DOI: http://dx.doi.org/10.26483/ijarcs.v14i5.7017

Volume 14, No. 5, September-October 2023

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

EXTRACTING THE NEEDS OF THE LABOR MARKET IN RIYADH THROUGH TWITTER USING TEXT CLASSIFICATION TECHNIQUES

Najlaa Musaad Alsadan*, and Mourad Ykhlef
Information Systems Department College of Computer and Information Sciences
King Saud University Riyadh,
Saudi Arabia

Abstract: Recently, Twitter has attracted a great deal of spread and attention. It is one of the most common social networking sites for sharing ideas, chats, and transfer of information and news through text. The labor market is where the supply and demand for jobs meet, with employees satisfying employer needs for certain services. On the one hand, certain jobs are being eliminated, while others are being replaced by new jobs that were not even possible a few years ago. In this paper, we focused on labor market classification of twitter data belonging to Riyadh city and written in Modern Standard Arabic. We want to classify Arabic jobs' tweets to determine the trending of required job in Riyadh city. Twitter's API was used to collect tweets related to labor market. Five different classifiers were used on the dataset namely; Support Vector Machine (SVM), Multinomial Naive Bayes (M-NB), Decision Tree (DT), Gradient Boosting Classifier (GBC), and Random Forest (RF).to classify the tweets based on their related job classes. We evaluated our work by four different measures which are Precision, Recall, Accuracy and F-measure. We made a comparison between the five classifiers based on those measures. The results show that RF achieved the best Accuracy and F-measure, and it equals 93.62%, 93% respectively. In addition, we found that the trend of labor market needed was administrative jobs "وظائف ادارية.". The percentages of that job class that related to 384 jobs about 19.2%.

Keywords: Text classification, machine learning, trends job, labor market.

I. INTRODUCTION

With the development of Internet technology, social media sites have grown in popularity and played a key role in reconstructing people life [1]. Social media provides people with news and information and make it possible to exchange feelings with many people. In addition, digital media content can be updated quickly and effortlessly [2].

We focused on Twitter data for several reasons. Twitter is now the primary platform for disseminating information in cases such as natural calamities, conflicts, or events that news reporters cannot cover [3].

Additionally, Twitter, a popular tweeting network, is a very influential place where individuals can express their thoughts on a variety of subjects, often articulating strong opinions in support of or against a particular subject. Twitter is a service that allows people to communicate and stay in touch by sending brief, quick messages to one another. Good. Twitter's design is kept incredibly simple, and the site is easy to use on both mobile devices and computers [4].

One of the most interesting features of Twitter is its real-time nature. Such that, at any given instance of time, millions of Twitter users can exchange views on several topics and events that are happening currently in any word country Hence the content posted in Twitter is extremely useful to gather real-time news on a variety of topics.

In other side, big data and its aggregation capabilities have opened new labor market analysis and decision-making [5]. Because today's labor markets are dynamic and complex, gathering data on current and future skill shortages might help better match education with job opportunities. Labor Market Intelligence refers to the design and definition of automated methodologies and tools for supporting real-time labor market monitoring at a very fine-grained level [6].

In this paper, we design a new system for extracting the needs of the labor market in Riyadh city using texting classification techniques. We collect the tweets from Twitter to analyze them for knowing the needs of the labor market. To do this analysis, texting classification techniques were utilized on a dataset of Arabic tweets that are associated with the account and the content. The proposed project has three main phases: data collection, raw data, and classification.

The rest of the paper is organized as follows. Section 2 gives some background information on the Labor Market Intelligence and discusses relevant prior work. Section 3 describes the labor market needs detection system. Results and discussion described in section 4. The study is concluded in Section 5.

II. RELATED WORKS

The aim of Labor Market Intelligence field is becoming increasingly relevant to Labor Market policies design and evaluation. Big data and its aggregation capabilities have opened new labor market analysis and decision-making [5-6].

Also, in existing studies, some scholars used unsupervised machine learning [7] and [8]to classify online recruitment information according to recruitment posts or job descriptions, to excavate labor market information or talent demand in professional fields. Also, reference [7] described an approach to Web Labor Market Intelligence along with three real-life application scenarios, focusing on the realization of a machine learning model for classifying job vacancies.

