A non-invasive machine learning mechanism for early disease recognition on Twitter: The case of anemia

Sarsam, S., Al-Samarraie, H., Alzahrani, A. I. & Shibghatullah, A. S.

Original citation:
Sarsam, S, Al-Samarraie, H, Alzahrani, AI & Shibghatullah, AS 2022, 'A non-invasive machine learning mechanism for early disease recognition on Twitter: The case of anemia', *Artificial Intelligence in Medicine*, vol. 134, 102428.
https://dx.doi.org/10.1016/j.artmed.2022.102428

DOI 10.1016/j.artmed.2022.102428
ISSN 0933-3657
ESSN 1873-2860

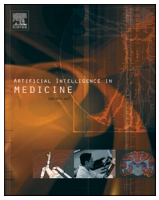
Publisher: Elsevier

Artificial Intelligence In Medicine 134 (2022) 102428

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed



A non-invasive machine learning mechanism for early disease recognition on Twitter: The case of anemia

Samer Muthana Sarsam [a,*], Hosam Al-Samarraie [b,c], Ahmed Ibrahim Alzahrani [d], Abdul Samad Shibghatullah [e]

[a] School of Strategy & Leadership, Coventry University, Coventry, UK
[b] School of Design, University of Leeds, Leeds, UK
[c] Centre for Instructional Technology & Multimedia, Universiti Sains Malaysia, Penang, Malaysia
[d] Computer Science Department, Community College, King Saud University, Riyadh, Saudi Arabia
[e] Institute of Computer Science and Digital Innovation, UCSI University, Kuala Lumpur, Malaysia

ARTICLE INFO

Keywords:
Health monitoring systems
Anemia recognition
Lexicon-based approach
Twitter
Machine learning

ABSTRACT

Social media sites, such as Twitter, provide the means for users to share their stories, feelings, and health conditions during the disease course. Anemia, the most common type of blood disorder, is recognized as a major public health problem all over the world. Yet very few studies have explored the potential of recognizing anemia from online posts. This study proposed a novel mechanism for recognizing anemia based on the associations between disease symptoms and patients' emotions posted on the Twitter platform. We used k-means and Latent Dirichlet Allocation (LDA) algorithms to group similar tweets and to identify hidden disease topics. Both disease emotions and symptoms were mapped using the Apriori algorithm. The proposed approach was evaluated using a number of classifiers. A higher prediction accuracy of 98.96 % was achieved using Sequential Minimal Optimization (SMO). The results revealed that fear and sadness emotions are dominant among anemic patients. The proposed mechanism is the first of its kind to diagnose anemia using textual information posted on social media sites. It can advance the development of intelligent health monitoring systems and clinical decision-support systems.

1. Introduction

Current efforts to diagnose and identify blood disorder diseases have been progressively popular. Blood disorder diseases are categorized into three categories: red blood cell disease, white blood cell disease, and platelet disease. The advantages of red blood cells include: (a) no surface antigens; (b) more e3szxconvenient to store than natural blood; and (c) function as effectively as hemoglobin [1]. Anemia is an example of red blood disease that affects more than two billion people around the world [2,3]. It is associated with impairment in oxygen transport which affects an individual's physical and mental wellbeing, and work performance [4]. Also, it is the main contributor to sustained fatigue—the most popular reported symptom among cancer patients [5,6]. Clinically, anemia disease can be categorized based on the morphology of red blood cells, underlying etiologic mechanisms, and discernible clinical spectra. The three main classes of anemia include excessive blood loss (acutely such as a hemorrhage or chronically through low-volume loss), excessive blood cell destruction (hemolysis), and deficient red blood cell production (ineffective hematopoiesis) [7]. Patients with anemia are likely to report different symptoms, including shortness of breath, weakness, fatigue, and arrhythmias [8]. The most common clinical signs of individuals with mild to moderate anemia are pale or sometimes yellow skin, pale cheeks, and lips, irritability, and mild weakness. Those with more severe forms of anemia may manifest more severe complaints such as shortness of breath, tachycardia, dizziness, headaches, and restless leg syndrome [9]. In general, anemia symptoms are vague and may result in a person feeling tired, weak, and poor ability to perform tasks [10]. Based on these, the prevention early-stage anemia has the potential to reduce the amount of hassle among patients and eliminate serious complications of severe anemia conditions.

The popular anemia recognition approach is accomplished by evaluating the level of hemoglobin concentration in the blood using a

* Corresponding author.
E-mail addresses: samer.sarsam@coventry.ac.uk (S.M. Sarsam), h.alsamarraie@leeds.ac.uk (H. Al-Samarraie), ahmed@ksu.edu.sa (A.I. Alzahrani), abdulsamad@ucsiuniversity.edu.my (A.S. Shibghatullah).

https://doi.org/10.1016/j.artmed.2022.102428
Received 9 October 2021; Received in revised form 10 September 2022; Accepted 13 October 2022
Available online 19 October 2022

complete blood count, which is an invasive, time-consuming, and costly technique [11]. It is also painful for patients as well as exposes the operators to the risk of infection [12]. Hence, the crucial need for a low-cost method for anemia detection has encouraged several scholars to propose non-invasive methods that allow diagnosing the disease safely within a short time and low budget. For instance, Chen, Miaou, and Bian [13] proposed an image-based method that did not relay on any sort of blood testing by only examining the color distribution of palpebral conjunctiva. The authors used two algorithms to diagnose anemia: The first algorithm was fast and simple (a two-stage classification), while the second one was sophisticated as it depends at the pixel value in the middle and minimum distance classifier (Mahalanobis distance). They found that the proposed method was able to predict anemia with reasonable accuracy (78 %). Another study by Chen and Miaou [14] proposed an anemia detection approach using a Kalman filter and a regression method. For this purpose, the Kalman filter was modified to fit the inserted time-independent data. The authors also computed the mean value of the red component of the palpebral conjunctiva image before applying a regression algorithm in which the corresponding levels of hemoglobin concentration showed an accuracy of 80 %. Another work by Jain et al. [11] used a neural network method to detect anemic patients from the images of eye conjunctiva. In this sense, backpropagation was implemented to adjust the weights for the utilized algorithm. The suggested method achieved 97 % prediction accuracy. Tamir et al. [15] used a support vector machine (SVM) approach to detect anemia from images of eye conjunctiva in which they achieved 78.9 % accuracy. Besides previous work, Dimauro et al. [16] applied another technique for predicting anemia using the k-nearest neighbor (KNN) algorithm, and their method achieved 90.26 % accuracy after being tested over images of several anemic patients and non-anemic patients. Bevilacqua et al. [17] developed an approach to capture images of eye conjunctiva then used these images to estimate hemoglobin level in blood and then predict (with 84.4 % accuracy) whether the patient is suffering from anemia or not using the SVM algorithm.

However, despite these efforts, it can be observed that images of eye conjunctiva of patients were mainly used in the process of anemia recognition. Such approach does not work well in many underdeveloped areas, due to a lack of medical facilities and capabilities to deal with small datasets [11]. In addition, the extensive use of images to detect anemia requires a huge number of images which was not fully accounted for by prior works [11]. An additional issue in anemia recognition is that patients' emotions were not considered in the detection process of previous studies [18]. This led us to assume that data from social media websites may somehow help/facilitate the recognition process of anemia. Therefore, this work aims at proposing a non-invasive mechanism to diagnose anemia from the Twitter platform. This study intended to answer two questions: "What are the main anemia-related topics shared by Twitter users?" and "What types of emotions should be used in the identification of anemia symptoms?". To answer these questions, we used a topic modeling approach to extract different anemia-related topics that are extensively shared between social media users. Also, we analyzed users' (or patients') sentiments existed in their tweets in order to determine the type of emotions (e.g., anger, fear, sadness, and joy) that can be associated with certain anemia-related symptoms.

2. Literature review

Patients with anemia experience different types of emotions according to their physical and mental conditions. Fear-related emotions are usually observed among anemia patients in relation to the severity of their health condition [19]. Furthermore, fear-related emotions are usually found among anemic women due to the reduced energy and capacity for work [20]. It is also assumed that fear is the dominant emotion of pregnant women with anemia since these women tends to be more concerned about facing poor pregnancy and risk of death—consequences of anemia among pregnant women [21]. Anemia is the most common hematologic problem in patients with rheumatoid arthritis (RA) who usually experience fatigue as one of their most annoying problems. According to Singh et al. [22], several studies have reported that pain, physical disabilities, impaired general health, limited physical activities, and comorbid conditions can be used to describe such fatigue.

In addition, sadness was also found to be another type of emotion among anemic pregnant women, as well as depression [23]. Anemia has been playing a significant role in postpartum depression; it causes depression by changing inflammatory cytokinins. This was confirmed by Parhizkar [24] who found that there is a relation between anemia and postpartum depression in pregnant women that is associated with sadness and anxiety. Also, the sadness emotion was found to be associated with patients who suffer from iron deficiency anemia [25]. Previous scholars found that depression and anxiety are more frequent in children with iron deficiency anemia [26], while joy-related emotions are found among patients who experience health improvements [25].

