pilgrimage sites [6]. However, there are papers that have worked on and discussed the busiest and most densely populated foot pilgrimage, known as the Arbaeen foot pilgrimage. In this regard, we can mention papers [7, 8, 9, 10]. Specifically, in [6], one can find valuable information about the motivations and experiences, the need for a worldview in foot pilgrimage studies, and the Arbaeen pilgrimage. There are also other references about the Arbaeen pilgrimage [11, 12]. This study will investigate the foot-pilgrims of the Arbaeen pilgrimage, which according to [13] attracts 20 million people annually, making it the world's largest annual gathering in one place.

The COVID-19 pandemic, with its associated quarantine measures and travel restrictions, impacted various forms of travel, including religious pilgrimages. This study's data was collected during the pandemic, a period in which COVID-19 remained a significant public health concern.

Analyzing sentiments related to the Arbaeen Walk, a significant religious event, offers insights into the public perception, emotions, and social dynamics surrounding the event. This can be useful for sociologists, anthropologists, and other researchers studying religious and cultural phenomena.

The Arbaeen event is also a subject of interest in Per-

*Independent Researcher, zkhani5319@yahoo.com

sian social media and, as a result, in natural language processing (NLP) research. Sentiment analysis of social media posts can provide insights into public opinions and emotions related to the event, helping to analyze different positions and points of view. This study contributes to the field of natural language processing (NLP) by focusing on the under-resourced Persian language. We address the unique challenges and opportunities associated with processing and analyzing Persian text, helping to fill a significant gap in existing research. By focusing on sentiment analysis of tweets, our study demonstrates the application of NLP techniques to real-world social media data. This can be useful for various stakeholders, including policymakers, community organizers, and social media companies interested in understanding public opinion and engagement.

Topic modeling is a crucial tool in analyzing text data, as it helps uncover hidden themes within a collection of documents. Numerous methods have been developed in this field, each with its own advantages and limitations. Among these methods, Latent Dirichlet Allocation (LDA) and BERTopic are notable. LDA is one of the most widely used topic modeling techniques, based on probabilistic distributions, and has proven effective in many text analysis projects. On the other hand, BERTopic is a newer approach that leverages deep embeddings and clustering to extract topics. Comparing Latent Dirichlet Allocation (LDA) and BERTopic for topic modeling in Persian text data can highlight the strengths and weaknesses of each method. This comparison can guide future researchers in choosing appropriate techniques for similar tasks and contribute to the ongoing development and improvement of NLP methods.

Research has focused on sentiment analysis in social networks, particularly Twitter, as well as on pilgrimage walks, especially the Arbaeen walk, and the comparison of models used in sentiment analysis. In this research, we analyze the sentiment of Persian tweets on Twitter (a language that has been less studied in this field) regarding the cultural and political event of Arbaeen, which holds significant interest. By constructing the Arbaeen Tweets Dataset focused on Persian text, our study tries to provide a new and valuable resource for researchers interested in social media analysis, sentiment analysis, and event-specific studies, especially in the context of Persian language and culture. We gathered tweets from July 20, 2021, to October 7, 2021. A significant portion of this research paper is devoted to the meticulous process of curating the "Arbaeen Tweets dataset" uniquely tailored for sentiment analysis. This dataset compilation involves the systematic collection and organization of tweets related to the Arbaeen event, aiming to facilitate in-depth sentiment analysis of the gathered textual data. Through the construction of this specialized dataset, we aim to delve into the nuanced sentiments expressed within the Twitter discourse surrounding the Arbaeen phenomenon, offering a comprehensive foundation for sentiment analysis studies within this specific context.

The paper's structure unfolds as follows. Section 2 reviews relevant literature on social media analysis, sentiment analysis, and topic modeling, particularly focusing on studies related to Persian language and the Arbaeen pilgrimage. Section 3 details the construction of the dataset, including data collection, annotation, and taxonomy. Section 4 delves into the preprocessing steps undertaken to prepare the Arbaeen Tweets dataset for analysis. Section 5 explores the methods and experiments conducted for topic modeling, comparing LDA and BERTopic. Section 6 evaluates the performance of the topic models using coherence and perplexity metrics. Section 7 discusses potential applications of the dataset, highlighting its value for researchers studying sentiment analysis, event-specific studies, and cultural phenomena. Section 8 provides information on accessing the dataset, ensuring its availability for further research and analysis. Section 9 discusses the results, focusing on the comparative performance of LDA and BERTopic. Finally, Section 10 concludes the paper by summarizing the key findings and contributions.

2 Literature Review

Today, social networks serve as platforms for communication, discussion, and the exchange of ideas. They also provide a voice for those who might not otherwise be heard. In this regard, Twitter has become a prominent space for discussions about various topics, including religious and political events. The Arbaeen Walk is an example where users, whether Muslim or not, use hashtags like those shown in Figure 1 to make their voices heard. Analyzing sentiment expressed on platforms like Twitter provides valuable insights into public opinions and emotions. Researchers have increasingly turned to Natural Language Processing (NLP) techniques to analyze the content of tweets, uncovering patterns, sentiments, and themes [14, 15, 16, 17, 18].

English hashtag	Persian hashtag	Arabic hashtag
#Arbaeen	#اربعین	#حب_الحسین_یجمعنا (Love for Hussein brings us together)
#Arbaeen2022		
#arbaeen2022		
#ArbaeenWalk		
#WhatIsArbaeen		

Figure 1: List of English, Persian, Arabic hashtags related to Arbaeen

There are researchers who have done studies on sentiment analysis and topic modeling in Persian language. [19] analyzes Persian/Farsi tweets related to the COVID-19 pandemic in Iran. The authors employed

topic modeling and manual annotation to identify the main themes and content of these tweets. Their findings revealed that the most popular topic was the experience of living under home quarantine. Other prominent topics included news and reports about the pandemic, satire, complaints, and discussions regarding the lifting of US sanctions against Iran. [20] introduces ITRC-Opinion, a new dataset for sentiment analysis of Persian social microblogs. This dataset comprises 60,000 informal and colloquial Persian texts collected from Twitter and Instagram. The authors propose a novel convolutional neural network (CNN)-based architecture designed to improve sentiment analysis on this type of colloquial data. The ITRC-Opinion dataset was used to evaluate this architecture, along with comparative analysis of several other models (LSTM, CNN-RNN, BiLSTM, and BiGRU) and various word embedding techniques (FastText, GloVe, and Word2Vec).

This study examines public sentiment expressed in Persian-language tweets by creating a new dataset focused on tweets containing specific hashtags (Figure 1). Analyzing hashtags offers valuable insights into public opinion regarding events and government policies, informing decision-making related to public wellbeing [21, 22, 23]. While this research focuses solely on textual content, other studies have explored the use of multimodal data (text, images, videos) for similar analyses [24, 25, 26, 27, 28, 29, 30, 31].

The cultural and religious background of the Arbaeen event profoundly influences the emotions expressed by users on Twitter. Arbaeen, which commemorates the martyrdom of Imam Hussein, holds deep spiritual significance for millions of Shia Muslims worldwide. This event is not only a display of religious devotion but also a manifestation of cultural identity and solidarity. The intense feelings of grief, reverence, and unity are often reflected in the tweets of participants and observers. Investigating these sentiments is crucial as it provides insights into how religious and cultural contexts shape public discourse and emotional expression on social media. Understanding these dynamics can enhance our comprehension of the social and psychological impacts of major religious events, contributing to more nuanced interpretations of online behavior and the societal significance of such commemorations. This analysis is essential for researchers and policymakers to appreciate the depth of cultural and religious influences in shaping online narratives and community sentiments.

3 Persian sentiment dataset construction

In the field of sentiment analysis, having access to a comprehensive and dependable data source is crucial. This study focuses on creating a novel dataset for sentiment analysis, specifically tailored to Twitter data related to the Arbaeen event from July 2021 to December 2022. The dataset, is called the "Arbaeen Tweets dataset," is designed to facilitate in-depth sentiment analysis of textual data from this event.