The rapid growth of twitter for advertising job positions provides a great opportunity for real-time labor market monitoring. As example, reference [9] made an analysis of the most demanded business lines through SVM as a machine

learning using the tweets posted after covid-19 on the social networking site.

In this work, the content of the Twitter texts published by people on different hashtags related to jobs in Riyadh using machine learning. Five different classification techniques were applied in Arabic labor market tweets, which are Support Vector Machine (SVM), Multinomial Naive Bayes (M-NB), Decision Tree (DT), Gradient Boosting Classifier (GBC), and Random Forest (RF).

III. SYSTEM DESCRIPTION

This section presents the methodology used to classify Arabic jobs' tweets to determine the trend of required job in Riyadh city. The key steps in the system have been illustrated in Fig 1. As an overview, we have collected tweets dataset and split it to training and testing and labeling the training dataset was performed. In the second phase, training and testing datasets preprocessing was applied.

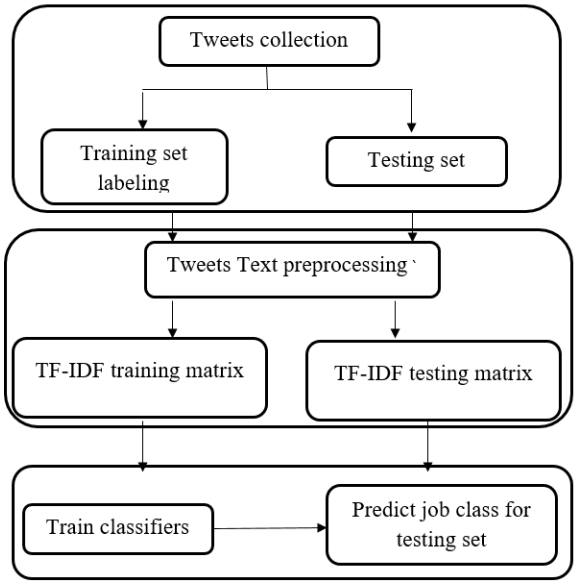

Figure. 1: Main steps in jobs' tweets classification

III.1. TWEETS COLLECTION

We collected the tweets dataset using the Twitter Search API during December 2022 and January 2023. The collected tweets included internal data that refers to unstructured data in the tweet content, i.e., the text of the tweet itself, and external data that refers to the structured data behind tweets such as tweet ID, retweets, in replay to user, tweet language and tweet location. The content of internal data was used to train and test the classifiers, while external data was removed. Only tweets explicitly mentioning 'jobs' or 'jobs in Riyadh' or 'work in Riyadh' ('وظائف , وظائف الرياض,عمل في الرياض') in Arabic were collected.

The number of tweets after removing the duplicated tweets is about 14000 tweets. We randomly chose 4000 tweets for the training process then we perform manual labelling to distinguish jobs' classes. The remainder 2000 tweets for testing process.

III.2. TWEETS TEXT PREPROCESSING

Most social features among Twitter users include hashtags, mentions, responses, and retweets. Before text classification it is necessary to perform preprocessing the tweets because of the informal nature of tweets' text. There are specific steps that the developers take [10-13] So, Preprocessing of tweets' text include the following:

- Removing URL and Hashtags. The researchers replaced them with a space string.
- Removing stop words helps in reducing the size of the dataset and the training time.
- Normalize words and remove punctuation marks, emotions, symbols, and Arabic diacritics.
- Stemming: By dropping the suffix, each word was reduced to its base. That was performed using AraNLP.
- Tokenization: Tokenization is a process that divides lengthy text strings into tokens, which are smaller units of text.
- Remove elongation and use a single occurrence instead.

After that we calculated the TF (Term Frequency) and DF (Document Frequency) to generate the training and testing matrix cells using the TF-IDF (Term Frequency Inverse Document Frequency) weighting method. TF-IDF assigns higher weights to distinguish terms in a document. The more a term occurs in a document, the more it represents the document's content. Also, the more the documents contain the terms, the less informative it becomes [14].

III.3. CLASSIFICATION

To extract the needs of the labor market firstly we performed two steps, Training, and testing. In training model step, the training matrix that contains the selected terms and their corresponding TF-IDF weights in each tweet of the training data is used to train the classification algorithms by learning the characteristics of every class from a training set of tweets. The training process constructs a classification model that will be tested.