The use of real-time methods becomes essential to both healthcare professionals and the public in infectious diseases. Social media networking sites have become important tools [27] because they allow the users to build their public personal profile, produce a list of users, and view a list of posts. These social communication mediums are critical real-time platforms for public and healthcare experts and to distribute and analyze medical information and alerts. Thus, using social media websites can be a robust approach to obtain users' opinions, emotions, and personal experiences in the health domain. This has motivated scholars like Lim, Tucker, and Kumara [28] to identify real-world hidden infectious diseases by analyzing social media data. The authors applied an unsupervised learning technique to mine social media that have temporal information to provide a bottom-up approach for latent infectious disease discovery in a specific location. Twitter was extensively used by health experts as a reliable source of information, mainly to gain insights about public health across the world. Twitter is a microblogging version of a social media site where users interact in real-time through 140 characters called tweets. It also allows users to easily interact with each other through updates, direct messaging, likes, and retweeting. This, as a result, led several scholars to rely on Twitter messages as a vital source of health information to examine numerous health-related topics. For example, Twitter was used in past work for monitoring the spread of the swine flu (H1N1) outbreak in 2009 [29]. Odlum and Yoon [30] also showed the possibility of analyzing health topics on Twitter. The researchers analyzed users' tweets and observed the Ebola outbreak in West Africa before the official outbreak announcements. Sarsam, Al-Samarraie, Ismail, Zaqout, and Wright [31] proposed a novel early-stage disease recognition method to track and detect migraine disease from users' emotions embedded in their tweets–based on the interconnection between certain emotional types and climatic factors that are associated with migraine. Another work by Karami et al. [32] analyzed the public opinion in aspects related to diabetes, diet, exercise, and obesity using Twitter messages. A multi-component semantic and linguistic framework was designed to find relevant topics that were used along with the Latent Dirichlet Allocation (LDA) technique. In addition to prior work on disease recognition from Twitter, Sarsam, Al-Samarraie, and Al-Sadi [33] developed a heuristic mechanism by mapping between diabetes-related terms and emotions related to diabetes diseases. Based on these, there seems to be a great potential in using Twitter to analyze individuals' thoughts and opinions in relation to anemia.

3. Method

The main steps of this study are presented in Fig. 1. This includes data collection, data pre-processing, cluster analysis, topic modeling, emotion extraction, part-of-speech tagging, and association rules mining. This research procedure allowed for the identification of certain users' emotions (expressed in tweets) that can be linked to anemia.

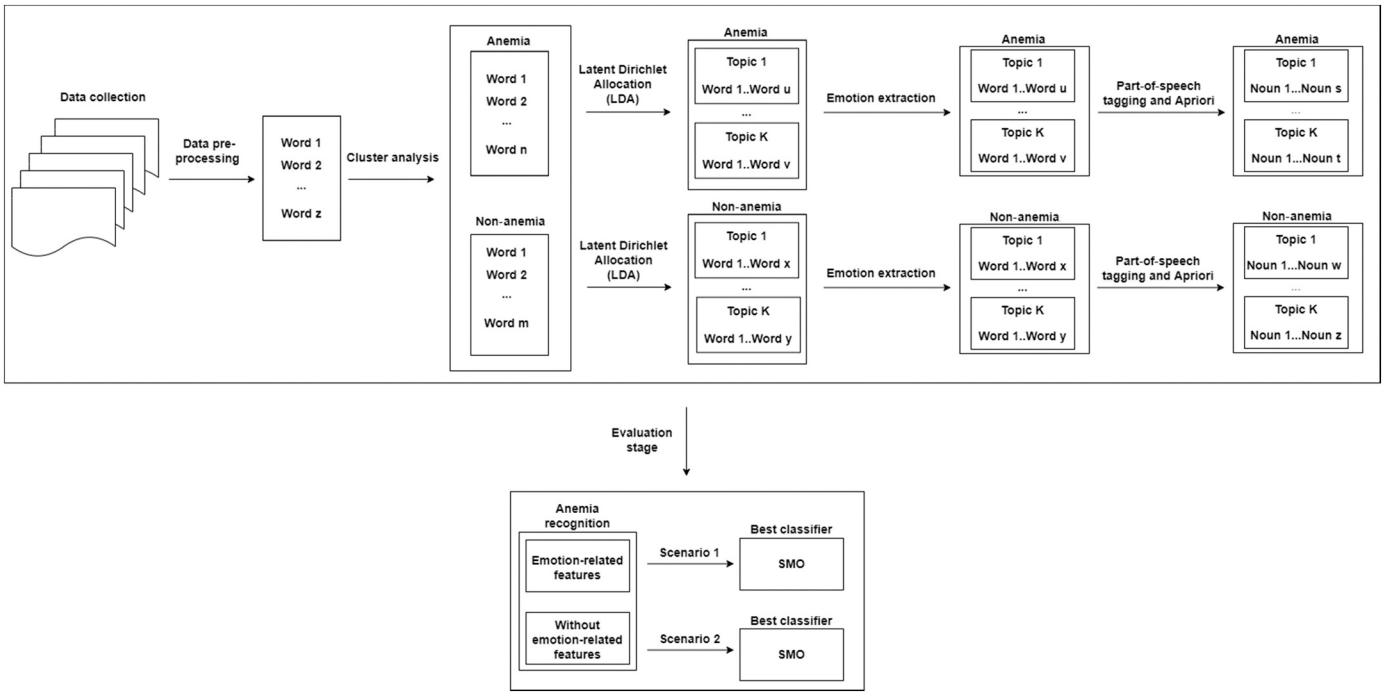


Fig. 1. General procedure.

Finally, the evaluation stage was designed to assess the merit of the discovered emotions. In this sense, two scenarios were considered—in the first scenario users' emotions were extracted and fed as an input for the classification task, while in the second scenario, emotions were not fed into the classifier.

3.1. Data collection

We collected and processed 1,738,759 English tweets within a time span of six months (December 1st 2019, till 31st May 2020). The Twitter free streaming Application Programming Interface (API) was applied based on the preious recommendations of Sarsam, Al-Samarraie, and Omar [34]. To collect the desired tweets, several keywords were used: 'I have anemia', 'anemia', 'cause of anemia'. After that, we performed several data preparation steps to make our data ready for the analysis stage.

3.2. Data pre-processing

At this stage, several pre-processing techniques were applied to extract solid knowledge out of it. The bag-of-words model and the "Tokenization" technique were implemented to extract the tweets features (words) and build the dictionary which we used to provide a numeric weight for each feature. Then, all the features, were converted to a lowercase form before applying the Stopwords list method. Stopwords list technique was used to keep the necessary words in the dictionary. After that, we utilized the L2 method to facilitate the normalization of the collected tweets. This method helped us to guarantee fair treatment to all the tweets by machine learning algorithms in the coming stages.

3.3. Cluster analysis

The K-means clustering algorithm was utilized [33,35] in order to group the processed tweets that share similar features. This was essential because some of the extracted tweets were not focusing on the causes and symptoms of anemia. It was assumed that by including the relevant anemia-related tweets we can increase the effectiveness of the proposed approach. For this reason, we used an unsupervised learning technique (k-means), which is well known for its superiority, to cluster (group) the relevant tweets based on the standard of similarity between data points (tweets) [36].

Due to the unsupervised learning nature of k-means, it is very challenging to determine the number of groups from the raw data. Therefore, the elbow method was utilized to tackle this issue. The results from utilizing the elbow method revealed two main groups (see Fig. 2a). To identify the anemia group from the collected tweets, we invited three specialists in blood diseases who recommended the labeling of the two groups as 'Anemia' and 'Non-anemia' groups. Based on Fig. 2b, the anemia group contains tweets that discussed anemia-related topics (the dark color represents the main features in the anemia group that could be used in the identification of the disease) as compared to the non-anemia group.

In this study, we argue that there are specific types of emotions related to anemia, which could facilitate the recognition of the disease. The relationship between users' emotions (expressed by anemic patients through their topics in the posted tweets) and the extracted anemia symptoms (expressed by anemic patients in these topics) was used in this study to aid the identification/recognition of the disease. This was achieved by the implementation of the association rules mining technique (see Section 3.7).

