Designing and Implementing an Emotion Analytic System (EAS) on Instagram Social Network Data

Seyed Faridoddin Kiaei*, Mohammad Dehghan Rouzi
Faculty of Engineering
University of Tehran
Tehran, Iran
{sfd.kiaei, dehghanr.mohammad}@ut.ac.ir

Saeed Farzi
Faculty of Computer Engineering
K. N. Toosi University of Technology
Tehran, Iran
saeedfarzi@kntu.ac.ir

Received: 2020/04/04 Revised: 2020/05/30 Accepted: 2020/07/04

***Abstract*— Being aware of people's attitudes and emotions about a specific person or an event can have a high impact on the decisions of individuals and organizations. With the rise of social networks, specifically Instagram, many people are sharing their attitudes on this social network. Analyzing the emotions of users of this social network can help managers make organizational decisions and predict essential events such as elections. In this research, the EAS system designed and implemented to extract emotions and visualize them. As a practical example, the Instagram users' feelings about the two main candidates for the 12th Iranian presidential election also examined. The data were Instagram Persian comments collected using a developed crawler. The result shows a more positive feeling about Rouhani in comparison with Raeisi. Also, the lexicon-based analysis of Rouhani revealed a high level of trust emotion, along with anger and disgust. The crawled and preprocessed dataset is publicly available at https://github.com/sfdk74/EAS.**

Keywords—Emotion Analysis; Visualization; Instagram; Election

1. INTRODUCTION

The first social network site created in 1997. These networks allow users to create a public profile page, list of friends, seeing the connection between them, and the list of their friends. [1]

These days, people share different varieties of information about their opinions, decisions, and feelings on social media. Gathering and analyzing these pieces of information has various applications, such as realizing the society's attitudes about an event or a person or forecasting the election.

In [2] has been shown that how social media expresses collective wisdom which, when properly used, can yield an extremely powerful and accurate indicator of future outcomes.

By expanding the use of social networks and the increase in people's desire to share their points of view and opinions, large social network companies such as Facebook, tend to dominate the social media landscape. In the meantime, with the advent of smartphones, Instagram obtained more popularity due to its high accessibility comparing Facebook. [3]

Sentiment analysis, which is called opinion mining, is a field of study that analyzes the opinions, evaluations, attitudes, and feelings of individuals according to entities such as products, services, organizations, individuals, events, issues, and their characteristics. Beliefs are important almost in every human activity. Because they are our most influential behaviors and whenever we need to make a decision, we want to know about the other's opinions. In the real world, businesses and organizations want to find out consumers' viewpoints of their products and services. Besides, consumers want to find out the previous buyers' opinions about a product to make the right choice. In elections, voters want to know the other's opinions about the candidates before they vote. In the past, individuals used to ask their family and friends about their viewpoints. Businesses and organizations used to use a questionnaire when they want to know about the individual's opinion. Obtaining public and consumers' opinion was being massive commerce for business marketing, public relations, and political advertising companies for a long time. With an increase in usage of the social networks, individuals and organizations increasingly use the content on these networks to make decisions. [4]

In this study, the EAS system has been designed and implemented to extract emotions in social network data. This system employs two significant methods for analyzing the sentiments in the Persian contents. The first method is based on emoji, and the second one is based on the lexicon. To check the performance of the designed system, the data of Instagram gathered in the range of the Iran presidential election in 2017. Because of the importance of Instagram, among other social networks, we chose Instagram.