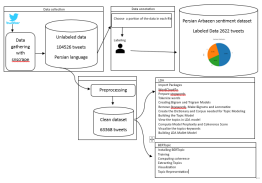

Figure 2: Overview of the process for constructing the dataset for Persian sentiment analysis of Twitter posts with the hashtag in Figure 1

3.1 Data collection

To create our dataset, we used the TwitterSearchScraper class within the snscrape Python module to collect Persian-language tweets containing the hashtags shown in Figure 1. Data collection spanned July 20, 2021, to October 7, 2021, encompassing the Arbaeen event. This targeted approach yielded a substantial dataset suitable for sentiment analysis. A representative sample of this dataset is included in the Appendix; details on the sampling method are provided there. The original dataset, created by combining data from multiple hashtag-specific CSV files, contained 104,526 tweets. Each tweet includes fields for date, ID, content, username, like_ count, and retweet_count.

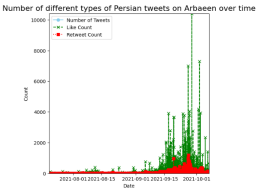

Figure 3: Time series plot depicting the daily counts of Persian tweets, likes, and retweets related to Arbaeen. The plot shows the number of tweets (sky blue line with circles), the total like count (green dashed line with crosses), and the total retweet count (red dotted line with squares) over time. The data highlights trends in social media activity and engagement related to the Arbaeen event

3.2 Annotating

3.2.1 Taxonomy for Titles

In the context of analyzing Persian Twitter sentiments about the Arbaeen Walk, taxonomy plays a crucial role in organizing and categorizing the diverse range of data

and sentiments expressed. Taxonomy, the systematic classification of elements into hierarchical categories, allows for a structured analysis by grouping tweets into predefined categories such as emotional sentiments (positive, negative, neutral), types of content (informational, personal experiences, religious messages), and thematic elements (logistics, solidarity, spiritual reflections). This structured approach facilitates a comprehensive understanding of the data, revealing patterns and insights that might be overlooked in an unorganized dataset. By applying taxonomy, researchers can systematically explore how cultural and religious contexts influence user sentiments, providing a clearer picture of the social and emotional impact of the Arbaeen event on its participants and observers. This method not only enhances the accuracy of sentiment analysis but also supports more nuanced interpretations, contributing to the broader field of cultural and social media studies.

Before any preprocessing, a subset of tweets was selected from each hashtag-specific file. These tweets were then manually annotated by the authors, who categorized them into themes such as “Love”, “Miss”, “Sadness”, “News”, “Corona”, “Question”, “Health advice”, “Surprise”, “Comedy”, “Humiliation”, “Insult”, “Reminiscing”, “Anger”, “Fear”, “Criticism”, “Longing”, “Encouragement”, “Media”, “Satire”, “Humor”, “Diary”, “left behind”, “Waiting for the coming of the Savior”, “Political”, “Cultural”, and “Neutral”. Table 1 provides an overview of the taxonomy of topics extracted from Persian tweets about the Arbaeen Walk. In cases where

Table 1: Taxonomy of Topics of Tweets

Main Category	**Subtopics/Keywords (LDA and BERTopic)**
Positive Emotions	Love, Miss, Encouragement, Longing, Surprise
Negative Emotions	Sadness, Humiliation, Insult, Anger, Fear
Neutral Emotions	Reminiscing, Waiting for the coming of the Savior
Health and Safety	Health advice, Corona
Communication	News, Media, Political, Cultural
Humor and Satire	Comedy, Satire, Humor
Personal Reflections	Diary, Criticism
General Categories	Question, Neutral, Left behind

multiple themes are present in a single tweet, we assigned more than one theme. Consequently, the predominant themes, along with their respective percentages, include corona (21%), reminisce (16.8%), grief-left (11.2%), comedy (10.2%), news (9%), love (8.5%), political (7.3%), insult (5.9%), sorrow (5.2%), and question (4.8%).

3.2.2 Labelling

We manually labeled a subset of the data from each file before merging the labeled data into a single CSV file containing 2,622 tweets with three columns: ’content,’ ’theme,’ and ’label.’ Two native Persian speakers (aged 30-40) with expertise in social media and Persian grammar performed the labeling. Positive tweets encompassed those that endorsed the Arbaeen pilgrimage, conveyed joy, or elicited feelings of nostalgia. Negative tweets included those advocating for harm, generating shame, instilling fear, discussing COVID-19, comprising insults, or employing negative humor. Neutral tweets comprised solely of news updates, inquiries, or responses with a neutral stance. Among the tweets, 54.5 were positive (marked as 1), 27.2 were negative (marked as -1), and 18.2 were neutral (marked as 0).

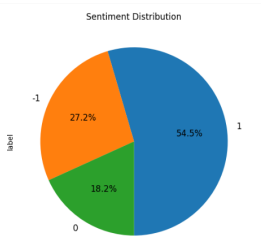

Figure 4: Assigned sentiment labels.

Figure 4 illustrates the breakdown of the dataset’s sentiment labels, including the proportion of positive (1), negative (-1), and neutral (0) sentiments.

As discussed the majority of tweets concerning the Arbaeen event convey positivity, centered around love for Imam Hussain AS, and regret over not participating in the pilgrimage. Following the commemoration of the Arbaeen, which marks the 20th day of Safar, the subsequent days of 28 and 30 Safar are dedicated to the remembrance of the martyrdom of the Prophet Muhammad (PBUH), Imam Hasan (AS), and Imam Reza (AS). Those who were unable to participate in the Arbaeen pilgrimage expressed their devotion through prayer and demonstrated their desire to at least make a visit to Imam Reza (AS), who is located in Iran, specifically in Mashhad. Conversations also touch on concerns related to the ongoing Coronavirus pandemic, revealing a notable divergence among users regarding the trip’s necessity and the perceived seriousness of the COVID-19 situation in Iraq. Users reminisce about past Arbaeen events, expressing deep emotions of longing and sorrow for those unable to attend. Contrasting perspectives emerge, with mentions of a celebrity marriage during the event, showcasing varying tones and themes within the discussions. Criticism surfaces towards fellow citizens for not treating the situation with adequate seriousness, with blame directed at individuals flouting quarantine rules by partaking in the Arbaeen walk or traveling to northern Iran for leisure. The tweets also exhibit a blend of pro- and anti-Iranian regime sentiments.

3.2.3 Evaluating Annotation

For verification and validation of users’s assigned polarity labels, we used the inter annotator agreement, which

calculates the agreement among the annotators. Some of the well-known measures for the calculation of inter-annotator agreement are Fleiss's K, Cohen's Kappa, Cronbach's Alpha, and Krippendorff's Alpha [43]. Cohen's Kappa is a statistical measure used to evaluate the agreement between two raters or observers on categorical items. Unlike simple percentage agreement calculations, Cohen's Kappa accounts for the possibility of agreement occurring by chance. It is defined as

$$k = (p_0 - p_e)/(1 - p_e) \quad (1)$$

where p_0 is the empirical probability of agreement on the label assigned to any sample (the observed agreement ratio), and p_e is the expected agreement when both annotators assign labels randomly. p_0 is estimated using a per-annotator empirical prior over the class labels [44]. The Kappa value ranges from -1 to 1, where 1 indicates perfect agreement, 0 signifies no agreement beyond what would be expected by chance, and negative values suggest disagreement.

Method	Agreement
Annotator's inter-agreement	0.517

Figure 5: Validation results

A cohen's Kappa value of 0.517 indicates that there is a moderate level of agreement between annotators. This is a positive result, indicating that there is a good level of consistency, but also highlighting areas where further refinement and contribution could enhance inter-rater reliability.

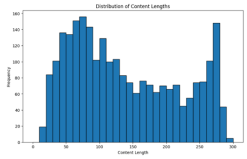

Figure 6: Distribution of comment lengths in the dataset.

Figure 6 illustrates the distribution of comment lengths within the dataset. The horizontal axis represents the number of characters per comment, while the vertical axis indicates the frequency of comments with a specific length. This distribution provides insight into the general structure of the comments, highlighting whether the dataset is skewed towards shorter or longer comments, and it can be used to inform decisions about text processing steps such as truncation or padding during model training.