3.4. Topic modeling

From the previous stage, two groups of tweets were found and labeled based on the nature of the topics. At this stage, the LDA algorithm was implemented via the LDAvis system [37,38] to extract the topics from both the anemia and non-anemia groups. LDA is an unsupervised generative probabilistic method that is commonly used in corpus modeling. It is also one of the most used topic modeling methods in the literature. LDA works by dealing with random mixtures over latent topics, where a single topic is characterized by a distribution over certain words. It assumes that each document can be represented as a probabilistic distribution over latent topics, and that topic distribution in all documents share a common Dirichlet prior. In addition, the LDA model represents every single latent topic as a probabilistic distribution

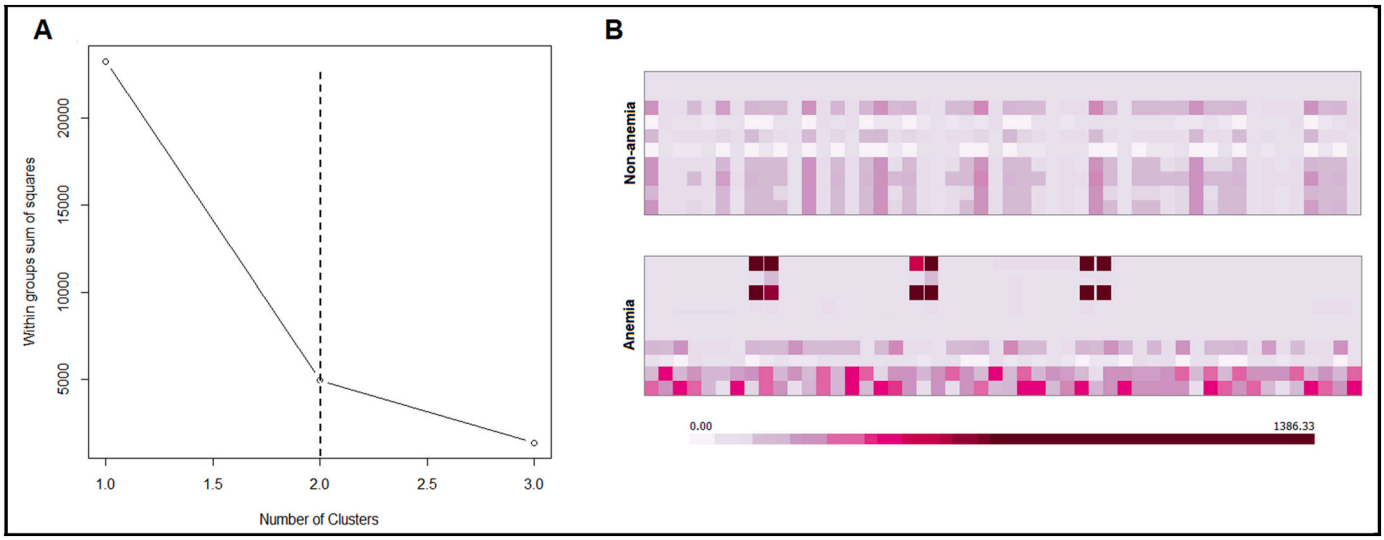


Fig. 2. K-means result. (A) Elbow method results. (B) Anemia and Non-anemia groups.

over certain words. The word distributions of topics are then linked with a common Dirichlet prior across the tweets.

3.5. Emotion extraction

After determining the key themes of the disease (e.g., anemia causes and anemia symptoms), we further identified the types of emotions associated with anemia. To extract the emotion from the tweets (texts), the lexicon-based method was applied using a predefined list of words where each word is associated with a specific type of emotion [39]. An example of the lexicon-based approach is NRC Affect Intensity Lexicon which we used to help us extract users' emotions from their textual data [40,41]. It includes a list of English words along with their associations that were used to reflect four types of emotions (anger, fear, sadness, and joy). The scores range from 0 to 1 was used to convey a given word and emotion X. A score of 1 represents the highest amount of emotion X. A score of 0 represents the lowest amount of emotion X. Then, the emotional features for each tweet were processed and identified by adding the relevant associations of the words for a given lexicon (see Section 4.1 for more details).

3.6. Part-of-speech tagging

To obtain anemia-related symptoms (noun words), we extracted grammatical constitution from the collected tweets. We used "Part-of-speech tagging" due to its popularity in social media analysis and in identifying terms that can be utilized in different parts of speech [42]. The Penn State Treebank tokenizer was applied to obtain words before using the probabilistic context-free grammar parser. This process helped us to extract 'noun' words from tweets which we later used for association rules mining. Some of these nouns were also used to form the terminologies of the anemia symptoms. The association rules approach was used to determine the relationship between the identified symptoms/terminologies and the type of emotions in the processed tweets.

3.7. Association rules mining

Since the main goal of this study is to recognize anemia from users' tweets, we attempted to link anemia symptoms with the patients' emotional experience of joy, sadness, or any other primary emotions. It was anticipated that after determining anemia-related topics and anemia-related emotions, the emotions of anemic patients in these topics can be further mapped with the key anima symptoms. For this purpose, the Apriori algorithm was used to find a meaningful set of relationships between anemia emotions and symptoms. We configured the Apriori algorithm by setting the delta value at 0.05 in order to reduce the support until minimum support is reached. The minimum metric score was set at 0.9, while the upper bound and lower bound support were set at 1.0. Then, we invoked the Apriori method on the data to predict the relationship between anemia-related emotions and terminologies.

3.8. Evaluation stage

In this stage we assessed the quality of the proposed method based on two scenarios. In the first scenario, we extracted and used users' emotions as an input for the machine learning model. In the second scenario, we did not relay on users' emotions as the main input for the machine learning model. In this sense, four classifiers were used and compared to select the best classifier with the highest prediction capability. These classifiers were: 1-rule classifier (OneR), RandomForest [43], Sequential Minimal Optimization (SMO) [44], and Bagging algorithm [45]. The classifiers were utilized within the Weka environment (Waikato Environment for Knowledge Analysis). The study utilized the stratified tenfold cross-validation method in the evaluation of the overall prediction process. A few evaluation metrics were applied to evaluate the prediction capability of each selected classifier (see Section 4.4).

4. Results

4.1. Emotion extraction result

After extracting the relevant emotions from the collected tweets, we attempted to determine the types of emotions that are associated with anemia. The main emotions of anger, fear, sadness, and joy were extracted from both anemia and non-anemia groups. The results (see Fig. 3) showed that both fear (81 %) and sadness (89 %) were the dominant types of emotions in the anemia group as compared with the non-anemia group (fear: 19 % and sadness: 11 %). In contrast, the percentages of anger (86 %) and joy (95 %) were found to be higher in the non-anemia group than their values in the anemia group (anger: 14 % and joy: 5 %). From these, it can be said that emotions related to fear and sadness found in the anemia group can aid the recognition/identification of the disease.

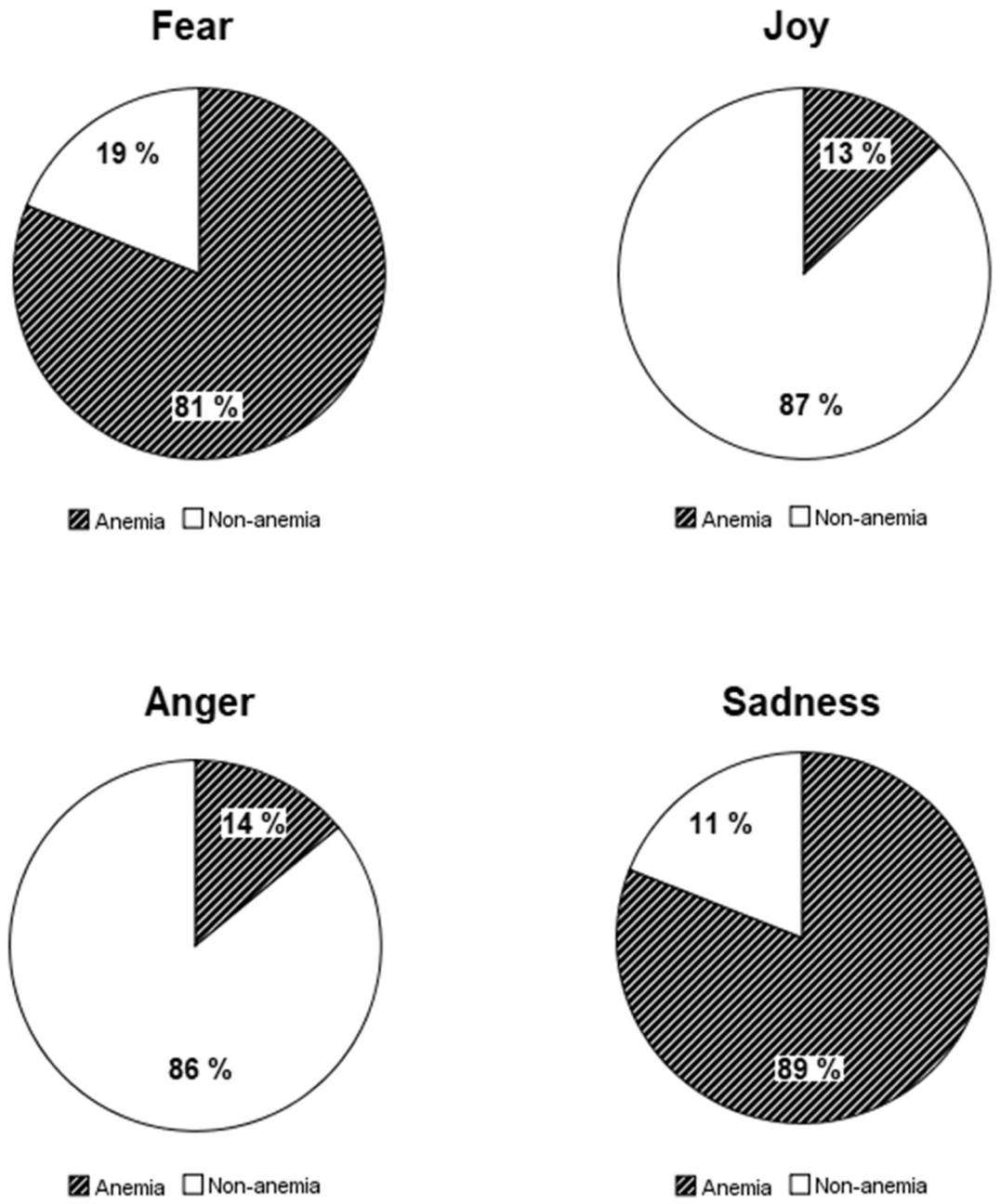


Fig. 3. The extracted emotions in anemia and non-anemia groups.