In this research, we used the most robust classifiers which are Random Forest Classifier (RF), Support vector machine (SVM), Naive Bayes (NB), Logistic Regression LR, and Gradient Boosting Classifier (GBC).

Random Forest derives from Decision Tree, this means, it shares all the benefits of decision trees [15]. Support Vector Machine (SVM) is a type of supervised ML algorithm that was coined by in 1995 [16]. Naïve Bayes classifier is a proven, simple and effective method for text classification [17] It has been used text classification since the 1950s [18]. Logistics regression (LR) is one of the popular and earlier methods for classification [18], and also Gradient Boosting [19].

Experiments performed included 4000 train tweets against 20 categories of jobs. Table 1 shows accuracy, recall precision, and F1-measure results that were obtained when running the selected classification algorithms. As shown in Table 1, the RF classifier has achieved the highest accuracy with 93.3%. The second highest accuracy are SVM and GBC, which are equal to 90.12%. The accuracy of LR classifier is 88.125. While the

performance of the M-NB classifier scored the lowest accuracy with 68.37%.

Table 1 Accuracy comparison among the selected classification techniques

	M-NB	LR	RF	SVM	GBC
Accuracy	**68.37**	**88.125**	**93.62**	**90.125**	**90.125**
Recall	0.68	0.88	0.93	0.91	0.90
Precision	0.74	0.87	0.93	0.90	0.91
F-measure	0.65	0.87	0.93	0.90	0.90

Fig. 2 illustrates the comparison results for the performance of five text classification techniques applied on our collected dataset.

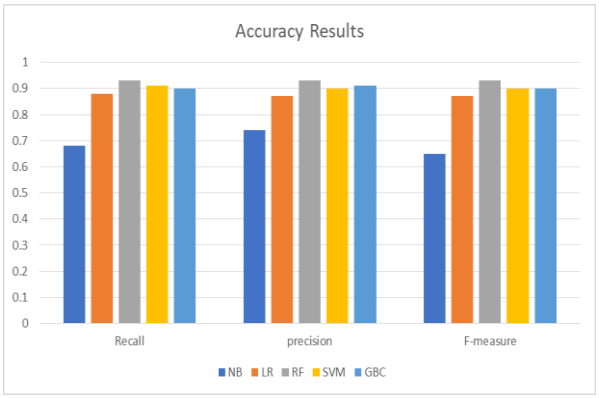

Figure 2. Comparison results

IV. RESULTS AND DISCUSSION

In the testing process the classification model predicts a class for jobs' tweets in the testing set using the predict function. The same terms extracted from the training data and the same weighing methods were used to test the classification model.

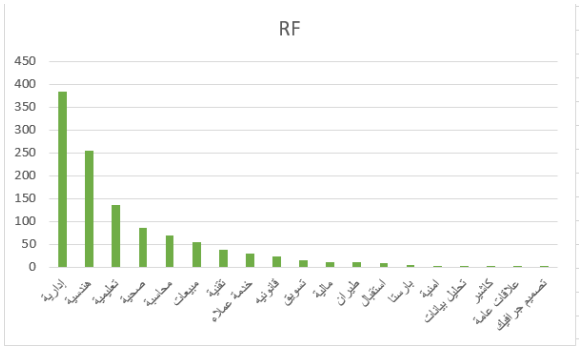

Figure.3 RF classifier prediction results.

As shown in Fig. 3, the trend of labor market needs is "وظائف ادارية", which have the maximum number of jobs equal 384. The second requested jobs are "هندسية " ،which has 255 requests.

The Figures 4 and 5 show the trend of labor market needs is "وظائف ادارية", according to M_NB and SVM. The second requested jobs are "هندسية".

According to the proposed system and the used techniques, we found that the "وظائف ادارية" is the trend of the labor market according to the Random Forest technique which has the maximum accuracy 93.6 from other techniques and precision and recall are 0.93, and 0.93.