4 Preprocessing step

Data preprocessing is a crucial step in the data mining and machine learning process, involving the transformation of raw data into a format that can be effectively analyzed by computational systems. This process is essential due to the inherent messiness and inconsistency of real-world data, which often contains errors, missing values, and lacks a uniform structure. To facilitate machine learning and analytics, data must be cleaned, integrated, transformed, reduced, and validated to ensure its quality and suitability for analysis.

Social media language is more similar to colloquial language than to formal writing. Therefore, compared to data from formal writing sources such as Wikipedia, the analysis of social media data presents unique challenges. Challenges arise from the similarity of social media language to colloquial language. In every language, spoken language is considered informal. People use contractions, potentially incorrect grammar, and varied word order. Formal language and informal (spoken) language in Persian exhibit distinct characteristics. Formal Persian, used in written documents, official communications, and academic contexts, adheres to strict grammatical rules, utilizes a more extensive vocabulary, and often includes classical Persian elements. It avoids colloquialisms and slang, maintaining a structured and polished tone. In contrast, informal spoken Persian is more relaxed and fluid, incorporating everyday slang, idioms, and regional dialects. It often features shorter sentences, contractions, and a conversational tone, reflecting the speaker's spontaneity and personal expression. This dichotomy is evident in the different registers used in various social contexts, highlighting the adaptability and richness of the Persian language.

Persian is among languages with complex and challenging preprocessing tasks [32],[33],[34]. Some of these challenges are listed in [35] consist in Imported letters from Arabic, Unicode ambiguity, Different spellings, Different spacing, Different writing prescriptions, Transliterations, New words, Irregular and compound verbs, Ezafe Construction.

In our study, the Persian language on social media exhibits these challenges. For these reasons, as with other data mining tasks, the preprocessing step is inevitable and even more important. Additionally, we encounter various stickers and emojis along with the text appears social medias especially twitter.

Several studies highlight the significant impact of emojis and emoticons on sentiment analysis accuracy. Byungkyu Yoo and Julia Rayz's research demonstrates that incorporating emojis into sentiment models substantially enhances accuracy, especially on platforms like Twitter [36]. Another study leverages millions of emoji occurrences for pretraining models, improving the detection of sentiment, emotion, and sarcasm across various domains [37]. V. Jagadishwari et al. found that while emoticons had a minimal impact on model accuracy, certain machine learning models like Bernoulli

Bayes performed better in their presence [38]. Petra Kralj Novak et al.'s Emoji Sentiment Ranking provides a language-independent tool for sentiment analysis, showing that most emojis are positive and their sentiment polarity increases towards the end of tweets [39]. A case study on Arabic texts reveals that emojis can emphasize, mitigate, or reverse sentiment depending on the context [40]. Additionally, an emoji-embedding model called CEmo-LSTM significantly improved sentiment classification accuracy in Chinese texts, particularly during the COVID-19 pandemic [41]. Lastly, the Multidimensional Lexicon of Emojis (MLE) offers a comprehensive tool to assess the emotional content of emojis, enhancing sentiment analysis by incorporating their nuanced contributions [42]. However, in this research, since we are focusing solely on text we omitted all emojis and stickers from the data, except for annotating the data.

4.1 Data Reduction

In the data reduction process, we specifically focus on eliminating duplicate rows to ensure the integrity and efficiency of our dataset. Duplicate entries can arise from various sources, such as data collection errors, redundant database entries, or multiple submissions of the same data. These duplicates can inflate the dataset size unnecessarily, leading to increased computational resources for processing and potential biases in analysis. By omitting duplicate rows, we aim to streamline the dataset, improve processing speed, and enhance the accuracy of our results. This step is crucial in maintaining the quality and reliability of the data, ensuring that each record is unique and reflective of the true underlying patterns. After eliminating duplicates, the cleaned dataset was reduced from 104526 to 63368 distinct tweets.

4.2 Data cleaning

Preprocessing steps included the removal of emojis, hashtags, irrelevant symbols, URLs, mentions, and non-English letters or words. Unnecessary columns like 'ID', 'date', 'username', 'like_count', and 'retweet_count' were dropped. The remaining data includes opinions in Persian, which are mostly informal and colloquial. By defining different functions, we removed elements in Figure 7 from our texts to prepare them for analysis. We tokenized and cleaned the text data by removing stop words and punctuation, and created our own stop words list.

4.3 wordcloud

WordCloudFa is a Farsi(Persian) word cloud generator designed to visually represent the frequency and significance of words in a given text, specifically for the Persian language, where the size of each word indicates its frequency or importance within the given dataset. It provides a user-friendly interface where users can input Farsi text, and the tool generates a word cloud where the size of each word correlates with its frequency or importance in the text. This visualization helps in identifying prominent themes and keywords quickly. WordCloudFa supports various customization options, including font styles, colors, and layouts, allowing users to tailor the word cloud to their preferences.

Figure 7: Items removed during the data cleaning process.

In our analysis, we generated a word cloud to highlight the most commonly occurring terms in the "Arbaeen Tweets dataset". The word cloud provides an intuitive overview of the predominant themes and topics discussed. Larger words appear more frequently in the dataset, suggesting their relevance or centrality to the content being analyzed. For example, words such as اربعین(Arbaeen),حسین(Hussain) ,and یجمعنا(unites us) are prominently featured, indicating their significant presence in the text. This visualization aids in quickly identifying key terms and patterns that may warrant further investigation. It serves as a complementary tool to quantitative text analysis methods, providing a more accessible and engaging way to understand the data.

Figure 8: Word cloud of Persian text in the dataset. WordcloudFa

Figure 9 shows word clouds visualizing the top 50 words for each of the six topics identified by topic model (Section 5). Word size reflects frequency within each topic. These Persian-language word clouds were generated using a custom implementation, incorporating normalization and stopword removal to improve clarity.

Figure 9: Word clouds representing the top 50 words for each of the 6 topics identified by the topic model. The size of each word reflects its frequency within the topic. The word clouds are generated using a Persian-specific WordCloud implementation with normalization and stopword removal applied.

4.4 Horizontal bar chart

Horizontal bar charts displaying the top 15 terms Figure 10, bigrams Figure 11, and trigrams Figure 12by frequency in the dataset. The terms are reshaped and reordered for correct right-to-left (RTL) display of Persian text. Each plot shows the frequency of the respective terms, with the highest frequencies at the top. The charts provide insights into the most commonly used terms, bigrams, and trigrams in the analyzed Persian text data.

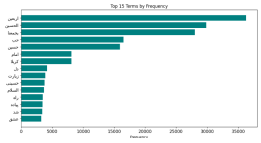

Figure 10: Top 15 terms by frequency in the dataset.

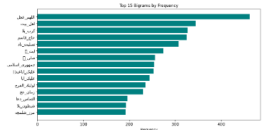

Figure 11: Top Bigrams by frequency in the dataset.

Figure 11 illustrates the top bigrams by frequency in the dataset. Bigrams are pairs of consecutive words, and this figure helps to identify common two-word combinations that may be significant in the context of the dataset. The horizontal axis displays the bigrams, while the vertical axis shows their frequencies. Figure 12 presents the top trigrams by frequency in the dataset. Trigrams, or three-word combinations, can reveal more complex patterns or phrases within the data. Because bigrams and trigrams frequently overlap, we will highlight some of the common word pairs and triplets identified in our analysis. اهل بیت(Ahl al-Bayt): The family of the Prophet (peace be upon them), حاج قاسم(Haj Qassem): A respectful reference to General Qassem Soleimani, جمهوری اسلامی(Jomhuri-ye Eslami): The Islamic Republic, التماس دعا(Eltimas-e Do'a): Request for prayers, مرز مهران(Marz-e Mehran): Mehran border, مرز شلمچه(Marz-e Shalamcheh): Shalamcheh border, امام زمان

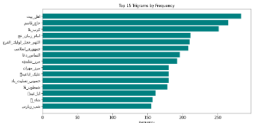

Figure 12: Top trigrams by frequency in the dataset.