4.2. Result of topic modeling

The LDA method was configured by performing several experiments. We firstly determined the number of topics based on the trade-off between information loss and information overload. After several experiments, we confirmed the number of topics to be examined. Fig. 4 exhibits the result of the LDAvis tool where each circle represents a specific topic from the collected tweets, while the size of the circle demonstrates the frequency of the topic. In addition, the distance between the circles reflects the similarity between these topics. From the figure, we can see that some topics are far apart (independent), whereas some topics are relatively close or even overlapping (a high level of similarity). However, the topic modeling approach is known to be limited in terms of labeling the resulted topics. In order to overcome this challenge, the labeling process was accomplished manually by humans based on the content of these topics. Therefore, we asked three specialists in blood diseases to individually inspect and label the resulting topics using relevant themes. After that, we measured the agreement between the three-labeling methods using kappa statistic which showed 0.96. According to the specialists, the main topics in the non-anemia group can be classified into: 1) personal opinions about anemia and 2) questions about anemia, whereas the anemia group had two themes: 1) anemia causes and 2) anemia symptoms. Consequently, to map the relationship between anemia-related emotions and anemia symptoms, we only analyzed the content found in both themes of anemia causes and anemia symptoms.

When inspecting the first theme of anemia causes, we observed that most of the cases were related to iron deficiency anemia, anemia related to pregnancy, and hemolytic anemia. These cases are shown (see Fig. 4) in topics 1, 3, and 2, respectively. On the other hand, our observation of the second theme (anemia symptoms) resulted in topics related to symptoms such as fatigue, weakness, and shortness of breath. Fig. 5 shows the top frequent words associated with disease-related themes.

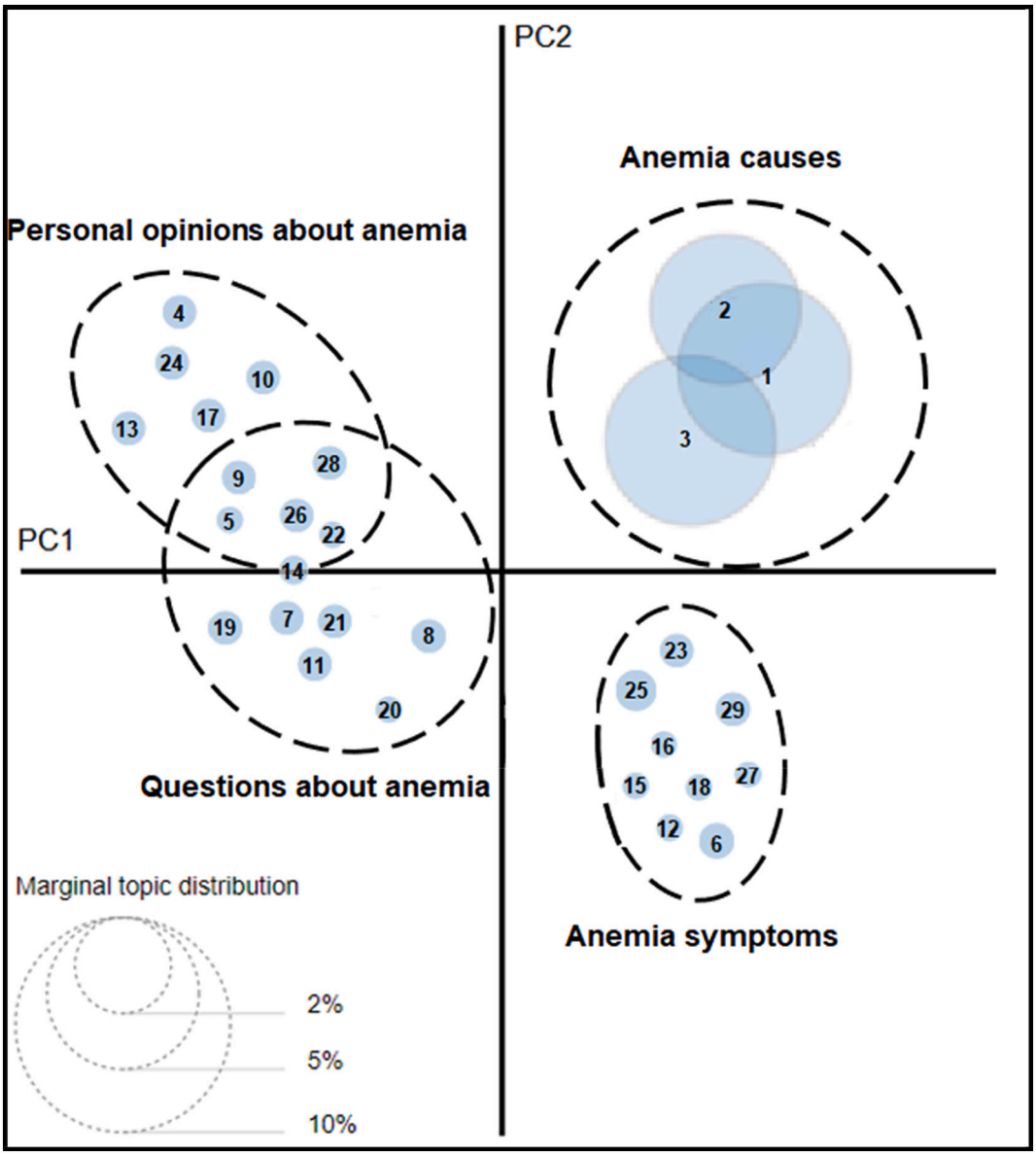


Fig. 4. Topic modeling of tweets related to anemia.

4.3. Results of association rules

To examine the possibility of using users' emotion as an investigative mechanism to recognize/identify anemia, we extracted these emotions using the NRC Affect Intensity Lexicon approach, where anger, fear, sadness, and joy emotions were extracted and mapped via the Apriori algorithm with anemia-related symptoms (see Fig. 6). From the figure, it can be clearly observed that emotions related to fear and sadness were associated with anemia symptoms. These symptoms are demonstrated in Fig. 6a: fatigue, weakness, shortness of breath, looking pale, and dizziness. From this, it can be concluded that there is a potential relationship between anima disease and specific types of emotions (fear and sadness).

4.4. Evaluation results

To evaluate the merit of the extracted emotions and their relations to anemia, we used four classifiers to evaluate the proposed approach (OneR, RandomForest, SMO, and Bagging) based on two scenarios (with emotions and without emotions). To assess the quality of the recognition results, several evaluation metrics were implemented based on the recommendations of Han, Kamber, and Pei [46], including:

$$\text{Accuracy} = \frac{TP + TN}{P + N} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

where True positives (*TP*), True negatives (*TN*), False positives (*FP*), False negatives (*FN*), Positive (*P*), and Negative (*N*) instances.

Besides, the kappa statistic was used to ensure the inter-rater reliability of the applied classification algorithms [47]. The optimal case of the kappa statistic is one (perfect agreement). The formula for the kappa statistic is represented as follows:

$$\text{Kappa statistic} = \frac{P_0 - P_e}{1 - P_e}$$

where P_0 is the actual observed agreement among the raters, while P_e is

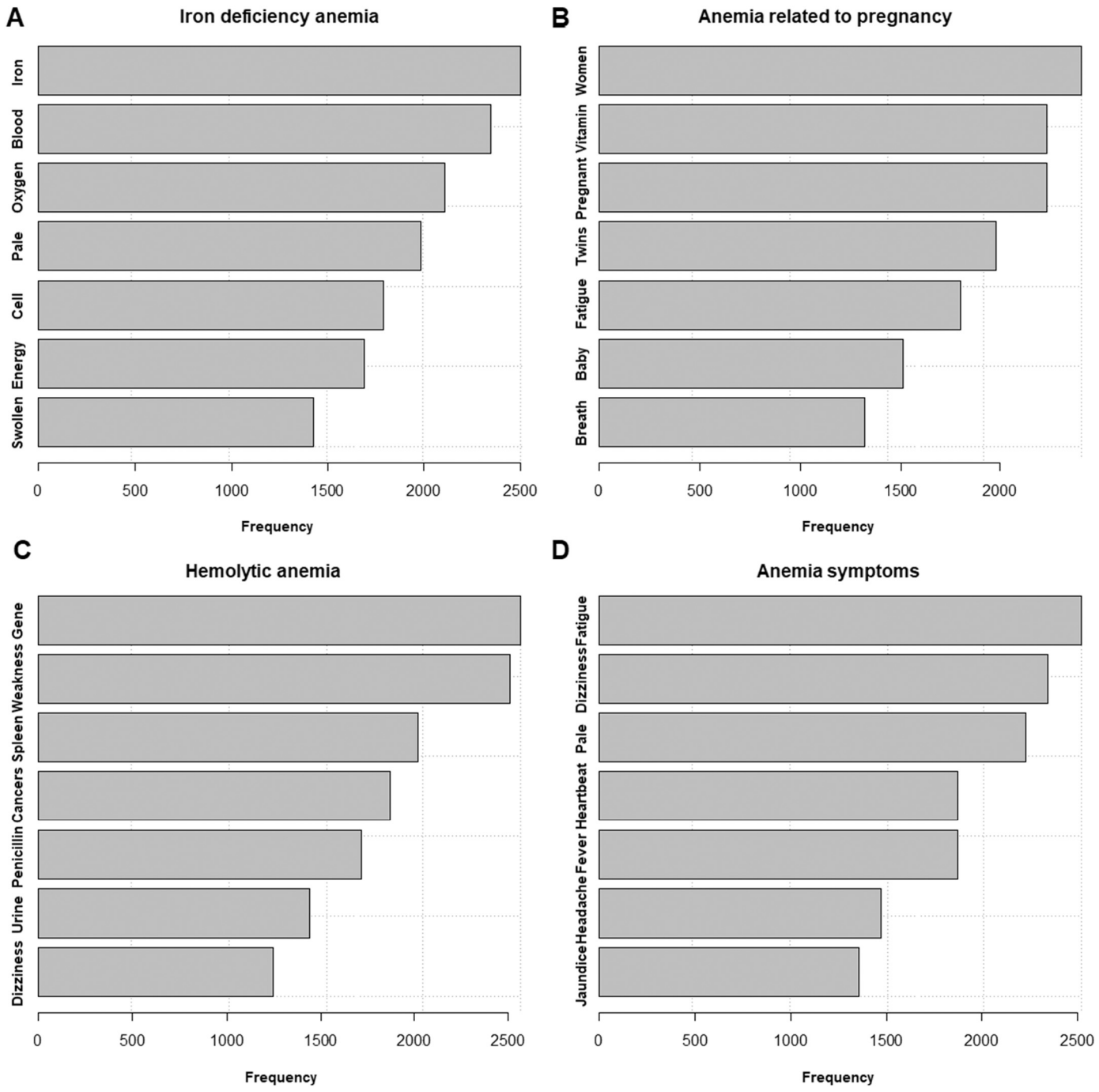


Fig. 5. Top frequent words in the anemia group.

the probability of chance agreement. Also, Root mean-squared error (RMSE) was computed, as recommended by Witten et al. [48], as follows:

$$\text{RMSE} = \sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}} \tag{4}$$

Here, n refers to the sample size. The predicted values on the test instances were p_1, p_2, …, p_n, while the actual values were a_1, a_2, …, a_n. According to Witten et al. [48], we computed the value of ROC curve using the formulas of False Positive Rate and True Positive Rate as shown below:

$$\text{False Positive Rate} = 100 \times \frac{FP}{FP + TN} \tag{5}$$

$$\text{True Positive Rate} = 100 \times \frac{TP}{TP + FN} \tag{6}$$

Our classification results are summarized in Table 1. The results from the first scenario (with emotions) showed that the SMO classifier achieved the highest classification accuracy (98.96 %), followed by Bagging (63.44 %), OneR (56.09 %), and RandomForest (41.82 %) algorithms. In addition, the SMO algorithm had the highest kappa statistic value (98 %) compared to Bagging (57 %), OneR (43 %), and RandomForest (35 %), respectively. However, the classification results illustrated that the RandomForest classifier produced the highest RMSE value (89 %),

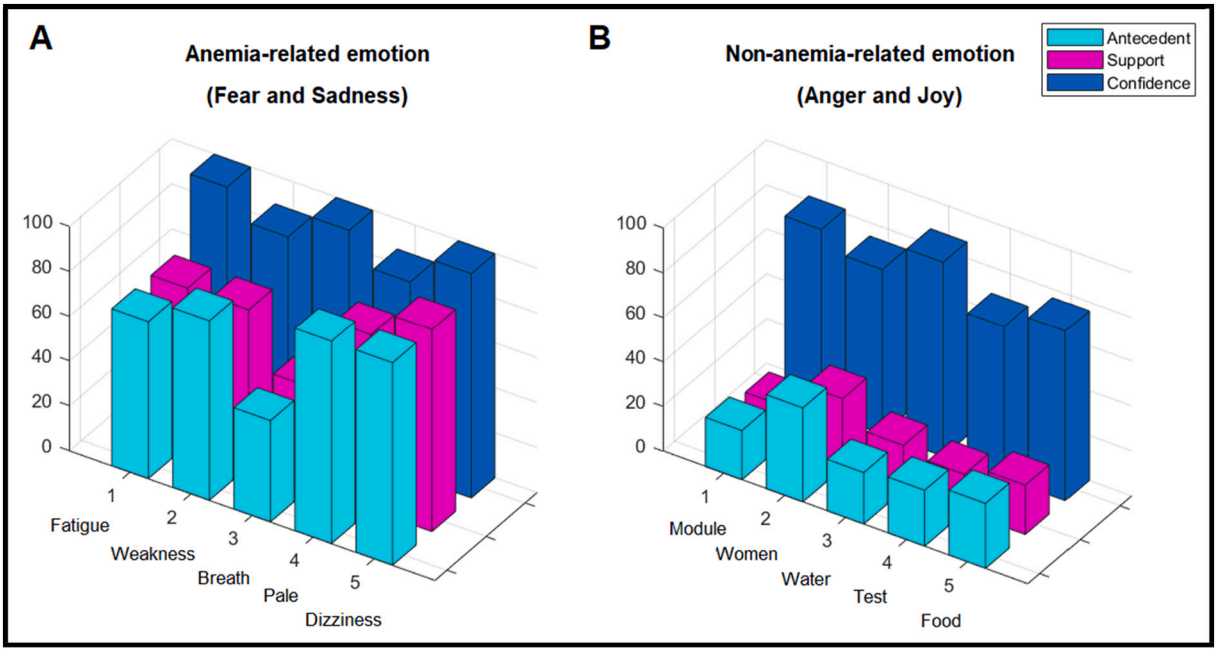


Fig. 6. Association rules results.

Table 1
Classification results.

Emotion	Algorithm	Accuracy (%)	Kappa statistic (%)	RMSE (%)	Precision (%)	Recall (%)
With emotions	SMO	98.96	98	5	96	89
	Bagging	63.44	57	48	85	82
	OneR	56.09	43	79	57	63
	RandomForest	41.82	35	89	48	17
Without emotions	SMO	64.12	65	24	71	62
	Bagging	56.08	57	66	64	56
	OneR	44.11	44	87	51	49
	RandomForest	35.90	36	96	42	42

followed by OneR (79 %), Bagging (48 %), and SMO (5 %), respectively. Moreover, SMO scored the highest precision (96 %) and recall (89 %) values, followed by Bagging (85 % precision, 82 % recall), OneR (57 % precision, 63 % recall) and RandomForest (48 % precision, 17 % recall).

The results from the second scenario (without emotions) indicated that SMO produced the highest classification accuracy (64.12 %), followed by Bagging (56.08 %), OneR (44.11 %), and RandomForest (35.90 %) schemes. We also found that SMO had the highest kappa statistic value (65 %) compared to Bagging (57 %), OneR (44 %), and RandomForest (36 %), respectively. In contrast, the RandomForest classifier scored the highest RMSE value (96 %), followed by OneR (87 %), Bagging (66 %), and SMO (24 %), respectively. SMO scored the highest precision (71 %) and recall (62 %) values, followed by Bagging (64 % precision, 56 % recall), OneR (51 % precision, 49 % recall) and RandomForest (42 % precision, 40 % recall).

To further understand the performance of the four classifiers, Fig. 7a and b showed the ROC curve results where it can be clearly observed that the SMO classifier outperformed other classifiers. Fig. 7c shows the accuracy, kappa statistic, RMSE, precision, and recall for each classification algorithm. Based on these results, it can be concluded that anemia-related emotions can significantly improve the recognition performance of anemia.

The "Confusion matrix" approach was used to provide an in-depth understanding of the relationship between the predicted and the actual instances along the diagonal of the confusion matrix. The final results (see Fig. 8) revealed that SMO classifier has the highest predictive capability between actual and predicted classes, i.e., 100 %, for both positive and negative categories.

5. Discussion

This study proposed a novel non-invasive mechanism to diagnose anemia disease from Twitter messages. Our result showed that both fear and sadness emotions were associated with anemia disease. This work confirms the feasibility of performing anemia recognition using a lexicon-based approach by producing sentimental features that are mapped with disease-related symptoms. In addition, our topic modeling result via the LDA method revealed that four main anemia-related themes that were frequently discussed on Twitter: anemia causes, anemia symptoms, personal opinions about anemia, and anemia-related questions. Discussing such themes is sensible since anima is one of the widespread diseases affecting the wellbeing of people around the globe [49].