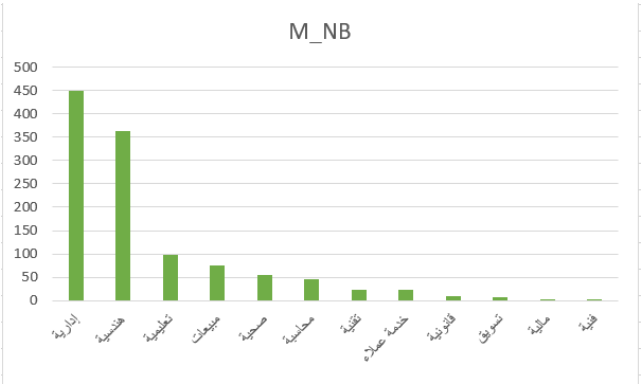

Figure. 4 the trend of labor market needs according to M_NB model

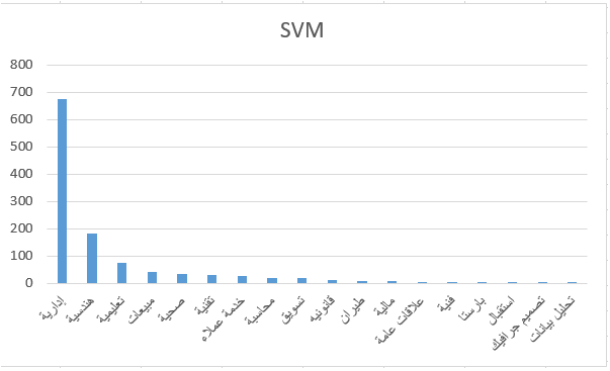

Figure. 5 the trend of labor market needs according to SVM model

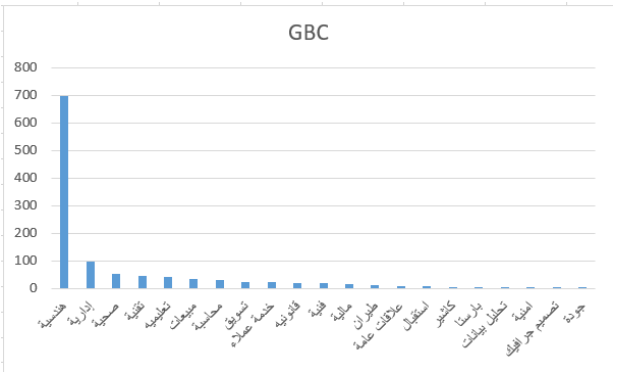

Figure. 6 the trend of labor market needs according to GBC model

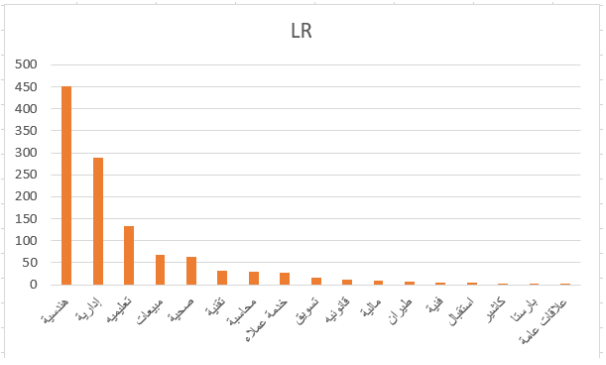

Figure. 7 the trend of labor market needs according to LR model

The Figures 6 and 7 show the trend of labor market needs is "هندسية", according to GBC and LR. The second requested job is "وظائف ادارية".

V. CONCLUSION

We collected 14 thousand tweets, but after cleaning and removing irrelevant data, we experimented the results in a set of 4000 tweets extracted from Twitter social network. To quantify the labor market needs, we applied five text classification techniques which are Support Vector Machine, Logistic Regression, Random Forest, Gradient Boosting Classifier, and Naive Bayes. We evaluated our work by five different measures which are Precision, Recall, Accuracy and F-measure.

The results showed that RF achieved the best accuracy and F-measure, and it equals 93.62%, 93% respectively. In addition, we found that the trend of labor market needed was "وظائف ادارية". The percentage of that job class that related to 384 jobs about 19.2%.

Finally, this paper presents a new system for extracting the needs of the labor market in Riyadh through Twitter using text classification techniques. This classification is done over Arabic language tweets.