ع(Imam Zaman A.J.): Imam Mahdi (May God hasten his reappearance), اللهم عجل لولیک الفرج(Allahumma Ajjil Li-Waliyyika al-Faraj): O Allah, hasten the reappearance of Your guardian (Imam Mahdi), حسینی تسلیت باد(Husseini Tasliat bad): Condolences to Hosseini.

5 Topic Modelling: Methods and Experiments

In today's data-rich environment, professionals in fields like digital marketing and social media (e.g., managers, business owners, politicians) face the challenge of analyzing large volumes of textual data—such as client feedback and social media opinions—on a daily basis. Manually processing this data is time-consuming and inefficient. Artificial intelligence (AI) and machine learning (ML) offer effective solutions. Topic modeling, a technique used to extract underlying semantic structures (topics) from text data, is particularly relevant. Topics are characterized by clusters of dominant keywords, enabling quick comprehension of textual themes. Popular topic modeling algorithms include Latent Dirichlet Allocation (LDA), Probabilistic Latent Semantic Analysis (PLSA), and more recent methods like top2vec and BERTopic. This study employs LDA (using Mallet implementation) and BERTopic to identify prominent topics within a collection of tweets.

5.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a popular algorithm for topic modeling, with excellent implementations available in Python's Gensim package. The challenge, however, lies in extracting high-quality topics that are clear, segregated, and meaningful. Achieving this depends heavily on the quality of text preprocessing and the strategy for finding the optimal number of topics [45].

LDA was initially proposed in the context of population genetics in 2000 and was applied to machine learning in 2003 [46]. It is a probabilistic model used for topic modeling, where documents are assumed to be mixtures of topics, and each topic is a mixture of words. The purpose of LDA is to uncover the hidden and underlying topic structure within a collection of texts. As an unsupervised learning method, LDA does not require labeled data, making it suitable for large text corpora. It assigns probabilities to words belonging to topics and topics belonging to documents. A

key limitation is its "bag-of-words" approach, ignoring word order. While Gensim's LDA employs faster Variational Bayes sampling, we opted for the Mallet (MAchine Learning for LanguagE Toolkit) implementation, utilizing Gibbs sampling for greater precision [47]. This choice, along with a custom stop word list, was tailored to our specific dataset and analysis needs, drawing upon resources such as [47] and [45].

5.2 BERTopic

After the advent of Transformers in 2020, Google introduced a state-of-the-art deep learning model called BERT (Bidirectional Encoder Representations from Transformers). As suggested by its name, this model considers the context from both directions (left-to-right and right-to-left) in all layers. In addition to processing words bidirectionally, it utilizes self-attention mechanisms to weigh the importance of different words in a sentence. The creators of this model pretrained it on a large corpus using two tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). BERT can then be fine-tuned for specific tasks such as sentiment analysis, question answering, and more.

BERTopic is an advanced topic modeling technique that leverages BERT embeddings (Bidirectional Encoder Representations from Transformers) and c-TF-IDF to generate easily interpretable topics from textual data [48, 49]. Unlike traditional methods like LDA, which rely on word frequencies and co-occurrence, BERTopic utilizes BERT's contextual understanding to capture semantic nuances. This enables the creation of more meaningful and contextually relevant topics, particularly in datasets with complex language. BERTopic combines UMAP (Uniform Manifold Approximation and Projection) for dimensionality reduction and HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) for clustering, resulting in robust topic representations well-suited for tasks such as document classification, sentiment analysis, and information retrieval.

5.3 Comparison of BERTopic and Latent Dirichlet Allocation (LDA)

BERTopic and Latent Dirichlet Allocation (LDA) are both popular methods for topic modeling, yet they exhibit distinct strengths and weaknesses.

Latent Dirichlet Allocation (LDA)

Strengths:

- *Simplicity:* LDA is relatively easy to implement and understand, making it accessible for a broad range of users.
- *Scalability:* It is less computationally intensive than BERTopic, thus more scalable for large datasets and suitable for environments with limited computational resources.
- *Interpretable Results:* LDA produces clear distributions of topics over documents and words over topics, facilitating straightforward interpretation.

Weaknesses:

- *Limited Contextual Understanding:* LDA relies on word frequency and co-occurrence patterns without considering the context in which words appear, leading to less coherent and contextually relevant topics.
- *Handling Polysemy and Synonymy:* LDA struggles with words that have multiple meanings and synonyms, as it cannot disambiguate them based on context.
- *Fixed Number of Topics:* It requires the number of topics to be specified in advance, which can be challenging to determine and may not adapt well to the data's inherent topic structure.

BERTopic

Strengths:

- *Contextual Understanding:* By leveraging BERT embeddings, BERTopic captures the semantic nuances and contextual relationships of words, resulting in more coherent and contextually relevant topics.
- *Semantic Richness:* It handles polysemy and synonyms effectively, providing richer topic representations.
- *Advanced Clustering:* BERTopic utilizes UMAP for dimensionality reduction and HDBSCAN for clustering, leading to robust and high-quality topic clusters.
- *Flexibility:* The method is well-suited for complex and diverse datasets, making it applicable to a wide range of tasks such as document classification and sentiment analysis.

Weaknesses:

- *Computational Complexity:* The use of BERT embeddings and advanced clustering techniques requires significant computational resources, which can be a limitation for large datasets or in environments with limited hardware capabilities.

- *Implementation Complexity:* Setting up and tuning BERTopic can be more complex compared to traditional methods, necessitating a deeper understanding of the underlying algorithms and parameters.
- *Pretrained Model Dependence:* The quality of topics generated by BERTopic is influenced by the pretrained BERT model, which may not always be optimal for specific domains or languages.

In summary, BERTopic offers advanced semantic understanding and robust clustering at the cost of increased computational and implementation complexity, while LDA provides a simpler and more scalable solution with limitations in contextual and semantic accuracy. The choice between these methods should be guided by the specific requirements and constraints of the project at hand.

LDA and BERTopic were selected for this study due to their distinct strengths: LDA's well-established probabilistic approach and BERTopic's incorporation of deep learning-based embeddings. These models align well with the research objectives, providing both quantitative and qualitative insights into the Persian text data. While other models exist, their inclusion was beyond the scope of this study.

6 Evaluation

One of the importance of Data Mining steps is Evaluation which allows us to understand how well our designing model is for the task sentiment analysis which we work with data of type text and topic modeling there are different evaluation techniques among them Topic coherence as an internal evaluation technique for topic models and perplexity as a confusion metrics are more famous and practical.

6.1 Coherence

Topic coherence is suggested as an internal evaluation technique for topic models, characterized by the mean or median of pairwise word similarities created by the leading words of a specific topic. Various versions of topic coherence metrics exist, but the most frequently utilized coherence metric in both LDA and BERTopic for determining the ideal number of topics is c_v. This c_v coherence metric relies on a sliding window, one-set segmentation of the top words, and an indirect confirmation method that employs normalized point-wise mutual information (NPMI) and cosine similarity. The c_v coherence measure is calculated using the following formula:

$$c_v = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \log \frac{D(w_i, w_j) + e}{D(w_i) + D(w_j) + e} \quad (2)$$

where n is the number of top words in the topic, $D(w_i, w_j)$is the number of documents that contain both words w_i and w_j , $D(w_i)$ is the number of documents that contain word w_i, and e is a smoothing parameter. The c_v coherence measure ranges between 0 and 1, a higher value indicates that the words in the topic are more coherent, and therefore better [50].

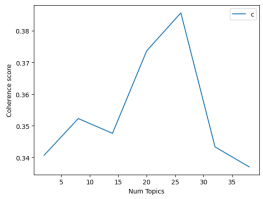

Figure 13: Coherence score for varying number of topics in Arbaeen dataset.

Figure 13 plots the coherence score of the LDA model against the number of topics. This visualization aids in determining the optimal number of topics by identifying the point of maximum coherence, indicating the most interpretable topics.

```
Num Topics = 2  has Coherence Value of 0.3407
Num Topics = 8  has Coherence Value of 0.3523
Num Topics = 14  has Coherence Value of 0.3476
Num Topics = 20  has Coherence Value of 0.3737
Num Topics = 26  has Coherence Value of 0.3857
Num Topics = 32  has Coherence Value of 0.3433
Num Topics = 38  has Coherence Value of 0.337
```

Figure 14: The Coherence results for different numbers of topics.