Our LDA results showed that 'iron deficiency anemia', 'anemia related to pregnancy', and 'hemolytic anemia' were commonly reported/shared among social media users as the main cause of anemia. This finding is in line with the literature that declared the importance of these topics as causal factors among anemic patients [50–53]. More precisely, the first topic (iron deficiency anemia) is found to be among the most important contributing factors to the global burden of anemia disease. This is associated with work of Mohamed [54] which stated that iron deficiency anemia is considered to be the top-ranking cause of the worldwide type of anemia. This is due to deficiency of food if intake of iron less or incomplete. In addition, iron deficiency anemia is

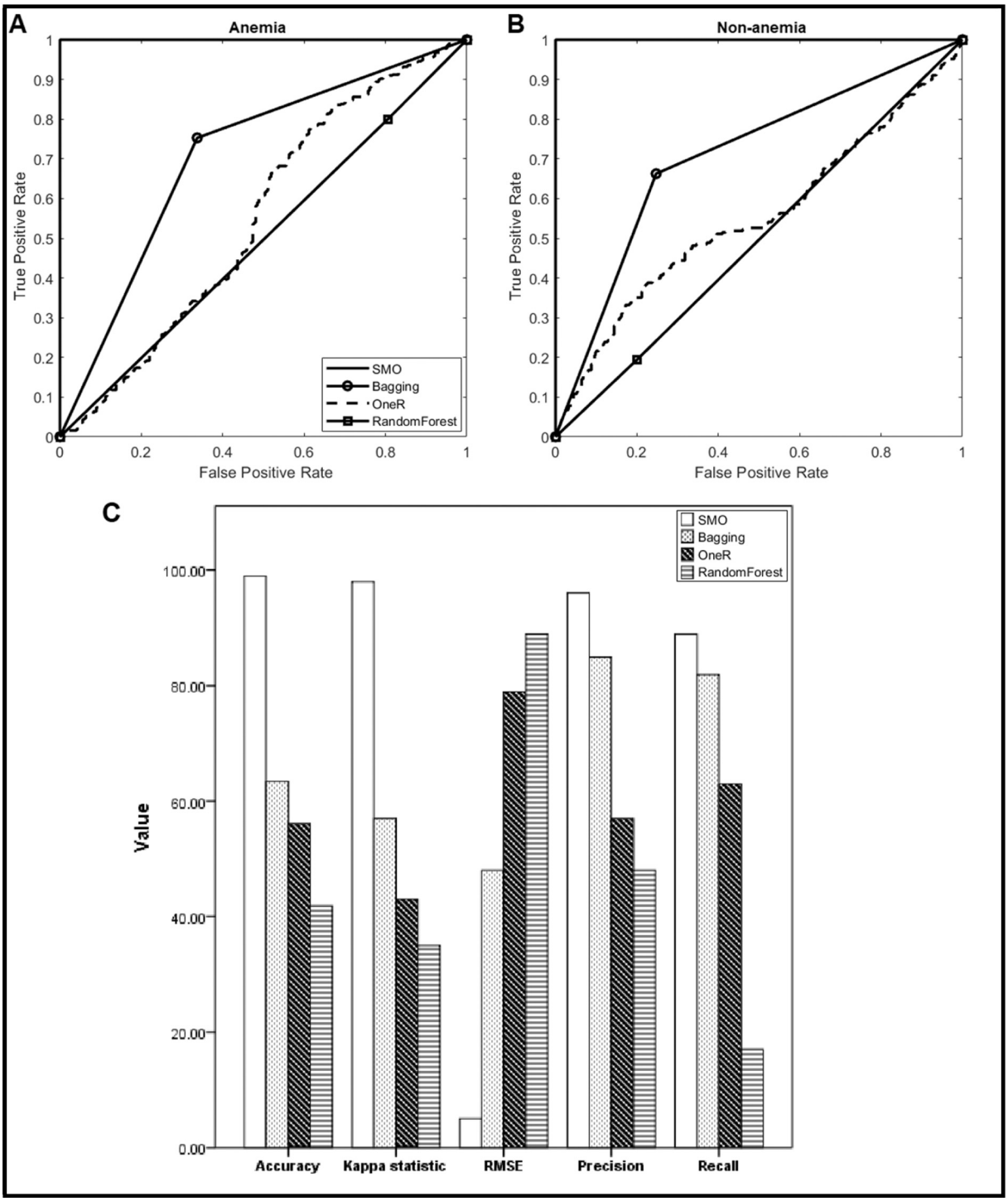


Fig. 7. Evaluation metrics of the four algorithms.

characterized by decreased hemoglobin synthesis leading to hypochromic and microcytic red blood cell production. The absolute iron deficiency arises when total body iron stores are low or exhausted; functional iron deficiency is a disorder in which total body iron stores are normal or increased, but the iron supply to the bone marrow is inadequate. Patients with iron deficiency anemia can experience symptoms such as pallor of the skin, while in severe cases patients could have dyspnoea at rest, angina pectoris, and haemodynamic instability [55]. In addition, iron deficiency affects epithelial cells with a rapid turnover, which result in dryness and roughness of the skin, hair damaged, and moderate alopecia. Iron deficiency anemia summarizes approximately 50 % of nutrient-scarred anemia cases, in which bleeding caused by gastrointestinal lesions is considered to be the first cause. All of that explains the importance of iron deficiency anemia that was observed in our results.

The second important topic found in our results was in the relationship between anemia and pregnancy. Anemia during pregnancy is a serious public health problem, as confirmed by prior work like Anggraeni and Fatoni [50] who stated that more than 50 % of pregnant mothers in developing countries are anemic. This is because during pregnancy, a number of changes take place in the body of the mother, including the blood [7]. For the period of pregnancy, iron needs are

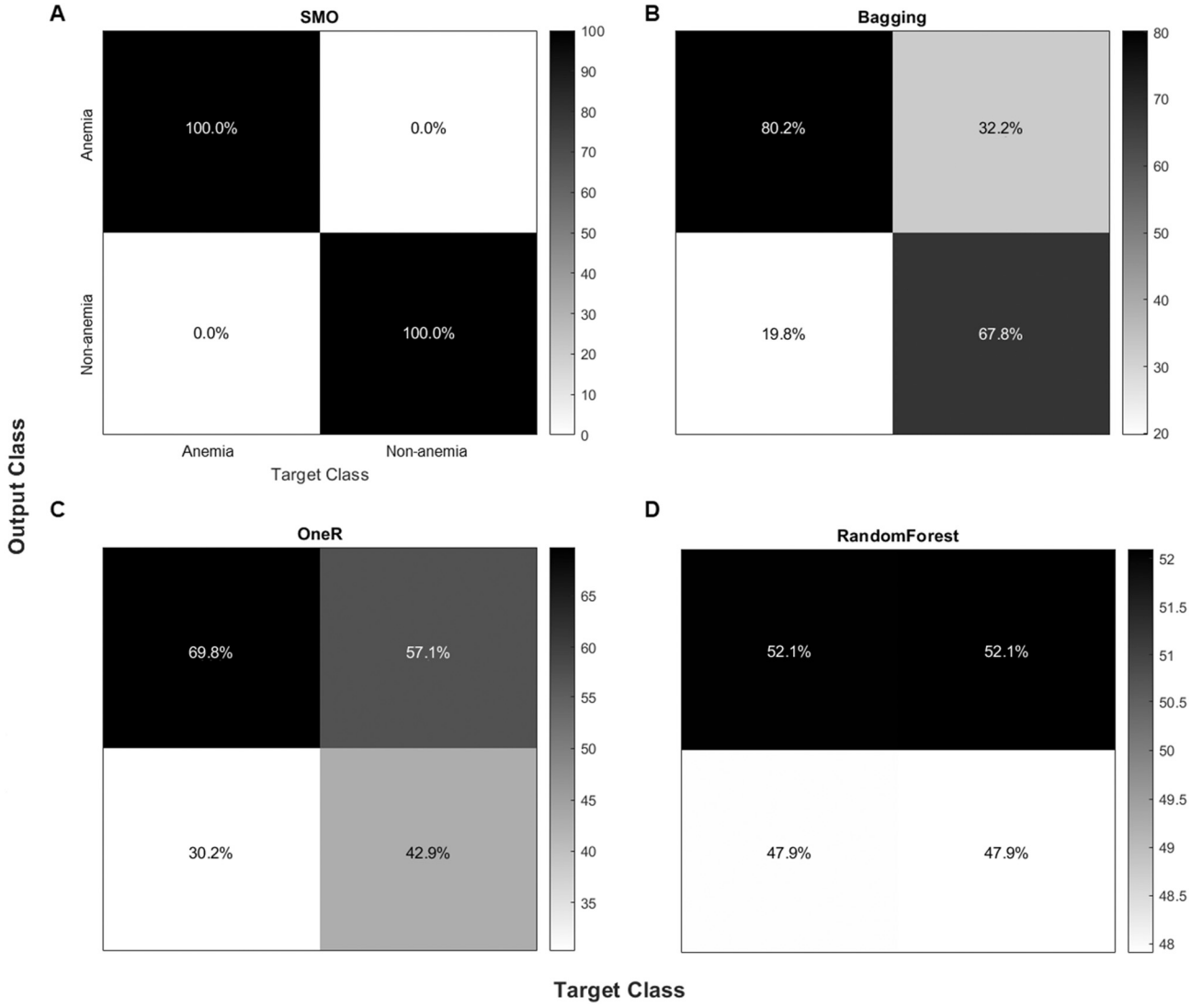


Fig. 8. Confusion matrix results.