2. RELATED WORK

In [5], a system designed for real-time visualization of Twitter microblogs and their analysis. For offering enhanced semantic insights, a weighted tag network has been designed. In [6], there is a system to find semantic patterns through heterogeneous data and without any social network structure like Instagram to detect events. In [7], not only Instagram users' posts, but also the combination of Instagram and Twitter users' posts employed to improve event detection quality. In [8], the numbers of four comment categories of Trump and Clinton, supporters and opponents have been studied. The type of classification accomplished by hashtags. The results of the study of comments one day before the U.S. presidential election demonstrates that Clinton's supporters were more than Trump. But since 60% of Instagram users were in the ages between 18 and 35, their prediction was different from the real results. This example clearly demonstrates the inability to generalize the results of Instagram to the whole society . Mohamed

in [9] studied the Malaysian politicians' storytelling. In this study, all the posts of politicians, including video, picture, and text content, have been analyzed. The posts are divided into six categories, and the results are compared to each other. Recently, it has been shown the high correlation between different Indonesian party influencers and their presidential candidates on Twitter [10]. Emotion analysis using text processing techniques also applied on social networks [11], [12].

In [13], city event detection using expandable initial event keywords has been studied. In this study, it found that Twitter is much better than Instagram in recognizing city events. In [14], a real-time visualizing system employed with Instagram and Twitter contents.

3. EAS SYSTEM

In this section, data collection, data preprocessing, data visualization, and system architecture has been explained.

Research method

The first stage (Figure 1) steps are choosing Instagram pages for collecting data, crawler implementation, and building a database of lexicon and emoji to store the data.

In the second stage, gathered data is going to preprocess. Emojis divided into three classes of positive, negative, and neutral. Lexicon database at first translated to Persian, then it is divided into ten groups.

In the last stage, information is going to visualize using emoji-based and lexicon-based methods.

Data crawling

In the beginning, a user account created on Instagram and 40 popular political pages are followed. At this point, we tried to have an equal number of each of the two political groups to have a more accurate analysis. The selected pages id with their followers counts are shown in Figure 2. The followers count number measured at the start time of data crawling which is 12 April 2017. Twenty-three pages for the group with fewer followers and seventeen pages for the other one selected. These pages were the most popular in the time of crawling data with the best of our knowledge.

Comments structure includes content, date and time, comments Id, the post Id, and the user Id. There are restrictions on getting the Instagram API. Besides, there are some limitations to use Ajax in the web version in showing all the comments since we employed reverse API. All the comments of the 40 pages were gathered in JSON format from 12 April 2017 to 29 July 2017 (Iran presidential election timeframe). Due to the political nature of the posts and the possibility of their removal, the data gathered several times a day. Then we removed the duplicate ones. We used Java programming language and Jaunt free library for crawling web pages.

Data preprocessing

The data stored in the Microsoft SQL server due to the large volume of the data, dynamic and stable access to the data.

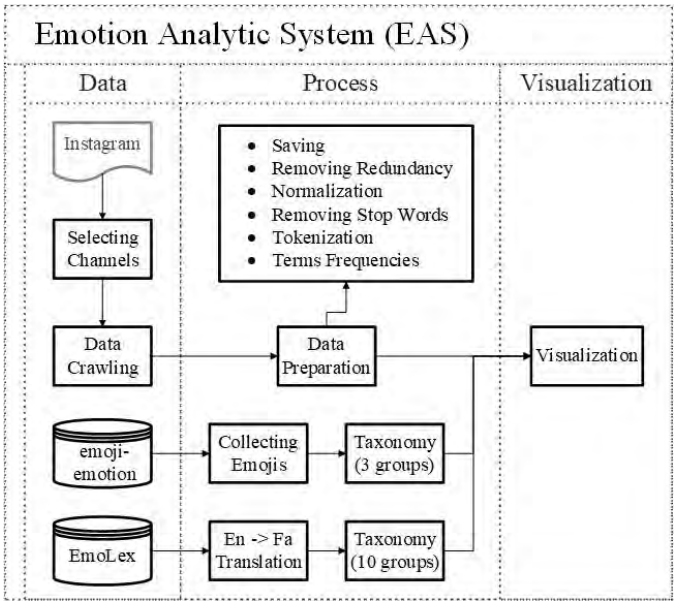


Fig 1**. Research outline process** Three main stages of research are data collection, data processing, and data visualization.