Analysis of Figure 14 reveals :

- As the number of topics increases, the coherence values fluctuate, indicating varying levels of topic interpretability.
- The highest coherence value, **0.3857**, is observed for **Num Topics = 26**. This suggests that 26 topics produce the most semantically interpretable and cohesive topics.
- **Num Topics = 20** yields the second-highest coherence value of **0.3737**, also indicating reasonably coherent topics.
- Lower coherence values are observed for **Num Topics = 2** (0.3407) and **Num Topics = 38** (0.3370), suggesting that having too few or too many topics leads to less semantically consistent topics.
- **26 Topics** appears to provide the best balance between topic granularity and coherence, offering the most interpretable set of topics.
- **Fewer Topics** (e.g., 2 topics) result in overly broad categories, merging distinct topics into less cohesive ones.

- **More Topics** (e.g., 38 topics) lead to over-segmentation, where meaningful patterns are split apart, resulting in more fragmented topics.

Given the highest coherence value of **0.3857** for **26 Topics**, it is recommended to use 26 topics for the analysis. However, depending on the specific objectives of the study, fewer or more topics may be chosen, with coherence considered as a guiding factor.

6.2 Perplexity

Perplexity is a confusion metric used to evaluate topic models and accounts for the level of "uncertainty" in a model's prediction result. It measures how well a probability distribution or probability model predicts a sample, and is used in topic modeling to measure how well a model predicts previously unseen data. Perplexity is calculated by splitting a dataset into two parts - a training set and a test set. The idea is to train a topic model using the training set and then test the model on a test set that contains previously unseen documents (i.e., held-out documents). The measure traditionally used for topic models is the perplexity of held-out documents, D_{test} which can be defined as:

$$\text{perplexity}(D_{\text{test}}) = \exp\left(-\frac{\sum_{d=1}^{M}\sum_{i=1}^{N_d}\log p(w_{di})}{\sum_{d=1}^{M}N_d}\right) \tag{3}$$

where w_{di} is the i-th word in document d, $p(w_{di})$ is the probability of word w_{di} given the topic model, N_d is the number of words in document d, and M is the number of documents in the test set D_{test}. A lower perplexity score indicates that the model is better at predicting new data [50].

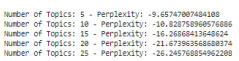

Figure 15: Perplexity scores for LDA models with varying numbers of topics.

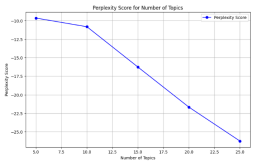

Figure 16: LDA model perplexity as a function of the number of topics.

Figures 15 and 16 illustrate the relationship between LDA model perplexity and the number of topics. Both figures show a general increase in perplexity with an increasing number of topics, indicating that the model's performance decreases as the number of topics rises. The lowest perplexity is observed with 5 topics (Figures 15 and 16)

7 Applications

The constructed dataset provides a valuable resource for several applications. For example, it can be used to train and benchmark sentiment analysis models specifically designed for Persian social media text. This resource allows for the analysis of public sentiment during events like Arbaeen, offering a unique opportunity to study how religious and cultural contexts shape online conversations. Further research could explore the influence of specific hashtags or user demographics on sentiment expression.

8 Access to Dataset

The dataset utilized in this study is publicly available and can be accessed through the link: `https://github.com/Didehkhani/Arbaeen_Dataset`. Users are requested to cite this work in any publications or presentations employing the data. This open-access approach aims to foster transparency, reproducibility, and collaboration within the research community.

9 Discussion and Results

- LDA

  In this study, the optimal settings for the key hyperparameters in Latent Dirichlet Allocation (LDA) were determined through a grid search. This process involved exploring various values for the number of topics (K). The initial search for the best number of topics ranged from 2 to 20, with a step of 1. The alpha parameter was set to 'auto' initially. During this process, only one hyperparameter was adjusted at a time, while the others remained constant until the highest coherence score was achieved. The coherence score, which measures the quality of the extracted topics, was observed to be 0.43 with a perplexity of -9.2 for 3 topics. For LDAMallet, the coherence score was 0.38 and the perplexity was -8.83 for 3 topics. To enhance the interpretability of the topics derived from the LDA model, pyLDAvis was employed to generate an intertopic distance map.

  Figure displays a topic visualization generated using PyLDAvis for the final Latent Dirichlet Allocation (LDA) model with 3 topics (T=3). In this visualization, each circle represents a topic, and the size of the circle corresponds to the prevalence of

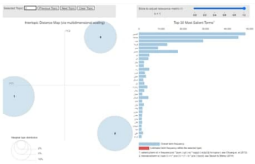

Figure 17: Topic visualization with PyLDAvis for the final LDA model with 3 topics (T=3)

that topic within the dataset. The distance between circles indicates the similarity between topics, with closer circles representing more closely related topics. The right panel displays the most relevant terms for each topic, allowing for a deeper understanding of the specific themes captured by the LDA model.

- BERTopic

  In the process of analyzing tweets related to Arbaeen, a BERTopic model was developed, yielding a coherence score of 0.53. This score reflects the model's capacity to categorize tweets into coherent topics, where a value closer to 1 signifies higher coherence. With a coherence score of 0.53, the model shows proficiency in identifying meaningful themes in the data, although there is potential for enhancement. In the BERTopic model, the default setting generates 731 topics. However, when the number of topics was reduced to 60, the coherence score improved to 0.66. This accomplishment highlights the efficiency of BERTopic in handling extensive amounts of unstructured text data, such as tweets, and deriving valuable insights from them. When working with the Persian language, we include language= "multilingual" in the model.

Table 2: Quantitative Evaluation of Topic Modeling Methods

Metric	**LDA**	**BERTopic**
Coherence Score (c_v)	0.38	0.53
Perplexity	-9.2	-8.8

10 Conclusion

This study introduces a novel dataset of 104,526 Persian tweets related to the 2021 Arbaeen pilgrimage, focusing on #Arbaeen hashtag usage. Through manual annotation (n=2,622) and topic modeling, we identified prominent themes, including 'Love for the Foot Pilgrimage' and 'Corona,' with the latter dominating overall discussion. This resource addresses the scarcity of publicly available Persian-language social media datasets for sentiment analysis, enabling future research into cultural and religious events, cross-linguistic sentiment analysis, and the evolving influence of social media on cultural discourse.

References

[1] C. Blacker The religious traveller in the Edo period. *Modern Asian Studies*, 18(04):593–608, 1984.

[2] R. González and J. Medina Cultural tourism and urban management in northwestern Spain: The pilgrimage to Santiago de Compostela. *Tourism Geographies*, :446–460, 2003.

[3] B. Kim S. S. Kim and B. King The sacred and the profane: Identifying pilgrim traveler value orientations using means-end theory. *Tourism Management*, 56:142–155, 2016.

[4] I. Reader Pilgrimage growth in the modern world: Meanings and implications. *Religion*, 37(3):210–229, 2007.

[5] X. M. Santos Pilgrimage and tourism at Santiago de Compostela. *Tourism Recreation Research*, 27(2):41–50, 2002.

[6] U. Mujtaba Husein phenomenological study of Arbaeen foot pilgrimage in Iraq *Tourism Management Perspective*, 26:9–19, 2017.

[7] A. Nikjoo, et al. What draws Shia Muslims to an insecure pilgrimage? The Iranian journey to Arbaeen *Iraq during the presence of ISIS, Journal of Tourism and Cultural Change*,2020.

[8] A. Nikjoo, et al. From Attachment to a Sacred Figure to Attachment to a Sacred Route: The Foot-Pilgrimage of Arbaeen in Iraq *mdpi Journal, Religions*, 11(45): 2020.