tripled due to the expansion of maternal red cell mass and growth of the fetus and placenta. Water weight and fluid gain during pregnancy dilute the blood, which might be reflected as anemia since the relative concentration of red blood cells is lower. Also, in pregnancy, iron deficiency is associated with an increased risk of preterm labor, low neonatal weight, and increased newborn and maternal mortality. All these details explained the importance of anemia related to pregnancy topic, and thus justify the reason behind gaining a high-level of attention among Twitter users, as we found in our result. Finally, the third topic that was frequently repeated in our result was hemolytic anemia. This topic seems to be very important as confirmed by prior work like Capriotti and Frizzell [56] because it is a blood disorder that can be a chronic and life-threatening cause. Basically, hemolytic anemia is the destruction of red blood cells before their normal 120-day life span [57]. Capriotti and Frizzell [56] revealed that hemolytic anemia accounts for 5 % of all existing anemias. It is a critical anemia type that happens due to the abnormal breakdown of red blood cells either in the blood vessels or elsewhere in the humans' body. The hemolytic anemia mostly occurs within the spleen and could occur in the reticuloendothelial system or mechanically. All of that provide clear evidence about the importance of hemolytic anemia-related symptoms in the recognition process.

This study also found that fear and sadness emotions were associated with anemia disease. This seems to be reasonable due to the recurrent anemia-related symptoms that were reported among anemic patients, including fatigue, dizziness or, pallor, chest pain, irregular heartbeat, cold hands and feet, and shortness of breath [49]. This finding extends the findings of previous work about how fear-related emotions are common among people who suffer from anemia, especially anemic women, due to the reduced energy and ability to perform the required effort [20]. Besides fear, sadness-related emotions were frequently recognized among anemia patients due to the high level of depression. This was confirmed in the literature since many anemic pregnant women were found to experience sadness [24]. The potential relationship between anemic people and the psychological state that they suffer from can, therefore, be used to unlock different recognition opportunities of diseases.

6. Implications

To the best of our knowledge, this is the first study to propose a non-invasive mechanism for diagnosing anemia disease from Twitter messages. The proposed mechanism contributes to the development of clinical decision support systems that process evidence-based guidelines to extract latent associations between the disease and its underlying features. In other words, our approach sheds a light on the importance of establishing intelligent associations between patients' emotions

embedded in their tweets and the anemia symptoms that they experience. The proposed mechanism is cost-effective and can be effectively used for anemia recognition in underdeveloped areas where medical facilities are insufficient. Furthermore, our method can be extremely time-effective and can be used not only for anemia recognition but also for other blood disorder diseases since it is based on posts available on social media platforms. Finally, the use of topic modeling showed an impressive result that can be used to explore health topics embedded in a massive amount of data.

7. Limitations and future works

Despite the efficiency of the proposed machine learning mechanism, there are still some limitations need to be tackled. For example, only tweets in the English language were analyzed since English is the most popular language in the world. Also, we used specific anemia keywords to collect disease-related tweets, so using other keywords could result in new features that may play an important role in the detection process. We used LDA in this study to model the embedded topics due to the popularity of this method. Future work could use a different technique to extract the hidden themes. An additional limitation is that four types of emotions (anger, fear, sadness, and joy) were extracted from the collected tweets since they are the common types of emotions in the contemporary theories of emotions. Hence, in the future, scholars can examine other emotions and examine their relation to the anemia-related symptoms. Finally, in the current study, we considered anemia disease as a popular disease around the world. Perhaps, future studies could adopt our mechanism to diagnose other diseases in an attempt to enrich the overall understanding of the role of disease-related emotions in the disease recognition process.

8. Conclusion

This study proposed a novel mechanism for recognizing anemia symptoms based on certain types of emotions that are expressed by anemic patients on social media sites. The technical contribution was in mining the collected tweets by establishing an association between anemia-related symptoms and anemia-related emotions in an attempt to accurately identify anemia. We used the k-means algorithm to group the tweets that share similar features. Then, we discovered the hidden disease-related topics via the LDA technique. After that, both disease emotions and symptoms were extracted from these topics and mapped together using the Apriori algorithm. In light of that, we were able to find the types of emotions that anemic patients expressed in their tweets during their illness time. These emotions can be used as a heuristic means to recognize certain disease symptoms that are associated with them. This study also evaluated the merit of the extracted emotions. The prediction results showed that the SMO classifier achieved the best accuracy in recognizing the disease (98.96 %). Besides, the results indicated that fear and sadness emotions are dominant among anemic patients. To our knowledge, the proposed non-invasive mechanism is the first of its kind to diagnose anemia from Twitter using textual information.

Declaration of competing interest

The author declares that no conflict of interest exists.

Acknowledgement

This work was funded by the Researchers Supporting Project number [RSP-2021/157], King Saud University, Riyadh, Saudi Arabia.

References

[1] Zhang N, Ming-Yuan Wei M, Ma Q. Nanomedicines: a potential treatment for blood disorder diseases. Front. Bioeng. Biotechnol. 2019;7:369.
[2] Chaudhari AS, Sontakke AN, Trimbake SB. Hba1c status in type ii diabetes mellitus with and without iron deficiency anemia. International Journal of Biochemistry Research & Review 2020;29(8):114–20.
[3] Kwon J-M, Cho Y, Jeon K-H, Cho S, Kim K-H, Baek SD, Oh B-H. A deep learning algorithm to detect anaemia with ecgs: a retrospective, multicentre study. The Lancet Digital Health 2020;2(7):e358–67.
[4] Bahrami A, Khorasanchi Z, Tayefi M, Avan A, Seifi N, Tavakoly Sany SB, Ghayour-Mobarhan M. Anemia is associated with cognitive impairment in adolescent girls: a cross-sectional survey. Appl Neuropsychol Child 2020;9(2):165–71.
[5] Hsu Y-L, Hung J-Y, Chiang S-Y, Jian S-F, Wu C-Y, Lin Y-S, Kuo P-L. Lung cancer-derived galectin-1 contributes to cancer associated fibroblast-mediated cancer progression and immune suppression through tdo2/kynurenine axis. Oncotarget 2016;7(19):27584.
[6] Stone P, Richards M, Hardy J. Fatigue in patients with cancer. Eur. J. Cancer 1998;34(11):1670–6.
[7] Khan S, Singh M, Gupta N, Singh B. The occurrence and hematological profile in anemic female1. The Institution of Engineers; 2014. p. 87–92.
[8] Baldwin C, Olarewaju O. Hemolytic anemia. 2020. StatPearls [Internet].
[9] Dika Haxhirexha F, Koxha S, Qatipi L, Haxhirexhaj A, Ademi A. Prevalence of iron deficiency anemia among children in the municipality of dibër, North Macedonia. Acta Medica Balkanica 2020;5(9–10):32–8.
[10] Johnson RL, Rubenstein SD. Anemia in the emergency department: evaluation and treatment. Emerg. Med. Pract. 2013;15(11):1–5.
[11] Jain P, Bauskar S, Gyanchandani M. Neural network based non-invasive method to detect anemia from images of eye conjunctiva. Int. J. Imaging Syst.Technol. 2020;30(1):112–25.
[12] Dimauro Baldari, Caivano Colucci, Girardi. Paper presented at the 2018 3rd International Conference on Smart and Sustainable Technologies (SpliTech). 2018.
[13] Chen Y-M, Miaou S-G, Bian H. Examining palpebral conjunctiva for anemia assessment with image processing methods. Comput. Methods Programs Biomed 2016;137:125–35.
[14] Chen Y-M, Miaou S-G. A Kalman filtering and nonlinear penalty regression approach for noninvasive anemia detection with palpebral conjunctiva images. J. Healthc. Eng. 2017:1–11. 9580385.
[15] Tamir A, Jahan CS, Saif MS, Zaman SU, Islam MM, Khan AI, Shahnaz C. Detection of anemia from image of the anterior conjunctiva of the eye by image processing and thresholding. In: In 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC). IEEE; 2017, December. p. 697–701.
[16] Dimauro, De Ruvo S, Di Terlizzi F, Ruggieri A, Volpe V, Colizzi L, Girardi F. Estimate of anemia with new non-invasive systems—a moment of reflection. Electronics 2020;9(5):780.
[17] Bevilacqua V, Dimauro G, Marino F, Brunetti A, Cassano F, Di Maio A, Guarini A. A novel approach to evaluate blood parameters using computer vision techniques. In: In 2016 IEEE International Symposium on Medical Measurements and Applications (MeMeA); 2016, May. p. 1–6. IEEE.
[18] Barsevick AM, Irwin MR, Hinds P, Miller A, Berger A, Jacobsen P, O'Mara A. Recommendations for high-priority research on cancer-related fatigue in children and adults. J. Natl. Cancer. Inst. 2013;105(19):1432–40.
[19] Feldthusen C, Björk M, Forsblad-d'Elia H, Mannerkorpi K, Care U. Perception, consequences, communication, and strategies for handling fatigue in persons with rheumatoid arthritis of working age—a focus group study. Clinical Rheumatology 2013;32(5):557–66.
[20] Neves PA, Castro MC, Oliveira CV, Malta MB, Lourenço BH, Cardoso MA. Effect of vitamin a status during pregnancy on maternal anemia and newborn birth weight: results from a cohort study in the western brazilian amazon. Eur. J. Nutr. 2020;59(1):45–56.
[21] Rahmati S, Azami M, Badfar G, Parizad N, Sayehmiri K. The relationship between maternal anemia during pregnancy with preterm birth: a systematic review and meta-analysis. J. Matern. Fetal. Neonatal Med. 2020;33(15):2679–89.
[22] Singh H, Arya S, Talapatra P, Lather K, Mathur R, Singhania A, Chaudhary V. Assessment of fatigue in rheumatoid arthritis (by functional assessment of chronic illness therapy–fatigue score) and its relation to disease activity and anemia. JCR. J. Clin. Rheumatol. 2014;20(2):87–90.
[23] Xavier ASG, Ferreira SL, de Santana Carvalho ES, de Araújo EM, Cordeiro RC. Perception of women suffering from sickle cell anemia regarding pregnancy: an exploratory study. Online Brazilian Journal of Nursing 2013;12(4):834–43.
[24] Parhizkar A. The relation between anemia and postpartum depression in pregnant women who referred to health and medical centers of Sanandaj in 2011–2012. Life Sci J 2013;10(7s):308–12.
[25] Powers JM, Nagel M, Raphael JL, Mahoney DH, Buchanan GR, Thompson DI. Barriers to and facilitators of iron therapy in children with iron deficiency anemia. J. Pediatr. 2020;219:202–8.
[26] Levy KN, Scala JW, Temes CM, Clouthier TL. An integrative attachment theory framework of personality disorders. In: Personality disorders: Toward theoretical and empirical integration in diagnosis and assessment; 2015. p. 315–43.
[27] Kullar R, Goff DA, Gauthier TP, Smith TC. To tweet or not to tweet—a review of the viral power of twitter for infectious diseases. Curr. Infect. Dis. Rep. 2020;22:1–6.
[28] Lim S, Tucker CS, Kumara S. An unsupervised machine learning model for discovering latent infectious diseases using social media data. J. Biomed. Inform. 2017;66:82–94.