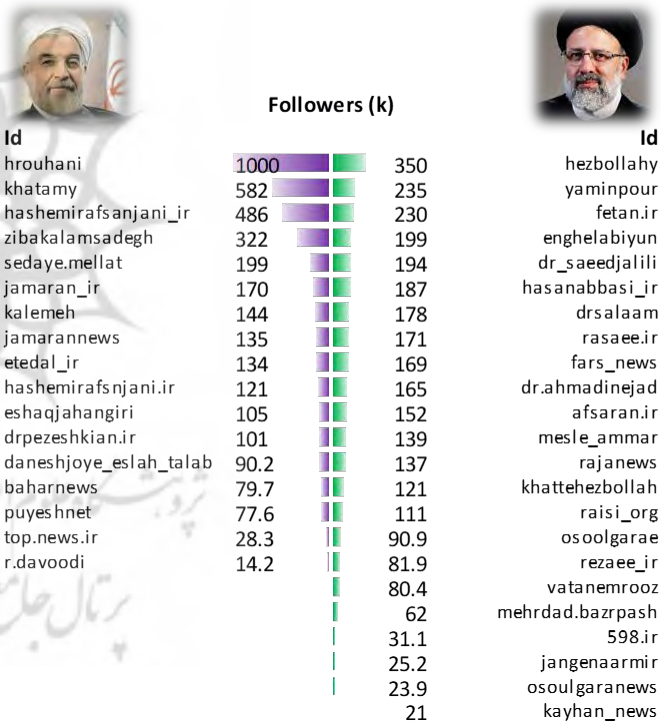


Fig 2 . **Instagram pages used for data crawling.** The pages are categorized into two groups based on their main political orientation. The pages id and followers count also presented here.

The content extracted in Unicode format. However, due to its Persian content, it has been converted to UTF-8 format by considering keeping emojis intact. Some words in the word recognition process could be misidentified for the computer. To solve this problem, we have employed the HAZM library [15] for data normalization. Data normalization includes removing duplicate half-spaces, correction suffixes spaces, converting Arabic or English numbers to Persian equivalent, etc.

Although stop words frequently used in the texts; however, they are not semantically significant. These words gathered [16], and their flaws are corrected and then removed for process results efficiency and speed efficiency. Then content tokenized, and words and emojis separated.

To create a graph to illustrate an abstraction of the data, we need to know the words which mostly repeated with each other. Therefore, the number of repetitions of all the pair words calculated separately and stored to use in the next processes.

Data visualization

Data is vital in analyzing different types of issues. Consequently, choosing the right graph to demonstrate the information in data is essential. Because it provides a powerful tool for a better understanding of data, and if not selected correctly, it can lead to misunderstanding of data.

1) Settings

In terms settings (Figure 3, Figure 4 -A), there is an ability to search at least one and at the most five terms. In the SoDA system [5], there was an ability to search only one term. However, EAS not only demonstrates the results of each term but also it is possible to compare up to five terms. Moreover, in the charts of (Figure 3, Figure 4 -E), there are options to choose between regular and stacked charts in this section. The graph in the system can be used to get familiar with the data and to select additional relevant terms. Furthermore, in time settings, there is an ability to choose the start and the end time (Figure 3, Figure 4-A). Eventually, the visualization process could be started.

2) Modeling and visualizing comments

In the beginning, there is an abstraction of the whole data in the database depicted in (Figure 3 -F) graph. Each node stands for a word, and the size of it shows its repetition number and importance. Each node connects others with edges. The thickness of each edge representing the repetition number of its head nodes in one comment.

In graph calculation, first of all, five words of most frequent words are chosen. Additionally, five words that have a high repetition number with each of the previous words chosen as well. Eventually, the graph depicted with that information. The drawn graph has the most important words of data, and it provides a general overview of the data. This graph is presented in the first part of the program to ease the user in choosing terms.