[9] F. Al Ansari, et al, Health Risks,Preventive Behaviours and Respiratory Illnesses at the 2019 Arbaeen: Implications for COVID-19 and Other Pandemics *mdpi, I nt. J. Environ. Res. Public Health*, 18: 3287 2021. https://doi.org/10.3390/ ijerph18063287

[10] F. Al Ansari, et al. Arbaeen health concerns: A pilot cross-sectional survey *Travel Medicine and Infectious Disease*, 2019.

[11] M. Karami Ghahi Comprehension of Iranian Women's Experience of Arbaeen Foot Pilgrimage *Quarterly Journal of Social Sciences, Allameh Tabataba'i University*,27(91): 2021.

[12] https://whoishussain.org/who-is-hussain/the-day-of-arbaeen/

[13] M. Piggot 20 million Shia muslims brave Isis by making pilgrimage to Karbala for Arbaeen *International Business Times. Retrieved from http://www.ibtimes.co.uk/20-million-shia-muslims-brave-isis-by-making-pilgrimage-karbalaarbaeen-* 1476618. 2014.

[14] T. T. Olga Kolchyna, P. Souza Methodology for Twitter Sentiment Analysis *Aste Published 2015 Business, Computer Science*

[15] N.F.F. da Silva, et al. Tweet sentiment analysis with classifier ensembles *Decision Support Systems,* 2014. http:// dx.doi.org/10.1016/j.dss.2014.07.003

[16] N. Chintalapudi, G.Battineni„ F. Amenta Sentimental Analysis of COVID-19 Tweets Using Deep Learning Models *Infect. Dis. Rep.*, 13: 329–339, 2021. https://doi.org/10.3390/idr13020032

[17] M. Surya Asriadie, M. Syahrul Mubarok, Adiwijaya Classifying emotion in Twitter using Bayesian Network *IOP Conf. Series: Journal of Physics: Conf. Series 971* , 2018. 012041

[18] S. Das, A. Dutta, G. Medina, L. Minjares-Kyle, Z. Elgart Extracting patterns from Twitter to promote biking *IATSS Research 43*, 51–59,2019.

[19] P. Hosseini, P. Hosseini, D. Broniatowski Content analysis of Persian/Farsi Tweets during COVID-19 pandemic in Iran using NLP *Association for Computational Linguistics, Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, 2020.

[20] M. Mazoochi, L. Rabiei, F. Rahmani, Z. Rajabi constructing colloquial dataset for persian sentiment analysis of social microblogs *https://arxiv.org/pdf/2306.12679*

[21] G. Kassem, E. Asfoura2, B. Alhuthaifi, J. J. C. Gallego and F. Balhaj Sentiment Analysis and Classifying Hashtags in Social MediaUsing Data Mining Techniques *Inf. Sci. Lett. 12*, 9:2153–2163, 2023.

[22] J. Weston, S. Chopra, K. Adams #TAGSPACE: Semantic Embeddings from Hashtags *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1822–1827

[23] S. Patil, S. Patil An overview of sentiment analysis of hashtags from social media *IJAR 2021* 7(10): 391–394, 2021.

[24] J. Rashid, J. Kim, and U. Naseem Coherent Topic Modeling for Creative Multimodal Data on Social Media *In Preceedings of the ACM Web Conference*, 3923–3927,2023.

[25] J. Bian, Y. Yang, and T-S. Chua Multimedia summarization for trending topics in microblogs. *In Proceedings of the 22nd ACM international Conference on information & knowledge management*,1807–1812, 2013.

[26] T. Chen, H.M.SalahEldeen, X. He, M. Kan, and D.L. Velda Relating an image tweet's text and images *In Twenty-Ninth AAAI Conference on Artifcial Intelligence*,2015.

[27] U. Naseem, J. Kim, M. Khushi, and A. G Dunn A Multimodal Framework for the Identifcation of Vaccine Critical Memes on Twitter *In Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*,706–714, 2023

[28] N. Rasiwasia, J.C. Pereira, E. Coviello, G. Doyle, G.R Lanckriet, R. Levy, and N. Vasconcelos A new approach to cross-modal multimedia retrieval *In Proceedings of the 18th ACM international conference on Multimedia*, 251–260, 2010.

[29] S. Thapa, A. Shah, F. Jafri, U. Naseem, and I. Razzak A multi-modal dataset for hate speech detection on social media: Case-study of russia-ukraine confict. *In Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*,1–6, 2022.

[30] F. Xue, R. Hong, X. He, J. Wang, S. Qian, and C. Xu Knowledge-Based Topic Model for Multi-Modal Social Event Analysis *IEEE Transactions on Multimedia* ,8:2098–2110, 2020

[31] F. Xue, J. Sun, X. Liu, T. Liu, and Q. Lu Social multi-modal event analysis via knowledge-based weighted topic model *Journal of Visual Communication and Image Representation* ,59:1–8, 2019.

[32] M. Shamsfard Challenges and Opportunities in Processing Low Resource Languages: A study on Persian *the Proceedings of the 1st International Conference on Language Technologies for All (LT4All)*, 291–295, 2019.

[33] M. Shamsfard Challenges and Open Problems in Persian Text processing *Unpublished manuscript*,

[34] B. QasemiZadeh, S. Rahimi, and M. Safaee Ghalati Challenges in Persian Electronic Text Analysis *CoRR abs/1404.4740*, 2014.

[35] M. Shamsfard, H. Jafari, M. Ilbeygi STeP1: A Set of Fundamental Tools for Persian Text Processsing In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA)., 2010

[36] B. Yoo and J. Rayz Understanding Emojis for Sentiment Analysis *the 34th International FLAIRS Conference (FLAIRS-34*, 2021.

[37] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, S. Lehmann Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm *Association for Computational Linguistics*, 1615–1625, 2017.

[38] V. Jagadishwari, A. Indulekha, K. Raghu, P. Harshini Sentiment analysis of Social Media Text-Emoticon Post with Machine learning Models Contribution Title *Journal of Physics: Conference Series*, 2070(012079), 2021.

[39] P. Novak, J. Smailovi´c, B. Sluban, and I.Mozeti˘c Sentiment of Emojis. *PLOS ONE*, 2015.

[40] S.A. Hakami, R.Hendley, and P. Smith Emoji Sentiment Roles for Sentiment Analysis:A Case Study in Arabic. Texts *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, 346–355, 2022.

[41] C. Liu, F.Fang, X. Lin, T. Cai, X. Tan, J. Liu, and X. Lu Improving sentiment analysis accuracy with emoji embedding *Journal of Safety Science and Resilience*,2: 246–252, 2021.

[42] R. Godard and S. Holtzman The Multidimensional Lexicon of Emojis: A New Tool to Assess the Emotional Content of Emojis *Human-Media Interaction, a section of the journal Frontiers in Psychology*,2022.

[43] L. Zhang, S. Wang, and B. Liu Deep learning for sentiment analysis: A survey *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4): e1253, 2018.

[44] R. Artstein and M. Poesio Inter-coder agreement for computational linguistics *Computational Linguistics*, 34(4):555–596, 2008.

[45] https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

[46] D.M. Blei, A.Y. Ng, and M.I. Jordan Latent Dirichlet Allocation *Journal of Machine Learning Research*, 3: 993–1022, 2003.

[47] https://towardsdatascience.com/basic-nlp-on-the-texts-of-harry-potter-topic-modeling-with-latent-dirichlet-allocation-f3c00f77b0f5

[48] M. Grootendorst BERTopic: Neural topic modeling with a class-based TF-IDF procedure *arXiv preprint arXiv:2203.05794*, 2022.

[49] https://maartengr.github.io/BERTopic/index.html

[50] H. Axel born, J. Berggren Topic Modeling for Customer Insights A Comparative Analysis of LDA and BERTopic in Categorizing Customer Call *Master's thesis, Umeå University, SE-901 87 Umeå, Sweden. Retrieved from http://umu.diva-portal.org/smash/get/diva2:1763637/FULL-TEXT01.pdf*, 2023.

Appendix Tweet Categories

In this appendix, we present visual examples of tweets from various categories within the Arbaeen Tweets dataset. These images illustrate the diversity and range of sentiments expressed by users on Twitter during the Arbaeen walk. Each category captures distinct themes and sentiments that were identified through our analysis.