As a future work we will develop the proposed architecture to explore the requirements of the local labor market in Arabic text and utilize an experiment-oriented approach to identify some of the tweets that have anything related to the needs of the local labor market.

VI. REFERENCES

[1] U. Manzoor, S.A. Baig, M. Hashim and A. Sami, Impact of social media marketing on consumer's purchase intentions: the mediating role of customer trust. International Journal of Entrepreneurial Research, Vol. 3,2020, pp. 41-48.

[2] G. Appel, L. Grewal, R.Hadi, and A.T. Stephen, The future of social media in marketing, Journal of the Academy of Marketing Science, Vol. 48, 2020, pp. 79-95.

[3] A. Hernández-Fuentes, and A. Monnier, Twitter as a Source of Information? Practices of Journalists Working for the French National Press, Journalism Practice 5, 2020, Vol. 16, pp. 920-937.

[4] N. Erskine and S. Hendricks, The use of Twitter by medical journals: systematic review of the literature, Journal of medical Internet research, Vol. 23, 7, 2021, , p. e26378.

[5] V. Alena, and I. Kalinouskaya, Better understanding of the labour market using Big Data, 3, s.l. : Ekonomia i Prawo, Vol. 20, 2021, , pp. 677–692.

[6] R. Boselli, M. Cesarini, F. Mercorio and M. Mezzanzanica, Labour Market Intelligence for Supporting Decision Making. s.l. : In SEBD, 2017, p. 74.

[7] R. Boselli M. Cesarini F. Mercorio M. Mezzanzanica, Classifying online Job Advertisements through Machine Learning, Future Generation Computer Systems-the International Journal of Escience , 86, 2018,pp. 319-328.

[8] T-L. Wong, W. Lam, B. Chen, Mining Employment Market via Text Block Detection and Adaptive Cross-Domain Information Extraction, s.l. : Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July, 2009, pp. 19-23.

[9] Y. Balcioglu, M. artar and O. irdil, Machine learning and analysis of twitter data identifying trend jobs after covid-19, S.l. : vii. International battalgazi scientific studies congress at: Malatya (2022), 2022.

[10] A. de Arriba Serra, M. Oriol Hilari, and J. Franch Gutiérrez, Applying sentiment analysis on Spanish tweets using BETO international Conference of the Spanish Society for Natural Language Processing: Málaga, Spain : CEUR-WS. Org., sep. 2021, In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021): co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), pp. 1-8.

[11] V. Kharde, and P. Sonawane, Sentiment analysis of twitter data: a survey of techniques, International Journal of Computer Applications , Vol. 139, 11, april 2016, pp. 975 – 8887.

[12] F. Resyanto, Y. Sibaroni, and A. Romadhony, Choosing the most optimum text preprocessing method for sentiment analysis: Case: iPhone Tweets. s.l. : IEEE, 2019. 2019 Fourth International Conference on Informatics and Computing (ICIC). pp. 1-5.

[13] M. Heikal, M. Torki, N. El-Makky, Sentiment Analysis of Arabic Tweets using Deep Learning, s.l. : Procedia Computer Science, 2018, pp. 114-122.

[14] J. Paralic and P. Bednar, text mining for document annotation and ontology support. s.l. : Intelligent Systems at the Service of Mankind, 2003, pp. 237-248.

[15] H. k. Kam, Random decision forest, Proceedings of the 3rd International Conference on Document Analysis and Recognition. Vol. 1416. Montreal, Canada, August, 1995.

[16] C. Cortes, and V. Vladimir ,Support vector networks, Machine learning, vol.20(3) 1995,pp. 273-297.

[17] B. Liu, E. Blasch, Y. Chen and D. Shen, Scalable sentiment classification for big data analysis using naive bayes classifier. Liu, B., et al. s.l. : In Proceedings of the 2013 IEEE International Conference on Big Data, Silicon Valley, CA, USA, ,9 October 2013 ; pp. 99–104.

[18] K. Kowsari, K., Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, Text classification algorithms, s.l. : A survey. Information 2019, 10(4), 150.

[19] J.H Friedman, Stochastic gradient boosting. " Computational stsistics & data analysis vol.38(4), 2002, pp.367-378.