3) Modeling and visualizing emotions distribution

After term(s) choosing, comments are classified into three classes: positive, negative, and neutral, based on their emojis. The results have shown in the bar plot in (Figure 3, Figure 4-D). The horizontal axis stands for class, and the vertical one designates the percentage of comments for each input term. In this plot, outcomes normalized, and the sum of categories for each term equal to one.

In the next section, comments divided into ten categories concerning their emotional words. In categorization, if the comment contains any of the ten emotional class words, it categorized into the class of that word. Therefore, each comment could be in more than one class. Ultimately, the result represented in the bar plot (Figure 4-D). The horizontal axis indicates the classes, and the vertical axis stands for comment percentage.

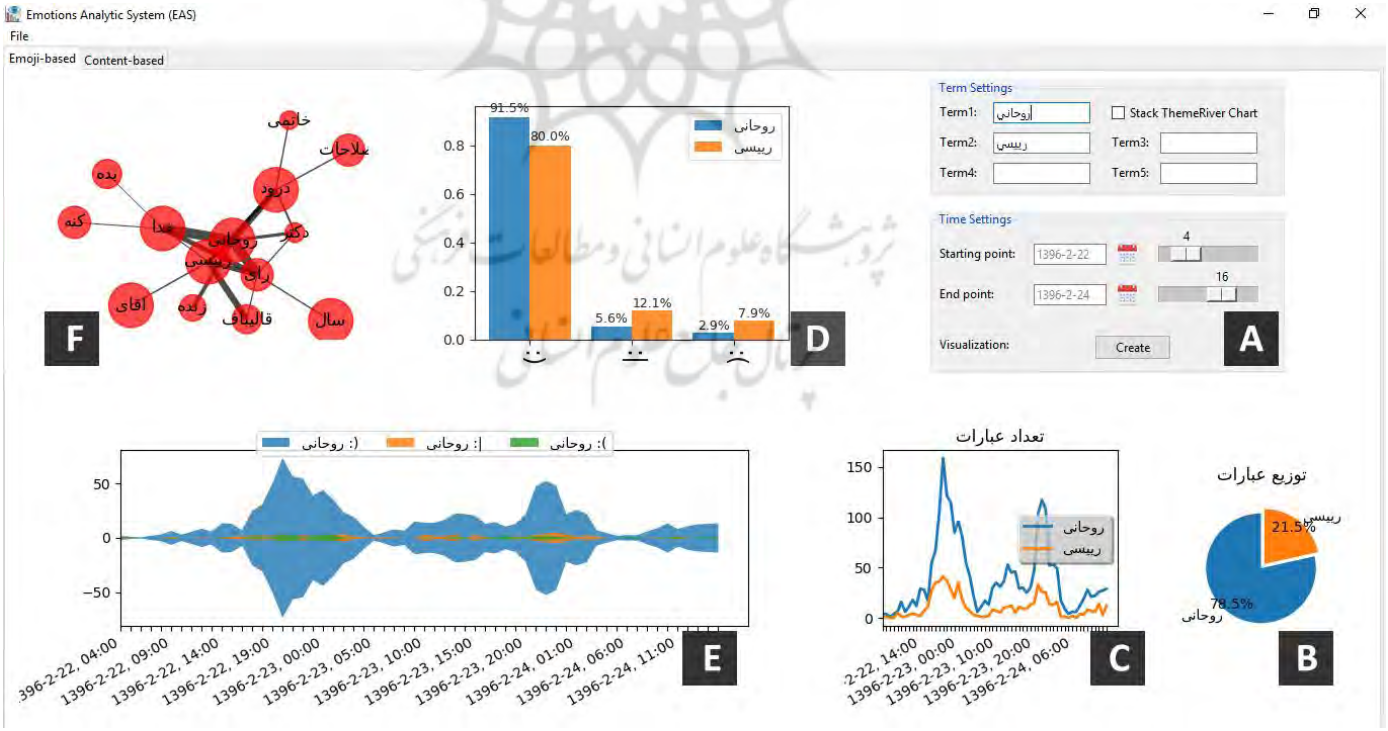


Fig 3 **The emoji-based output of EAS for 'Rouhani' and 'Raeisi' terms.** A few days before Iran presidential election, based on analysis and visualization of 2D (content and time) Instagram comments. Graph of information summary (F) performs essential words of data, including 'Rouhani' and 'Raeisi', two major election rivals. The bar plot (D) demonstrates comments distribution per term in three classes of positive, negative, and neutral. This plot reveals more popularity for 'Rouhani' that has elected. Theme river chart (E), demonstrates the distribution of the first term through time.