- Category 1: Positive Sentiments

  The first set of images showcases tweets that express positive sentiments regarding the Arbaeen walk. These tweets often include messages of unity, spiritual fulfillment, and community support.

  Example Image 1: A tweet expressing gratitude for the shared pilgrimage experience.

- Category 2: Negative Sentiments

  This category includes tweets that convey negative sentiments. Such tweets may reflect concerns, criticisms, or logistical challenges faced during the pilgrimage.

  Example Image 2: A tweet discussing difficulties encountered during the journey.

- Category 3: Neutral Sentiments

  Tweets in this category are neutral, providing factual information or personal observations without strong emotional tones.

  Example Image 3: A tweet describing the route and arrangements for the pilgrimage.

- Category 4: Political Commentary

  These tweets contain political commentary related to the Arbaeen walk. They often highlight geopolitical implications or socio-political discussions influenced by the event.

  Example Image 4: A tweet commenting on the political significance of the Arbaeen pilgrimage.

- Category 5: Religious and Spiritual Reflections

  This category includes tweets that focus on the religious and spiritual aspects of the Arbaeen walk, sharing prayers, reflections, and religious teachings.

  Example Image 5: A tweet sharing a religious quote related to Arbaeen.

Waiting, longing, left, sadness:		**Label**
Those of us who stayed from the Arbaeen walk... I wanted to say that, God willing, we will be in Mashhad at the end of Safar.... 🙁💔 We are heartbroken sir; I know you care about us :)	...ما که از پیاده روی اربعین جاموندیم خواستم بگم ان شاء الله آخر صفر مشهد باشیم که.... 🙁💔 دلشکسته ایم آقا میدونم حواست بهمون هست :)	Positive
Political, insult, corona		
Arbaeen and the terrifying ghost of the sixth peak From a political point of view, the Velayat system has turned the Arbaeen procession and the sending of large crowds to Iraq into a symbol of its presence and influence in Iraq, and in this way it tries to intimidate the Iraqi people and its political rivals in this country. #Rise_to_overthrow #coronavirus	اربعین و شبح هولناک پیک ششم به‌لحاظ سیاسی، نظام ولایت، راهپیمایی اربعین و گسیل جمعیت انبوه به‌عراق را، به‌نمادی از حضور و نفوذ خود در عراق تبدیل کرده و از این طریق تلاش می‌کند مردم عراق و رقبای سیاسی‌اش در این کشور را مرعوب کند #قیام_تا_سرنگونی #کرونا	Negative

News		
A strange violation of a certain tourism company in #Arabeen Offices such as "Moin Gesht Mellat", despite being informed of the cancellation of all the flight permits of the Iraqi Airline, have proceeded to register the mass of Arbaeen pilgrims, which caused the pilgrims to wait 24 to 48 hours at the airport and stop traveling!	تخلف عجیب یک شرکت گردشگری خاص در #اربعین دفاتری مانند «معین گشت ملت» با وجود اطلاع از ابطال تمام مجوزهای پروازی هواپیمایی العراقیه، اقدام به ثبت‌نام انبوه زائران اربعین کرده‌اند که معطلی ۲۴ تا ۴۸ ساعته زائران در فرودگاه و بازماندن از سفر را رقم زدند!	Neutral
Commander Naja: Today, my colleagues in different parts of the country, after serving the Arbaeen Hosseini pilgrims, are creating facilities in the holy city of Mashhad	فرمانده ناجا: امروز همکاران من دراقصی نقاط کشور، پس ازخدمت به زائران اربعین حسینی، در شهرمقدس مشهد باآمادگی کامل وتمهیدات ویژه، تسهیلات ایجاد می‌کنند و باتمام توان برای امنیت که مهمترین نعمت	Positive

with full preparation and special preparations, and they are trying with all their might for security, which is the most important blessing, and we should thank the officers of order and security, whom the people are also praying for	است درتلاش هستند وباید ازخادمان نظم وامنیت که مردم هم دعاگویشان هستند، تشکر داشته باشیم	
Paradise dream exhibition Arbaeen photo and arrangement exhibition visit from 4 to 15 Mehr (the last day of Safar) 15:00 to 21:00 Location: Tehran, Abbas Abad cultural and tourism area, Haqqani highway (west to east), entrance of the National Library, Shahada Bank Central Bank, Garden art	نمایشگاه رؤیای بهشت نمایشگاه عکس و چیدمان اربعین بازدید ۴ تا ۱۵ مهر (آخرین روز ماه صفر) ساعت ۱۵ تا ۲۱ مکان: تهران، منطقه فرهنگی گردشگری عباس آباد، بزرگراه حقانی (غرب به شرق)، ورودی کتابخانه ملی، بلوار شهدای بانک مرکزی، باغ هنر	positive

These visual examples provide a comprehensive view of the range of sentiments and topics covered in the Arbaeen Tweets dataset. They help illustrate the rich and varied discourse surrounding the Arbaeen pilgrimage on social media.

Corona		
Peace be upon you or Abdullah 📉 Karbala without corona deaths a week after the gathering of 17 million Arbaeen this year	السلام علیک یا ابا عبدالله 📉 کربلا بدون فوتی کرونایی یک هفته پس از تجمع ۱۷ میلیونی اربعین امسال	Positive
Both the one who goes to the north helps in the spread of Corona and the one who goes to Arbaeen	هم اونی که میره شمال تو شیوع کرونا کمک میکنه هم اونی که میره اربعین	Negative
Arbaeen and the terrifying ghost of the sixth peak From a political point of view, the Velayat system has turned the Arbaeen procession and the sending of large crowds to Iraq into a symbol of its presence and influence in Iraq, and in this way tries to intimidate the Iraqi people and its political rivals in this country. #Rise_to_overthrow	اربعین و شبح هولناک پیک ششم به‌لحاظ سیاسی، نظام ولایت، راهپیمایی اربعین و گسیل جمعیت انبوه به‌عراق را، به‌نمادی از حضور و نفوذ خود در عراق تبدیل کرده و از این طریق تلاش می‌کند مردم عراق و رقبای سیاسی‌اش در این کشور را مرعوب کند قیام_تا_سرنگونی#	Negative
Satire		

When they say let's go to a naughty Arbaeen, then they know what news we don't know 🤔😅	وقتی میگن یه اربعین شیطون بلا بریم پس 😅🤔 میدونن چه خبری ماها خبر نداریم	Negative
Insult		
🤮🤮🤮🤮🤮 Wild animals and, of course, according to you, unclean pigs do not eat like this A bunch of dirty Shiites, thrown in #Carnaval_Arabeen Shia is responsible for the spread of #Corona 🤮🤮🤮🤮🤮🤮🤮	🤮🤮🤮🤮🤮 حیوانات وحشی و البته بقول خودتون خوک نجس هم اینطوری غذا 🤮🤮🤮🤮🤮نمیخوره یه مشت شیعه‌سان کثیف ، پلشت در #کارناوال_اربعین شیعه عامل گسترش #کرونا 🤮🤮🤮🤮🤮🤮🤮	Negative
Addressing those sheep who in the midst of Corona became a means of earning money for their Ansar and going to Karbala for Arbaeen: instead of mourning 1400 years ago and making puppets, beat yourself up for the year 1400 with so much poverty. and the corruption, cruelty, death and shame of the Islamic Republic in our country	خطاب به اون گوسفندانی که وسط کرونا تبدیل به ابزار ارز آوری ج ا و اعوان انصارشون شدن و برای اربعین به کربلا رفتن: به جای عزای ۱۴۰۰ سال پیش و ادای عروسکهای خیمه شب باری در آوردن،برای سال ۱۴۰۰ بزنید توی سر خودتون با اینهمه فقر و فساد و ظلم و مرگ . و نحسی ج ا در کشورمون	Negative