[29] Kostkova P, De Quincey E, Jawaheer G. The potential of social networks for early warning nad outbreak detection systems: the swine flu twitter study. Int. J. Infect. Dis. 2010;14:e384–5.
[30] Odlum M, Yoon S. What can we learn about the ebola outbreak from tweets? Am. J. Infect. Control. 2015;43(6):563–71.
[31] Sarsam SM, Al-Samarraie H, Ismail N, Zaqout F, Wright B. A real-time biosurveillance mechanism for early-stage disease detection from microblogs: a case study of interconnection between emotional and climatic factors related to migraine disease. NetMAHIB 2020;9(1):32.
[32] Karami A, Dahl AA, Turner-McGrievy G, Kharrazi H, Shaw Jr G. Characterizing diabetes, diet, exercise, and obesity comments on twitter. Int. J.Inf. Manag. 2018; 38(1):1–6.
[33] Sarsam SM, Al-Samarraie H, Al-Sadi A. Disease discovery-based emotion lexicon: a heuristic approach to characterise sicknesses in microblogs. Network Modeling Analysis in Health Informatics and Bioinformatics 2020;9(1):1–10.
[34] Sarsam SM, Al-Samarraie H, Omar B. Geo-spatial-based emotions: A mechanism for event detection in microblogs. In: Proceedings of the 2019 8th international conference on software and computer applications; 2019, February. p. 1–5.
[35] Sarsam SM, Al-Samarraie H. A first look at the effectiveness of personality dimensions in promoting users' satisfaction with the system. SAGE Open 2018;8 (2). 2158244018769125.
[36] Ran X, Zhou X, Lei M, Tepsan W, Deng W. A novel k-means clustering algorithm with a noise algorithm for capturing urban hotspots. Appl. Sci. 2021;11(23):11202.
[37] Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, Zhao L. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. Multimed. Tools Appl. 2019;78(11):15169–211.
[38] Sievert C, Shirley K. LDAvis: A method for visualizing and interpreting topics. In: Proceedings of the workshop on interactive language learning, visualization, and interfaces; 2014, June. p. 63–70.
[39] Almatarneh S, Gamallo P. A lexicon based method to search for extreme opinions. PloS one 2018;13(5):e0197816.
[40] Mohammad SM. Word affect intensities. arXiv preprint arXiv:1704.08798. 2017.
[41] Sarsam SM, Al-Samarraie H, Alzahrani AI, Alnumay W, Smith AP. A lexicon-based approach to detecting suicide-related messages on twitter. Biomed. Signal Process. Control 2021;65:102355.
[42] Aquino PA, López VF, Moreno MN, Muñoz MD, Rodríguez S. In: Opinion mining system for twitter sentiment analysisInternational conference on hybrid artificial intelligence systems. Cham: Springer; 2020, November. p. 465–76.
[43] Cheng L, De Vos J, Zhao P, Yang M, Witlox F. Examining non-linear built environment effects on elderly's walking: a random forest approach. Transp. Res. Part D: Transp. Environ. 2020;88:102552.
[44] Corazza M. A note on "portfolio selection under possibilistic mean-variance utility and a smo algorithm". Eur. J. Oper. Res. 2020;197(2):693–700.
[45] Al-Samarraie H, Sarsam SM, Alzahrani AI, Alalwan N. Personality and individual differences: the potential of using preferences for visual stimuli to predict the big five traits. Cogn. Tech.Work 2018;20(3):337–49.
[46] Han I, Kamber M, Pei J. Data Mining: Concepts and Techniques Third Edition. Elsevier; 2012.
[47] Cohen J. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 1960; 20(1):37–46.
[48] Witten IH, Frank E, Hall MA, Pal CJ. Data mining: practical machine learning tools and techniques. Morgan Kaufmann; 2016.
[49] Kasiviswanathan S, Bai Vijayan T, Simone L, Dimauro G. Semantic segmentation of conjunctiva region for non-invasive anemia detection applications. Electronics 2020;9(8):1309.
[50] Anggraeni MD, Fatoni A. Non-invasive self-care anemia detection during pregnancy using a smartphone camera. In IOP Conference Series: Materials Science and Engineering 2017, February;172(1):012030. IOP Publishing.
[51] Jamwal M, Sharma P, Das R. Laboratory approach to hemolytic anemia. The Indian J. Pediatr. 2020;87(1):66–74.
[52] Kumar S, Anukiruthika T, Dutta S, Kashyap A, Moses JA, Anandharamakrishnan C. Iron deficiency anemia: a comprehensive review on iron absorption, bioavailability and emerging food fortification approaches. Trends Food Sci. Technol. 2020;99: 58–75.
[53] Wolf M, Rubin J, Achebe M, Econs MJ, Peacock M, Imel EA, Brandenburg V. Effects of iron isomaltoside vs ferric carboxymaltose on hypophosphatemia in iron-deficiency anemia: two randomized clinical trials. JAMA 2020;323(5):432–43.
[54] Mohamed MSA. Evaluation of plasma Interleukin-10 level in sudanese iron deficiency anemia patients among Khartoum state. Sudan University of Science and Technology); 2019 (Doctoral dissertation.
[55] Camaschella C. Iron-deficiency anemia. N. Engl. J. Med. 2015;372(19):1832–43.
[56] Capriotti TM, Frizzell JP. Pathophysiology: introductory concepts and clinical perspectives: FA Davis company. Faculty Bookshelf 2015;75. Retrieved from, https://hsrc.himmelfarb.gwu.edu/books/75.
[57] Phillips J, Henderson AC. Hemolytic anemia: evaluation and differential diagnosis. Am. Fam. Physician 2018;98(6):354–61.

Samer Sarsam is a senior lecturer at the Department of Business Analytics, Sunway University Business School, Sunway University. His border research area is in Human-Computer Interaction (HCI) with emphasis on Natural Language Processing (NLP) and patterns recognition. He is familiar with various machine learning approaches including supervised and unsupervised learning techniques. Sarsam is interested in applying data mining and machine learning schemes for processing different neurological signals and physiological measures. His recent projects have mainly focused on event detection in microblogs, reinforcing the decision-making process in chemometrics, and enhancing the User Experience (UX) with an interface.

Hosam Al-Samarraie is an associate professor in digital innovation design, University of Leeds. His research brings together knowledge from Human-Computer Interaction, Human-Centred Design, and Psychology into UI/UX, education, healthcare, and services–with emphasis on exploring, understanding, comparing, and predicting individuals' use of digital technology.

Ahmed Ibrahim Alzahrani is an associate professor at the computer science department, community college, King Saud University. His main research interests span over the area of Information Systems including adoption and diffusion of emerging information technologies, IT management and human inter-action with social media and latest technologies.

Abdul Samad Shibghatullah is an Associate Professor at the Institute of Computer Science and Digital Innovation, UCSI University as an Associate Professor. He does research in Artificial Intelligence, Optimization and Simulation.