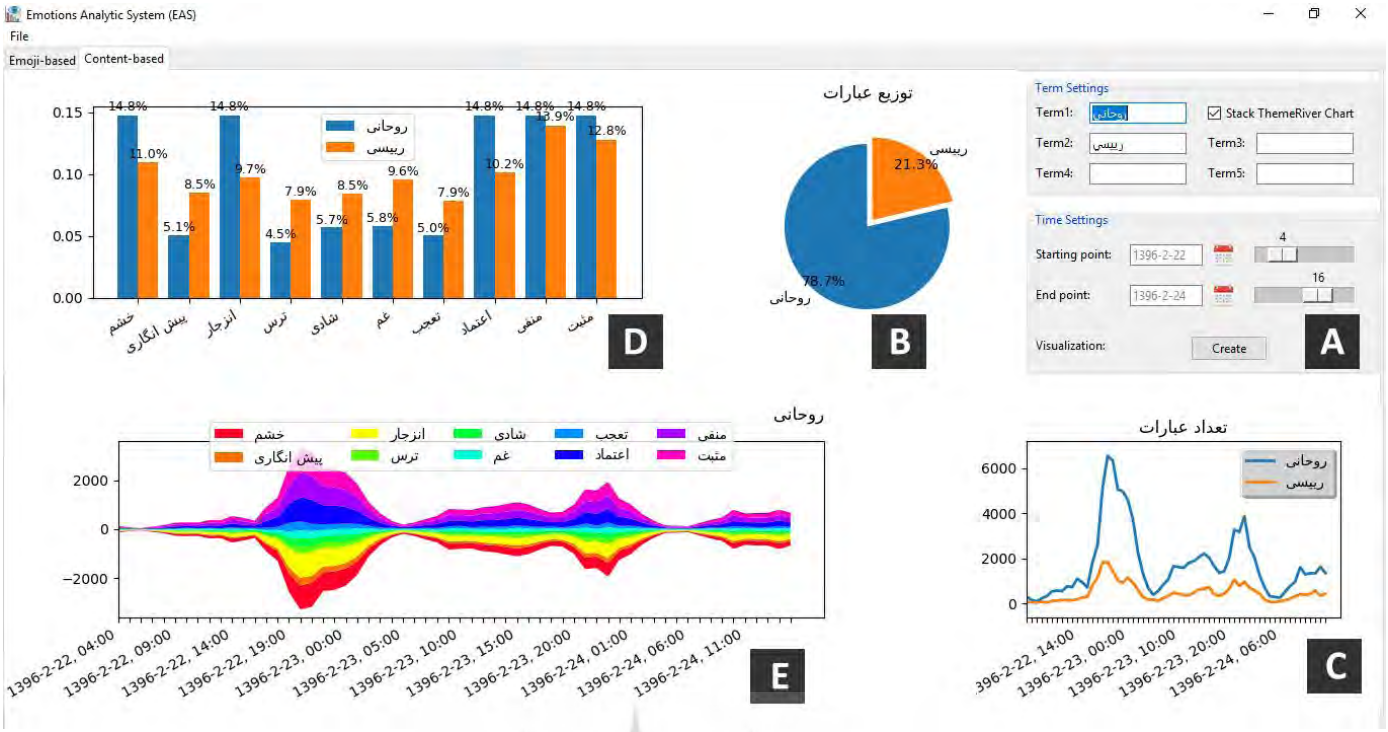


Fig 4 **The lexicon-based output of EAS for 'Rouhani' and 'Raeisi' terms.** A few days before Iran presidential election, based on analysis and visualization of 2D (content and time) Instagram comments. Bar plot (D) depicts comment distribution per each term in ten categories. This plot dispenses higher trustability and less fear for 'Rouhani' that later become a president. The theme river chart (E) depicts the first term results distribution through time.

4) Modeling and visualizing time

To have a closer look at the first term, which is the essential one, the theme river chart employed. Because it's the best plot in displaying data flow through time. In the first section plot (Figure 3-E), the horizontal axis stands for a time, and the vertical axis stands for the number of comments in each class. In this plot, the stack type plot not been used, and it's easy to compare the results of it. In the second section (Figure 3, Figure 4-E), the axes are likewise the previous section. In (Figure 3, Figure 4-E), stacked mode employed, and each color represents a specific class.

5) Visualizing number and distribution of comments

For awareness of the number and distribution of terms among all comments, two plots supported in the system. For the overall comparison of terms, the distribution of comments for each searched term, pie plot depicted in (Figure 3, Figure 4-B). Pie plot is convenient for the investigation of the homogeneous data at a particular time. The line chart in (Figure 3, Figure 4-C) is used. Because it's appropriate for the data study and awareness of the exact number of the comments which given term(s) used in them through time and their change flow.

System architecture

The proposed system has implemented successfully with the Tkinter free library in python. For the user-friendliness of the system, the calendar module employed for date designation (Figure 3, Figure 4-A). Besides, all the plots have zoom in and zoom out options with scrolling. This program, due to its application in Persian pages, its date format is a solar calendar.