Reminisce		
I remember in the Arbaeen walk a few years ago, there was a procession of Azerbaijan and Türkiye. How respectable they are and how cool and classy they are	یادمه تو پیاده‌روی اربعین چند سال پیش، یه موکب بود مال آذربایجان و ترکیه. چقد آدمای محترمی بودن و چه پذیرایی باحال و باکلاسی داشتن.	Positive
Today, we went with the cultural activists of our city to see the Friday imam of our city, who had just come from Karbala. What serenity he had when he talked about the memories of Karbala and Ziarat I felt the mood of Arbaeen walking. 😪😪😪	امروز با فعالهای فرهنگی شهرمون رفتیم دیدن امام جمعه شهرمون که تازه از کربلا اومده بود ، چه صفایی داشت وقتی از خاطرات کربلا 🥰🥰 و زیارت می گفت دلم حال و هوای پیاده روی اربعین رو 😪😪😪 .گرفت	Positive

He graduated from the Salafi seminary and officially became a Maulvi, but the attraction of the Prophet's family drew him towards him and he has been a missionary and promoter of the Ahl al-Bayt school for several years. He is not Seyed, but he always wears a green shawl around his neck. Be the host of Balochistan's Arbaeen movement always healthy and healthy	فارغ التحصیل حوزه علمیه سلفیت بوده و رسما مولوی شده بود اما جذبه خاندان پیامبر او را به سمت خود کشاند و چندین سال است مبلغ و مروج مکتب اهل بیت شده. سید نیست اما شال سبز را همیشه به گردن دارد. میزبان حرکت اربعین بلوچستان همیشه سالم و تندرست باشی.	Positive
Critical		
These friends who have been grumbling for so long that you don't go to Arbaeen, you don't go to Hayat, why don't you get vaccinated and yesterday you were stuck in traffic for 7-8 hours in the north, now the weather in the north is good and people are cultured and intelligent?!	این دوستانی که این چند وقته هی غر می‌زدن که آی نرید اربعین، نرید هیات، واکسن چرا نمی‌زنید و دیروز توی ترافیک هفت - هشت ساعته شمال مونده بودن، الان هوای شمال خوبه و مردم با فرهنگ و !باشعور شدن؟	Positive

Where does Radio and Television get this amount of creativity in programming? Arba'in becomes the shrine of Imam Hossein, two chairs, one presenter, one expert or guest 28 Safar, the location of Imam Reza Shrine, two chairs, one presenter, one expert 🤦‍♀️🤦‍♀️	صداو سیما این حجم ازخلاقیت توی برنامه !! سازی رو از کجا میاره آخه اربعین میشه حرم امام حسین دوتا صندلی یه مجری یه کارشناس یا مهمان صفر لوکیشن حرم امام رضا دوباره ۲۸ دوتا صندلی یه مجری یه کارشناس 🤦‍♀️🤦‍♀️	Negative
Nagd Ali, Member of Parliament: The need for pilgrimage is not less than the material needs 🔶 The government should be concerned about the issues of Arbaeen and the spiritual needs of the people.	نقد علی، نماینده مجلس: نیاز به ✅ زیارت از نیازهای مادی کمتر نیست دولت نسبت به مسائل اربعین و نیاز 🔶 .معنوی مردم دغدغه مند شود 🔶 نیاز به زیارت از نیازهای مادی کم تر نیست. جای آن بود که تصمیمات بهتری گرفته می‌شد.	Positive

The need for pilgrimage is not less than material needs. It was the place where better decisions were made.		
Media		
But the time of Radio Arbaini is warm, may God bless their end, God willing, they have made our hearts so much better in this period of 20 and 30 days, may God make your hearts better. #Radio_Arabin	ولی دمه رادیو اربعینی ها گرم خداوند آخر عاقبتشون رو ختم بخیر کنه ان شاء الله،تو این مدت ۲۰ سی روزه خیلی حال دلمونو خوب کردن خدا حال دلتونو خوب کنه #رادیو_اربعین	Positive
Broadcasting of the documentary "Arabin Jihadi" A corner of the Jihadi activities of the Fedayeen of Hussain (peace be upon him). Thanks to the efforts of Martyr Omid Akbari camp Now from the third channel of Sima	"پخش مستند "اربعین جهادی گوشه ای از فعالیت های جهادی هیئت فدائیان حسین علیه السلام به همت قرارگاه شهید امید اکبری هم اکنون از شبکه سوم سیما	Positive

Cultural		
The understanding of people like you about Arbaeen is a question for me. The effects of Arbaeen with Türkiye and the concert are the same?? I am sorry for this kind of insight. Arba'in is the basis of spirituality in individuals, then in the family, and then in the society. a million people don't go. This number is nothing to insure the spirituality of the society. You yourself need the spirituality of the society to go out.	درک امثال شما از اربعین برام سوال. اثرات اربعین با ترکیه و کنسرت یکی؟؟ متاسفم از این نوع بصیرت. اربعین زمینه معنویت در افراد و بعد خانواده وبعد در جامعه است..میلیونی که نرفتن.این تعداد برای بیمه شدن معنویت جامعه که چیزی نیست.خود شما برای بیرون رفتن احتیاج به معنویت جامعه دارید	Positive
Political		
Noble people of #Iraq 15 million of you will participate in the great march of #Arbain.	مردم شریف #عراق میلیون نفر از شما در راهپیمایی ۱۵ .بزرگ #اربعین شرکت میکنه	Positive

Choosing representatives for Imam Hosseini, anti-American and anti-corruption, I don't think it will be a difficult task for you! Throw the killers of Soleimani and Abu Mahdi out of your country. #Iraq_Elections #Majlis_Al-Shura_Elections	انتخاب نمایندگانی امام حسینی، ضد آمریکایی و ضد فساد فکر نمیکنم کار !سختی برای شما باشه قاتلان سلیمانی و ابومهدی را از کشورتان .به بیرون پرت کنید #انتخابات_العراق انتخابات_مجلس_الشوری#	
Like Ashura and Shab al-Qadr, etc., we have witnessed the occurrence of a deadly upsurge, continuously in relation to the Arbaeen ceremony, for which large crowds gather every year.	مانند عاشورا و شب قدر و... شاهد وقوع یک خیز مرگبار بوده‌ایم، پیوسته نسبت به‌برگزاری مراسم اربعین که رژیم هر ساله جمعیتهای انبوه برای آن	Negative
Waiting for Savior		

Waiting for Savior		
#Allahham_Ajl_Lolik_Alfarj 👈 Even after a few days of #Arbain... #Furj_Chairy, which is one of the best deeds of human beings, continues with full intensity under the holy dome of Hazrat Abu Abdullah al-Hussein, peace be upon him... Let's pray for the moon to appear, which is one of the best deeds	اللهم_عجل_لولیك_الفرج# حتی بعد از گذشت چند روز از 👈 ...#اربعین فرج_خواهی که از بهترین اعمال انسان # هاست، تحت قبه مقدسه حضرت اباعبدالله الحسین علیه السلام، با شدت تمام ادامه ...داره بیایید ماهم برای ظهور دعا کنیم، که از بهترین اعمال است 🌺	Positive
God will definitely take advantage of all these great works of Arbaeen pilgrimage The simplest analysis is that "the Arbaeen walk is a training exercise for the armies of Imam Zaman (a.s.)" Professor Panahian #eighth_shelter #Khadem_Al-Rezaeeim	خدا قطعاً از این‌همه آثار باعظمت زیارت اربعین، بهره‌برداری خواهد کرد ساده‌ترین تحلیل این است که «پیاده‌روی اربعین، رزمایشی برای لشکریان امام زمان(ع)» است استاد پناهیان پناه_هشتم# خادم_الرضاییم#	Positive
Imam Hassan, peace be upon him: Our Qaim will rule over the east and	امام حسن علیه السلام : قائم ما چهل سال بر شرق و غرب جهان حکومت خواهد	Positive

west of the world for forty years, blessed is the one who understands him and listens to his words in those days. Yamlek ma bain na al-khafiqain e-arbaein aamaam; It is good for those who understand his days and hear his words. protest	کرد، خوشا به حال کسی که او را در آن روزها درک کند و سخنانش را بشنود؛ یَمْلِکُ مَا بَیْنَ الْخَافِقَیْنِ أَرْبَعِینَ عَاماً؛ فَطُوبَى لِمَنْ أَدْرَکَ أَیَّامَهُ وَ سَمِعَ کَلَامَهُ. الإحتجاج	