Emotion analysis

To recognize emotions in sentences, the emoji-based method, and the lexicon-based method employed.

1) Emoji based

To recognize emotions in a sentence using the emoji-based method, we used a particular set of data [17]. This set of data contains the maximum combination of emojis in them, and also it has polarity between -5 and 5. According to this, the polarity of the sentence divided into three classes of positive, negative, and neutral. Afterward, whole comments that contain emojis, the class of that emoji recognized, and one unit added to its repetition. Eventually, the sum of all the classes for the searched term(s) is demonstrated.

2) Lexicon based

In this method, to recognize emotions, we have employed a lexicon [18] that can divide English words into eight emotional categories and two positive and negative classes. In the dataset, there are ten classes for each word. If the word contains each of them, the value of that class would be one and zero otherwise. The classes are as follows:

- Anger
- Anticipation
- Disgust
- Fear
- Joy
- Sadness

- Surprise
- Trust
- Negative
- Positive

In the beginning, using google translate engine, we have translated them to Persian, and then the duplicate words, produced by translation, are removed. Eventually, numbers and ratios of each class per term, in a particular time, calculated and depicted.

4. RESULTS

Using the proposed system and the collection of Instagram comments, the Persian equivalence of two keywords, 'Rouhani' and 'Raeisi', that were the major rivals in Iran's 12th presidential election, was given to the system. The results of the emoji-based analysis are given in Table 1. This table confirms that the happiness ratio for 'Rouhani' is higher than 'Raeisi', and besides, it has less unhappiness. Negative and neutral emotions for 'Rouhani' among their comments are less than half in comparison with 'Raeisi' among their comments. Positive emotion in all comments for 'Rouhani' is 10% greater than 'Raeisi'. On the other hand, negative emotion about 'Raeisi' is more than double in comparison with 'Rouhani'.

Furthermore, results based on lexicon-based analysis demonstrates more happiness for 'Rouhani' (Table 2). It's noticeable that not only the trust and positive emotions are high for 'Rouhani,' but also it has got higher disgust and anger. On the other hand, negative is the most used emotion in comments about 'Raeisi'. Thus, based on this evidence, it could be concluded that 'Rouhani' will be the president of Iran despite some unhappy people. It means 'Rouhani' supporters are not completely satisfied with him but prefer him to the other one.

Fear about 'Rouhani' is the smallest emotion among other emotions. But it's not that small for 'Raeisi'. Disgust about 'Rouhani' is greater than this emotion about 'Raeisi'. It can be understood that fear is more important for people to reject a candidate than disgust.

The negative sentences about 'Rouhani' are more used in comments than negative emojis.

5. RESEARCH CHALLENGES

The discovery of emotion in contents is possible using emoji or emotional words; however, the error is not low. To achieve better results, using machine learning and NLP methods could be helpful.

Considering the detection of fake users and removing them could lead to getting more reliable results.

Real-time visualization of information on the web couldn't be done on Instagram because of the instability of the Instagram API. In the next stages, this consideration could be done for more reliable results.

Moreover, by getting location data of the users, that they rarely share it on their profiles, location-based information could be added.

Table 1 Emoji-Based Analysis Results

			☻
'Rouhani'	91.5%	5.6%	2.9%
'Raeisi'	80%	12.1%	7.9%

Table 2 Lexicon-Based Analysis Results

	'Rouhani'	**'Raeisi'**
Anger	14.8%	11%
Anticipation	5.1%	8.5%
Disgust	14.8%	9.7%
Fear	4.5%	7.9%
Joy	5.7%	8.5%
Sadness	5.8%	9.6%
Surprise	5%	7.9%
Trust	14.8%	10.2%
Negative	14.8%	13.9%
Positive	14.8%	12.8%

6. CONCLUSION

An emotional analysis system, which acts in both emoji-based and lexicon-based, implemented as named EAS. The graph depicted in this system demonstrates the essential words among the data, including their connections and quantity. The user using this information is able to study one or more term(s) through an arbitrary period of time.

The emoji-based section of EAS has the explained graph and four charts. The pie chart used to present comments distribution. The line chart used to present comments count in the flow of time. The theme river chart used to present the emotion of comments in the flow of time. The bar chart used to present summarized data of the theme river chart. The lexicon-based section has those four charts too. In the emoji-based section, three main emotions that are positive, negative and neutral presented. In the lexicon-based section, ten emotions investigated.

In this study, we have tried to predict the 12th Iranian presidential election result in 2017 using the Instagram comments dataset. Besides, a variety of Instagram users' emotions for two candidates has studied. The results clarify the fact that the positive emotions were higher for 'Rouhani' in comparison with 'Raeisi'. It should be noted that not all comments about 'Rouhani' were positive, and some negative ones exist. But positive emotions about 'Rouhani' are more than 'Raeisi' on average. It's noteworthy that the presented results are based on the Instagram user's data. So, it couldn't be generalized to the whole society.

The collected and preprocessed dataset is available on https://github.com/sfdk74/EAS for researchers.

REFERENCES

[1] [1] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *J. Comput. Commun.*, vol. 13, no. 1, pp. 210–230, Oct. 2007.

[2] [2] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010*, 2010, vol. 1, pp. 492–499.

[3] [3] R. Martinez-Pecino and M. Garcia-Gavilán, "Likes and problematic instagram use: The moderating role of self-esteem," *Cyberpsychology, Behav. Soc. Netw.*, vol. 22, no. 6, pp. 412–416, Jun. 2019.

[4] [4] B. Liu, *Sentiment Analysis: A Fascinating Problem*. *Sentiment Analysis and Opinion Mining*. 2012.

[5] [5] S. Hassan, J. Sänger, and G. Pernul, "SoDA: Dynamic visual analytics of big social data," in *2014 International Conference on Big Data and Smart Computing, BIGCOMP 2014*, 2014, pp. 183–188.

[6] [6] K. Polous, A. Freitag, J. Krisp, L. Meng, and S. Singh, "A general framework for event detection from social media," in *Advances in Geographic Information Science*, 2015, vol. 19, pp. 85–105.

[7] [7] P. Giridhar and T. Abdelzaher, "Visualization of events using Twitter and Instagram," in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2017*, 2017, pp. 82–84.

[8] [8] H. Schmidbauer, A. Rösch, and F. Stieler, "The 2016 US presidential election and media on Instagram: Who was in the lead?," *Comput. Human Behav.*, vol. 81, pp. 148–160, Apr. 2018.

[9] [9] S. Mohamed, "Instagram and political storytelling among Malaysian politicians during the 14th general election," *J. Komun. Malaysian J. Commun.*, vol. 35, no. 3, pp. 353–371, 2019.

[10] [10] O. Monica, F. W. Wahida, and H. Fakhruroja, "The Relations between Influencers in Social Media and the Election Winning Party 2019," in *Proceeding - 2019 International Conference on ICT for Smart Society: Innovation and Transformation Toward Smart Region, ICISS 2019*, 2019.

[11] [11] F. Mozafari and H. Tahayori, "Emotion Detection by Using Similarity Techniques," in *2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems, CFIS 2019*, IEEE, 2019, pp. 1–5.

[12] [12] F. Ghanbari-Adivi and M. Mosleh, "Text emotion detection in social networks using a novel ensemble classifier based on Parzen Tree Estimator (TPE)," *Neural Comput. Appl.*, vol. 31, no. 12, pp. 8971–8983, 2019.

[13] [13] S. Liu and P. Jansson, "City event detection from social media with neural embeddings and topic model visualization," in *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, 2017, vol. 2018-Janua, pp. 4111–4116.

[14] [14] C. Xia, R. Schwartz, K. Xie, A. Krebs, A. Langdon, J. Ting, and M. Naaman, "CityBeat: Real-time social media visualization of hyper-local city data," *WWW 2014 Companion - Proc. 23rd Int. Conf. World Wide Web*, 2014, pp. 167–170.

[15] [15] HAZM, "Python library for digesting Persian text," *Sobhe*, 2014. [Online]. Available: https://github.com/sobhe/hazm. [Accessed: 14-Feb-2020].

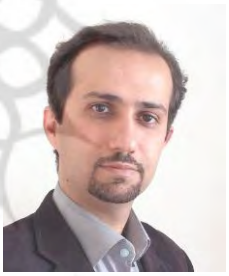
[16] [16] "kharazi/persian-stopwords: Persian (Farsi) Stop Words List." [Online]. Available: https://github.com/kharazi/persian-stopwords. [Accessed: 14-Feb-2020].

[17] [17] T. Wormer, "List of emoji rated for valence in JSON." [Online]. Available: https://github.com/words/emoji-emotion. [Accessed: 14-Feb-2020].

[18] [18] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," in *Computational Intelligence*, , vol. 29, no. 3, pp. 436–465, 2013

Seyed Faridoddin Kiaei received his B.S. in Computer Software Engineering from K. N. Toosi University of Technology, Tehran, Iran. He graduated in 2018 with the first rank. His research interests include Data Fusion, Social Network Analysis, Visualization, and Bioinformatics.

Mohammad Dehghan Rouzi received his B.S. in Petroleum Engineering from the Sharif University of Technology, Tehran, Iran, in 2018. His research interests include Data Fusion, Medical Image Analysis, and Machine Learning.



Saeed Farzi is an assistant professor of Computer Engineering, at the K. N. Toosi University of Technology, where he has served since 2015. He received a Ph.D. in Software Engineering from the University of Tehran in 2015. His current research involves Statistical Natural Language Processing, Intelligent Information Retrieval (IIR), Data-Intensive Computing and Big Data. He is head of TRLab since 2017.